EXPLORING THE EXTINCTION OF RETROTRANSPOSONS IN MAMMALIAN GENOMES

A Dissertation

Presented in Partial Fulfillment of the Requirement for the

Degree of Doctor of Philosophy

with a

Major in Bioinformatics and Computational Biology

in the

College of Graduate Studies

University of Idaho

by

Lei Yang

December 2013

Major Professor: Holly A. Wichman

AUTHORIZATION TO SUBMIT DISSERTATION

This dissertation of Lei Yang, submitted for the degree of Doctor of Philosophy with a Major in Bioinformatics and Computational Biology and titled "Exploring the Extinction of Retrotransposons in Mammalian Genomes" has been reviewed in final form. Permission, as indicated by the signatures and dates below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor:

_____ Date: _____
Dr. Holly A. Wichman

Committee Members:

_____ Date: _____
Dr. Wenfeng An

_____ Date: _____
Dr. Celeste J. Brown

_____ Date: _____
Dr. James A. Foster

Department Administrator:

_____ Date: _____
Dr. Eva M. Top

Discipline's College Dean:

_____ Date: _____
Dr. Paul Joyce

Final Approval and Acceptance by the College of Graduate Studies:

_____ Date: _____
Dr. Jie Chen

ABSTRACT

LINEs and SINEs are mobile genetic elements in mammalian genomes which move by retrotransposition, although SINEs are dependent on LINEs. Together LINEs and SINEs comprise approximately a quarter of a typical mammalian genome. However, the major family of LINEs, L1, was found to have become inactive in the whole megabat family ~24 MYA and in a large group of South American rodents ~8 MYA. Besides, in these rodents, a family of SINEs, B1, lost its activity prior to that of L1 despite its dependency on L1s. Examination of the evolutionary history of L1 in these L1-extinct groups revealed a surprising diversity. Megabat L1s are unique in that two parallel and fairly synchronized L1 lineages persisted in the genome and both underwent extinction soon after a significant wave of L1 deposition occurred. Reconstructions of the most recent common ancestor of the extinct megabat L1 were tested in tissue culture assays and actively retrotransposed. The evolutionary history and reconstruction of the megabat L1 suggests that L1 extinction is unlikely the consequence of degenerative L1 sequence or long-term L1 quiescence. The L1 and B1 evolutionary histories in the South American rodents show that L1s maintained activity until after the split of the basal group carrying active L1s but inactive B1s. B1 retrotransposition tempo is comparable in the L1-extinct clade and the basal group; the most recent wave of B1 retrotransposition is prior to the separation of the basal group and this wave is the largest one detected. Thus, in both the megabat and rodent cases there was a large wave of retrotransposition prior to L1 extinction, suggesting that completion between elements, or between elements and the host, may have contributed to L1 extinction. The study of mammalian genome evolution in non-model organisms has become increasingly viable in the current genomic era and will continue to broaden our understanding of the complex regulatory mechanisms of life.

ACKNOWLEDGMENTS

I could not have reached this point without the help of my family, colleagues and friends. During my Ph.D., I have learned and experienced much more than I originally expected. These years will be a treasure of my life forever.

I would like to thank my advisor Dr. Holly Wichman for opening my door into the amazing world of transposable elements and bioinformatics, for her sincere dedication on my path to successful career development and her kind care for my personal life. I would like to thank my committee members Drs. Wenfeng An, Celeste Brown and James Foster for offering valuable comments and discussions on my dissertation, for Dr. An's help for us to establish the tissue culture lab and technical support on L1 cloning, and for Drs. Brown and Foster's training on bioinformatics and ideas on developing the pipelines.

I would like to thank our lab manager LuAnn Scott for her training and technical support in the lab as well as her help outside the lab. I would like to thank the Wichman lab members for helpful discussions and support. I would like to thank IBEST (Institute for Bioinformatics and Evolutionary Studies) members for having me in this great family. I learned so much in this interdisciplinary environment.

I would like to thank Dr. Astrid Roy-Engel for providing the backbone plasmid used in Chapter 1 and her technical support; Dr. Jerzy Jurka for the bioinformatics training on transposable elements and his welcome advice and generously allowing us to perform the L1 evolutionary history analysis in Chapter 1 on the clusters of Genetic Information Research Institute; the IBEST Genomic Resources Core and Computer Resources Core for helping us to generate the data and perform the analysis for Chapter 2.

I would like to thank my parents for their support all the way across the ocean. You are the power for me to insist on my way to realize my dream. We are so close that it feels like that I have never left you. Your son has grown up to be a scientist that you can be proud of. I would like to thank my friends in Moscow. You made my Ph.D. career the most amazing time in my life.

My dissertation research is funded by NIH-R01-GM38737 to Holly Wichman and Chapter 2 is also supported by NSF-DDIG-1210694 to Holly Wichman and Lei Yang. I have been supported by an IBEST fellowship, the Bioinformatics and Computational Biology program fellowship and a teaching assistantship of the Department of Biological Sciences at the University of Idaho.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Human genome sequencing [1-3] propelled us into the genomics era. Since then, our understanding of mammalian genomes has grown steadily [4-13]. Two of the most unexpected features of the human genome are the number of genes (~20,000 per haploid genome) and percent of the genome responsible for the genes (~3%), in that these numbers are much smaller than expected [3,14]. Half of the human genome is composed of transposable elements, with the remainder largely being made up of unidentified sequences which are presumably old transposable elements beyond the recognition threshold [1,3]. This landscape has become better defined as more mammalian genomes are annotated [4-13], and it is commonly recognized that transposable elements are the major contributors to the size of a mammalian genome [15].

Transposable elements are mobile genetic elements and their movement within the genome is termed transposition. Transposable elements are classified according to their structure and manner of transposition [16,17]. DNA transposons mobilize by either a "copy-and-paste" or "cut-and-paste" mechanism with the aid of a transposase. Retrotransposons mobilize through an RNA intermediate with the aid of reverse transcriptase and follow the "copy-and-paste" mechanism termed retrotransposition. This results in increasing copy numbers so that retrotransposons now comprise a large proportion of most eukaryotic genomes. Retrotransposons are further classified into LTR (Long Terminal Repeat) and non-LTR elements based on the presence or absence of LTRs on their ends, which serve as regulatory elements and also indicate their viral ancestry. Within each class, transposable elements are also classified into autonomous and non-autonomous families based on their dependency on other transposable

elements to mobilize, with the transposition of a non-autonomous element being dependent on that of an autonomous one.

Although transposable element loads in mammalian species are comparable, the landscape of transposable element classes and families varies between different mammalian species. For example, no recent insertion events of DNA transposons have been observed in primates and most other mammals with well-characterized genomes, whereas microbats experienced recent bursts of DNA transposons [18], presumably via horizontal gene transfer [19]. LTR retrotransposons have not been recently active in humans, but were shown to be actively retrotransposing in rodent genomes [6,20-23].

While transposable elements were traditionally considered as "junk" DNA, studies in the genomics era have revealed the impact of transposable elements on their host genomes. Barbara McClintock's insightful view of transposable elements offering the opportunity for reorganizing the genome [24] is supported by increasingly more data with the advances of the genomics era. As transposable elements mobilize and recombine in the genome, they introduce instability [25], cause diseases [23,26] and are occasionally co-opted to serve host functions [27-35]. Transposable elements affect the expression of genes in their vicinity through various methods [36] such as regulatory properties [37,38], coding biases [39] and plentiful splice sites.

Because of the deleterious effects of transposable elements on the genome, they are subject to strict defense mechanisms developed by the host [40-42]. This regulation is strongest in germline cells by means of germline-specific small RNAs [43] and epigenetic silencing [44-47]. The strong germline regulation of transposable elements allows them to accumulate slowly but steadily in the host genome, with rare insertion events in each generation of the host.

Among the mammalian transposable elements, a family of LINEs (Long INterspersed Elements), L1 (LINE-1), which is the focus of this dissertation, distinguishes itself from its come-and-go counterparts by its long-term persistence in the mammalian host genomes. L1s have been co-evolving with the genome since before the divergence of eutherian and metatherian mammals [48,49], ~160 MYA (million years ago) [50]. SINEs (Short INterspersed Elements) are non-autonomous elements dependent on LINEs for their movement. The evolutionary history of independently evolved SINE families is not as long as that of L1s, but different mammalian orders tend to have long co-evolutionary histories with their specific SINE families. For example, the Alu family is responsible for the majority of SINEs in primates and has been co-evolving with its hosts since the radiation of primates, ~65 MYA [51]; multiple families of SINEs, including B1, B2, B4 and ID elements have been found to dominate various rodent species [52].

The prevalence of LINEs and SINEs, their low excision rate and neutral evolution after insertion enabled them to record the evolutionary history of their mammalian hosts [53]. These elements provide a natural genetic fossil record for mammalian genome evolution studies. Although no currently active SINE family traces back to the common ancestor of all mammals, SINEs are major contributors to mammalian genome size and the evolutionary pattern of specific SINE families provides a peek at the evolutionary history of the corresponding mammalian clades.

As our understanding of L1s expands beyond model organisms, their variability in different genomes becomes apparent. Given L1s' ability to introduce instability to the genome and the strong defenses their hosts impose, L1 quiescence or extinction may be expected. Indeed, several occurrences of extinction or current quiescence of L1s have been documented

[54-60]. However, few of these cases have been examined in a phylogenetic context to convincingly demonstrate that extinction, and not simply quiescence, best explains the lack of recent L1 insertions into the genome. Because L1s are transmitted vertically with no evidence of horizontal transmission among mammals, ancient L1 extinctions would affect all subsequent species and should be the most easily identified and confirmed. Early L1 extinctions would have covered large clades of mammals, which have not yet been observed, whereas recent L1 extinctions are difficult to discern from quiescence because the fossil elements have not yet accumulated sufficient divergence from their active ancestors to be recognized as inactive. Thus, either most L1 extinctions are recent or mammalian lineages subject to ancient L1 extinctions do not persist or give rise to few new species. Understanding the dynamics of L1 extinction will be as important as understanding the dynamics of L1 activity in sorting out the impact of L1s on mammalian genome evolution. Two ancient L1 extinction events have been well-documented, one affecting the whole megabat family Pteopodidae ~24 MYA [55] and the other in a group of South American rodents covering the majority of the Sigmodotinae superfamily ~8 MYA [56-58]. These two L1 extinction events are the focus of the two following chapters.

Investigating the evolutionary history of L1s and their corresponding SINEs in different mammalian clades contributes to our understanding the evolution of mammalian genomes. Because fixed L1 and SINE retrotransposition events are permanently recorded in mammalian genomes, they serve as great fossil records for the genome evolutionary history of their hosts. As transposable elements, including LINEs and SINEs, are proposed to derive from ancient acquisition of alien DNA, studying their dynamics, especially their activity in different historical time windows, has the potential to reveal the co-evolution history of LINEs and SINEs with their host cells. Although L1 and SINE evolution has been extensively investigated in human, mouse

and their closely related species, data on non-model organisms is scarce. Besides the evolutionary history of L1s and SINEs related to their extinction, this dissertation work also offered the opportunity to investigate the properties of genomes of non-model organisms and reveal more clearly the diversity of mammalian genomes.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.

2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304-1351.

3. (2004) Finishing the euchromatic sequence of the human genome. Nature 431: 931-945.

4. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438: 803-819.

5. Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, et al. (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326: 865-867.

6. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520-562.

7. Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, et al. (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science 324: 522-528.

8. Miller W, Drautz DI, Ratan A, Pusey B, Qi J, et al. (2008) Sequencing the nuclear genome of the extinct woolly mammoth. Nature 456: 387-390.

9. Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, et al. (2007) Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature 447: 167-177.

10. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69-87.

11. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, et al. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428: 493-521.

12. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, et al. (2010) A draft sequence of the Neandertal genome. Science 328: 710-722.

13. Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, et al. (2007) Evolutionary and biomedical insights from the rhesus macaque genome. Science 316: 222-234.

14. Pennisi E (2012) Genomics. ENCODE project writes eulogy for junk DNA. Science 337: 1159, 1161.

15. Kidwell MG (2002) Transposable elements and the evolution of genome size in eukaryotes. Genetica 115: 49-63.

16. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, et al. (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8: 973-982.

17. Kapitonov VV, Jurka J (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet 9: 411-412; author reply 414.

18. Ray DA, Feschotte C, Pagan HJ, Smith JD, Pritham EJ, et al. (2008) Multiple waves of recent DNA transposon activity in the bat, Myotis lucifugus. Genome Res 18: 717-728.

19. Gilbert C, Schaack S, Pace JK, 2nd, Brindley PJ, Feschotte C (2010) A role for host-parasite interactions in the horizontal transfer of transposons across phyla. Nature 464: 1347-1350.

20. Cantrell MA, Ederer MM, Erickson IK, Swier VJ, Baker RJ, et al. (2005) MysTR: an endogenous retrovirus family in mammals that is undergoing recent amplifications to unprecedented copy numbers. J Virol 79: 14698-14707.

21. Erickson IK, Cantrell MA, Scott L, Wichman HA (2011) Retrofitting the genome: L1 extinction follows endogenous retroviral expansion in a group of muroid rodents. J Virol 85: 12315-12323.

22. Zhang Y, Maksakova IA, Gagnier L, van de Lagemaat LN, Mager DL (2008) Genome-wide assessments reveal extremely high levels of polymorphism of two active families of mouse endogenous retroviral elements. PLoS Genet 4: e1000007.

23. Maksakova IA, Romanish MT, Gagnier L, Dunn CA, van de Lagemaat LN, et al. (2006) Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. PLoS Genet 2: e2.

24. McClintock B. Mechanisms that rapidly reorganize the [maize] genome; 1978.

25. Hedges DJ, Deininger PL (2007) Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. Mutat Res 616: 46-59.

26. Belancio VP, Hedges DJ, Deininger P (2008) Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health. Genome Res 18: 343-358.

27. Cantrell MA, Carstens BC, Wichman HA (2009) X chromosome inactivation and Xist evolution in a rodent lacking LINE-1 activity. PLoS One 4: e6252.

28. Chow JC, Ciaudo C, Fazzari MJ, Mise N, Servant N, et al. (2010) LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. Cell 141: 956-969.

29. Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, et al. (2005) Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature 435: 903-910.

30. Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, et al. (2009) L1 retrotransposition in human neural progenitor cells. Nature 460: 1127-1131.

31. Frendo JL, Olivier D, Cheynet V, Blond JL, Bouton O, et al. (2003) Direct involvement of HERV-W Env glycoprotein in human trophoblast cell fusion and differentiation. Mol Cell Biol 23: 3566-3574.

32. Mi S, Lee X, Li X, Veldman GM, Finnerty H, et al. (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403: 785-789.

33. Dupressoir A, Marceau G, Vernochet C, Benit L, Kanellopoulos C, et al. (2005) Syncytin-A and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae. Proc Natl Acad Sci U S A 102: 725-730.

34. Carbone L, Harris RA, Mootnick AR, Milosavljevic A, Martin DI, et al. (2012) Centromere remodeling in Hoolock leuconedys (Hylobatidae) by a new transposable element unique to the gibbons. Genome Biol Evol 4: 648-658.

35. Cordaux R, Udit S, Batzer MA, Feschotte C (2006) Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. Proc Natl Acad Sci U S A 103: 8101-8106.

36. Rebollo R, Romanish MT, Mager DL (2012) Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu Rev Genet 46: 21-42.

37. Bourque G, Leong B, Vega VB, Chen X, Lee YL, et al. (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res 18: 1752-1762.

38. Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, et al. (2010) Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet 42: 631-634.

39. Han JS, Szak ST, Boeke JD (2004) Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature 429: 268-274.

40. Wissing S, Montano M, Garcia-Perez JL, Moran JV, Greene WC (2011) Endogenous APOBEC3B restricts LINE-1 retrotransposition in transformed cells and human embryonic stem cells. J Biol Chem 286: 36427-36437.

41. Gasior SL, Roy-Engel AM, Deininger PL (2008) ERCC1/XPF limits L1 retrotransposition. DNA Repair (Amst) 7: 983-989.

42. Suzuki J, Yamaguchi K, Kajikawa M, Ichiyanagi K, Adachi N, et al. (2009) Genetic evidence that the non-homologous end-joining repair pathway is involved in LINE retrotransposition. PLoS Genet 5: e1000461.

43. Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ (2007) Developmentally regulated piRNA clusters implicate MILI in transposon control. Science 316: 744-747.

44. Yoder JA, Walsh CP, Bestor TH (1997) Cytosine methylation and the ecology of intragenomic parasites. Trends Genet 13: 335-340.

45. Walsh CP, Chaillet JR, Bestor TH (1998) Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. Nat Genet 20: 116-117.

46. Bourc'his D, Bestor TH (2004) Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature 431: 96-99.

47. Rebollo R, Miceli-Royer K, Zhang Y, Farivar S, Gagnier L, et al. (2012) Epigenetic interplay between mouse endogenous retroviruses and host genes. Genome Biol 13: R89.

48. Smit AF (1996) The origin of interspersed repeats in the human genome. Curr Opin Genet Dev 6: 743-748.

49. Smit AF, Toth G, Riggs AD, Jurka J (1995) Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J Mol Biol 246: 401-417.

50. Luo ZX, Yuan CX, Meng QJ, Ji Q (2011) A Jurassic eutherian mammal and divergence of marsupials and placentals. Nature 476: 442-445.

51. Batzer MA, Deininger PL, Hellmann-Blumberg U, Jurka J, Labuda D, et al. (1996) Standardized nomenclature for Alu repeats. J Mol Evol 42: 3-6.

52. Deininger PL, Tiedge H, Kim J, Brosius J (1996) Evolution, expression, and possible function of a master gene for amplification of an interspersed repeated DNA family in rodents. Prog Nucleic Acid Res Mol Biol 52: 67-88.

53. Huff CD, Xing J, Rogers AR, Witherspoon D, Jorde LB (2010) Mobile elements reveal small population size in the ancient ancestors of Homo sapiens. Proc Natl Acad Sci U S A 107: 2147-2152.

54. Boissinot S, Roos C, Furano AV (2004) Different rates of LINE-1 (L1) retrotransposon amplification and evolution in New World monkeys. J Mol Evol 58: 122-130.

55. Cantrell MA, Scott L, Brown CJ, Martinez AR, Wichman HA (2008) Loss of LINE-1 activity in the megabats. Genetics 178: 393-404.

56. Casavant NC, Scott L, Cantrell MA, Wiggins LE, Baker RJ, et al. (2000) The end of the LINE?: lack of recent L1 activity in a group of South American rodents. Genetics 154: 1809-1817.

57. Grahn RA, Rinehart TA, Cantrell MA, Wichman HA (2005) Extinction of LINE-1 activity coincident with a major mammalian radiation in rodents. Cytogenet Genome Res 110: 407-415.

58. Rinehart TA, Grahn RA, Wichman HA (2005) SINE extinction preceded LINE extinction in sigmodontine rodents: implications for retrotranspositional dynamics and mechanisms. Cytogenet Genome Res 110: 416-425.

59. Waters PD, Dobigny G, Pardini AT, Robinson TJ (2004) LINE-1 distribution in Afrotheria and Xenarthra: implications for understanding the evolution of LINE-1 in eutherian genomes. Chromosoma 113: 137-144.

60. Platt RN, 2nd, Ray DA (2012) A non-LTR retroelement extinction in Spermophilus tridecemlineatus. Gene 500: 47-53.

CHAPTER 2

Reviving the Dead: History and Reactivation of an Extinct L1

Lei Yang[1,2], John Brunsfeld[1], LuAnn Scott[1] and Holly Wichman[1,2]

[1]Department of Biological Sciences & [2]Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, Idaho, United States of America

Abstract

Although L1 sequences are present in the genomes of all placental mammals and marsupials examined to date, their activity was lost in the megabat family Pteropodidae ~24 million years ago. To examine the characteristics of L1s prior to their extinction, we analyzed the evolutionary history of L1s in the genome of a megabat, *Pteropus vampyrus*, and found a pattern of periodic L1 expansion and quiescence. In contrast to the well-characterized L1s in human and mouse, megabat genomes have accommodated two or more simultaneously active L1 families throughout their evolutionary history, and major peaks of L1 deposition into the genome always involved multiple families. We compared the consensus sequences of the two major megabat L1 families at the time of their extinction to consensus L1s of a variety of mammalian species. Megabat L1s are comparable to the other mammalian L1s in terms of adenosine content and conserved amino acids in the open reading frames (ORFs). However, the intergenic region (IGR) of the reconstructed element from the more active family is dramatically longer than the IGR of well-characterized human and mouse L1s. We synthesized the reconstructed element from this L1 family and tested the ability of its components to support retrotransposition in a tissue culture assay. Both ORFs are capable of supporting retrotransposition, while the IGR is inhibitory to retrotransposition, especially when combined with either of the reconstructed ORFs. We dissected the inhibitory effect of the IGR by testing truncated and shuffled versions and found that length is a key factor, but not the only one affecting inhibition of retrotransposition. Although the IGR is inhibitory to retrotransposition, this inhibition does not account for the extinction of L1s in megabats. Overall, neither the evolution of the L1 sequence itself nor the long-term quiescence of L1 is a likely the reason of L1 extinction.

Author Summary

Most of a typical mammalian genome is occupied by transposable elements, which have played an important role in shaping these genomes, and L1s account for approximately half of this transposable element load. Mammals have evolved several mechanisms to control L1 retrotransposition, and yet L1s remain active in almost all mammalian lineages. However, L1s were found to have gone extinct in the megabat family ~24 million years ago. We were able to trace megabat L1s to the ancestral L1 families shared by all mammals as well as identify bat-specific L1 families. Unlike most well-characterized mammals which have a single active L1 lineage, multiple L1 lineages have persisted in megabats throughout their evolutionary history. When the L1 extinction occurred in megabats, two active lineages lost their ability to retrotranspose almost simultaneously after a burst of activity. We synthesized the L1 from the most active family at the time of extinction and found a long intergenic spacer between its two protein coding genes. Tissue culture assays of the reconstructed megabat L1 revealed that both genes supported retrotransposition, but that the spacer is inhibitory. Despite the inhibition, this family accounted for 18% of the L1s detected in the megabat genome.

Introduction

L1 (LINE-1, Long INterspersed Element-1) belongs to the superfamily of autonomously replicating, retrotransposable elements that lack long terminal repeats. Functional L1s are 6,000-7,000 bp long and made up of a 5' untranslated region (5'UTR), two non-overlapping open reading frames (ORFs) known as ORF1 and ORF2, an intergenic region (IGR) usually less than 100 bp and a 3'UTR followed by a poly-adenosine sequence [1]. The proteins encoded by both ORFs are strictly required for L1 retrotransposition and have very strong *cis*-preference [2,3]. The function of the IGR is less well characterized, but it is known to be indispensable for the translation of human ORF2 protein [4] and to serve as an internal ribosome entry site (IRES) in mice [5].

There is considerable evidence that transposable elements, including L1s, have significant effects on the genome. L1 retrotransposition is one of the major sources of mutagenesis and genome instability [6,7]. Besides their copy-and-paste retrotransposition mechanism that interrupts genes and disrupts the normal splicing of messenger RNAs [8], L1s also cleave genomic DNA with the endonuclease they encode [9-13] and are sites of ectopic recombination due to their homology to each other and prevalence throughout the genome [14-18]. L1s and their dependents may be occasionally co-opted to provide host functions. For example, they may serve as the source of new genes [8] or structural chromosome components [19], or regulate genes in their vicinity by various mechanisms [20-22]. They have also been proposed to play a role in X chromosome inactivation [23-25], neuro-plasticity [26-28] and regulatory functions [29].

L1s have been coevolving with their mammalian host genomes since before the eutherians and metatherians diverged [30] more than 160 million years ago (MYA) [31]. The tempo of L1 retrotransposition can vary both between species and at different time intervals within species [32-35]. They evolve as master lineages such that closely related active L1 copies succeed the older masters and become new major contributors to the total retrotransposition events [33,36-38]. Most species are dominated for long periods of time by a single such master lineage [1], although multiple lineages are occasionally active at the same time [32,35,39]. Retrotransposition of the L1 population is extremely inefficient and few new active elements are produced, with the vast majority of new inserts being 5' truncated pseudogenes. There are over 500,000 copies of L1 in the human reference genome [40], but only 80-100 of the L1s in an average human genome are estimated to be full-length and retrotranspositionally competent, with just six of these contributing more than 80% of the total L1 activity. These six elements are closely related; all belong to the youngest family of human L1s, and four of them belong to the youngest clade within that family [41]. Because there is no known mechanism for precise excision of L1s from the genome, old elements accumulate and make up 15-20% of a typical mammalian genome [40,42]. These 'fossil' sequences make it possible to track the activity of L1s within a particular mammalian clade back many millions of years.

One possible reason for this unusual pattern of L1 evolution is that L1s are epigenetically silenced [43,44] and highly regulated by a set of host defense mechanisms [45-48], especially in germline cells. Given the strong host defenses controlling L1 activity, it might seem reasonable to expect L1 extinctions among mammalian lineages. To clarify the terms related to loss of L1 activity in this work, we refer to a prolonged period of low L1 activity as "quiescence" and complete loss of L1 activity as "extinction". Indeed, quiescence or extinction of L1 has been

proposed several times in the literature [32,49-54], but few of these cases have been examined in a phylogenetic context to convincingly demonstrate that extinction, and not simply quiescence, best explains the lack of recent L1 insertions into the genome. Because L1s are transmitted vertically with no evidence of horizontal transmission among mammals, ancient L1 extinctions would affect all subsequent species and should be the most easily identified and confirmed. One well-documented case of L1 extinction occurred in the ancestor of the megabat family, Pteropodidae, which is the focus of this study. The L1 extinction was verified in 11 sampled genera within Pteropodidae, but did not affect other families of bats. The ancestor of the megabats had two active L1 lineages, both of which became extinct at about the same time at least 24 MYA [50].

In this study, the evolutionary history of L1s prior to their extinction in megabats was explored by data-mining the unassembled genome of *Pteropus vampyrus*, the first publicly available genome trace file of the megabat family. At the time of L1 extinction, *P. vampyrus* contained two active L1 lineages. We determined that these lineages likely diverged before the origins of bats. We reconstructed the master element of the more active lineage at the time of L1 extinction and compared its structure to other active L1s, noting particularly that the IGR between the two ORFs is dramatically longer than that of the well-characterized L1s of human and mouse. Finally, we created chimeric L1s between the reconstructed megabat L1 and a human L1 to test the ability of the extinct megabat L1 to support retrotransposition in tissue culture and we manipulated the IGR to explore its effect on retrotransposition.

Results

To be clear about nomenclature used in this paper, we refer to clades of closely related L1s identified by shared, co-segregating sites as *subfamilies*. Closely related subfamilies are grouped into *families* that represent a window of L1 deposition into the genome. These families replace each other sequentially within a clade to form a *lineage*.

Evolutionary history of L1 in megabats

To investigate the history of L1 retrotransposition in the megabats, we identified subfamilies using COSEG in RepeatMasker [55] based on shared, co-segregating sites within 575 bp of the 3' end of ORF2. These were designated subfamilies 0-63 using the convention of the program. The consensus sequences of these subfamilies were subjected to phylogenetic analysis and the phylogenetic relationships were used to identify families with the stipulation that the pairwise distances between subfamilies within a family be no greater than 3.5%. This distance was determined operationally based on the divergence among phylogenetically clustered subfamilies. Given that the L1 masters are constantly being replaced during evolution, perfect designation within large families is not possible. The 3.5% threshold was chosen according to practical observations to cluster closely related subfamilies without inflating the number of families. This method identified 16 L1 families that account for the peaks of L1 fixation in the megabat genome (Figure 1 and Table S1).

Previous work indicated that two major lineages of L1 were active at the time of L1 extinction in megabats [50]. Full-length consensus sequences from two time points in the evolution of each lineage can be found in RepBase [56], designated L1-1_PVa to L1-4_PVa.

COSEG analysis confirms and extends this history. Lineage 1 corresponds to families 1A (L1-2_PVa), 1B (L1-3_PVa) and 1C. Lineage 2 corresponds to families 2A (L1-1_PVa), 2B (L1-4_PVa), 2C and 2D. It is clear that these two lineages existed prior to the emergence of the bats since families 2C and 2D are not bat-specific, but are closely related to elements found in various Laurasiatheria species. The older L1 families identified in our work (5-11) have high identity to the L1 families shared by all placental mammals [57] and by the Laurasiatheria superorder [58]. Smit *et al.* [57] designated the ancestral mammalian L1 families from most recent to oldest as L1MA, L1MB, L1MC, L1MD and L1ME. Subfamilies within each family are identified by number, with 1 being the most recent. The bottom panel of Figure 1 places megabat L1 dynamics in the context of these ancestral L1 families and the extant L1 lineages of primates and rodents. The relationship between the COSEG subfamilies, families and the ancestral L1s are summarized in Table S1.

Tempo of L1 activity and extinction in megabats

To examine the activity and extinction of L1s in megabats, we extracted 79,978 L1 sequences from the ORF2 of L1s in the ~2x unassembled shotgun sequence of the *P. vampyrus* genome (Baylor College of Medicine) and assigned them to one of the subfamilies described above based on sequence similarity. The age of each sequence was approximated by its percent identity to the subfamily consensus – the higher the percent identity, the younger the sequence. Subfamilies were combined into their designated families as determined by phylogenetic analysis (described above) and the age distribution was determined for each family. Taking all families together, we observed periodic fluctuations in the number of L1s fixed in the genome (Figure 1, top).

At least two large waves of L1 fixation in megabats can be identified in the lineages described above with peaks at 92-93.5% and 87.5-89% similarity to subfamily consensus sequences (Figure 2). Each peak corresponds to activity of two or more families and to multiple lineages. The most recent peak, accounting for 25% of the L1s detected in the megabat genome, corresponds to families 1A and 2A and is megabat-specific. No more recent waves of retrotransposition can be identified, consistent with the extinction of L1 retrotransposition in the common ancestor of megabats ~24 MYA [50]. The next peak, accounting for 13% of detected L1s, corresponds to activity in families 1B, 2B and 2C. A third peak, accounting for 12% of detected L1s, resides at 84.5-85.5% and corresponds to families 2D and 3; this peak likely represents retrotransposition prior to the origin of bats. Older waves of L1 fixation are also evident and correspond to ancestral mammalian L1 families.

The dynamics of families within lineages 1 and 2 are not perfectly consistent with short bursts of retrotransposition followed by long periods of quiescence. Given the evolutionary pattern of L1 as master lineages, most L1 sequences evolve neutrally after their insertion into the genome. Therefore, the distribution of mutations in elements inserted at the same time should follow a Poisson distribution (*i.e.*, the mean divergence from the consensus is expected to be equal to the variance of the distribution). However, the mean of each family is 1-2% larger than the peak, indicating that the variance of the distribution is higher than that of a Poisson distribution. This increased variance could be due to sequence differences between active L1s in the same subfamily at the time of transposition, a wave of retrotransposition over an extended period of time, errors introduced during L1 retrotransposition, technical noise in the analysis or some combination of these. Interestingly, the highest copy number peak is for family 1A, one of

the two youngest detectable lineages active just prior to L1 extinction. This peak accounts for 18% of the total L1s detected in the megabat genome.

Reconstruction of an extinct L1

We sought to reconstruct a full-length version of the more active L1 lineage in megabats at the time of L1 extinction, synthesize it and test its activity in a tissue culture assay. It was not possible to reconstruct the less active lineage with confidence because the copy number, especially in the 5' end, is too low. Since the extinction of megabat L1 retrotransposition happened in the common ancestor of the family, the retrotransposition history of L1 in *P. vampyrus* represents that of the whole Pteropodidae family.

Reconstruction was conducted on the *P. vampyrus* genome using a consensus-based method, with curated correction of CpG sites. We performed this reconstruction independently, without reference to RepBase [56], thus the RepBase reconstruction served as a way to assess the quality of our reconstruction and a benchmark for problematic areas. Our reconstructed megabat L1 (GenBank accession number KF796623) has 99.7% identity to the RepBase reconstruction (RepBase Reports 10:(3), 474-474, 2010, available at http://www.girinst.org/2010/vol10/issue3/L1-2_PVa.html) at the nucleotide level, with six differences (two in ORF1 and four in ORF2) at the amino acid level. The amino acid differences were examined individually in the original alignments: three resulted from ambiguous nucleotides or frame shifts in the RepBase reconstruction, one from CpG site correction and two from variable sites which we called differently than RepBase. None of these differences were at sites of conserved amino acids (see below). Note that although RepBase designation L1-2_PVa

suggests that this sequence falls within lineage 2, we follow the precedence of Cantrell *et al.* [50] to designate it as a member of lineage 1.

We compared the reconstructed L1 to the most recently active consensus sequences from 31diverse mammalian species (Table 1 and Text S1 and S2). Sequences are taken from RepBase except five which we reconstructed from trace files. As noted in the Materials and Methods, several sequences were edited to restore ORFs. These alterations were generally within A-rich tracts, which are common in L1s and difficult to reconstruct with confidence. Since the 5' end of ORF1 can be non-homologous in different mammalian species [1,59], we used only the conserved region of ORF1 (amino acids 123-321, bp 1273-1869 of L1rp, GenBank accession number AF148856) as well as the region corresponding to full-length ORF2 of L1rp (bp 1987-5814) for this comparison. The orthologous region of the reconstructed megabat ORF1 retains all the conserved amino acid sites, while the reconstructed ORF2 has two private changes (L418V and V671T, bp 3238-3240 and 3997-3999, respectively). These differences are consistent between our reconstruction and L1-2_PVa in RepBase and were verified in the original alignment to assure that they are not ambiguous in our reconstruction.

We investigated the adenosine content of the reconstructed terminal members of megabat lineages 1 and 2 and 31 additional L1 consensus sequences from the mammalian species listed in Table 1. L1 A-content of the two ORFs and the intergenic region (IGR) ranged from 39% to 44.5%, with a mean of 41.9%. Megabat L1 A-content was high among the species examined: lineage 1 ranked fifth at 43.7% and lineage 2 ranked second at 44.3%.

To our surprise, the length of the megabat L1 IGR set it apart from the well-characterized L1s of rodents and primates. The IGR lengths of the surveyed L1 sequences from 31 species are listed in Table 1 and range from 18 to 580 bp. At 445 bp, the IGR of the reconstructed L1 is

dramatically longer than either the median (63 bp) or mean (172 bp) among the species examined. Long IGRs were found among marsupials, Laurasiatheria (which includes bats) and Afrotheria species, but not among Euarchontoglires. Long IGRs are found in megabat families 1A (445 bp) and 1B (481 bp), but the IGR length of families 2A (38 bp) and 2B (26 bp) is comparable to that of the majority of mammalian species. The IGR lengths in the remaining megabat L1 families are unknown. When multiple sequences were available in RepBase, we used the consensus of the most recently active L1 from each species for comparison; therefore, long IGRs could have existed in older or less active clades, or in sequences for which only partial reconstructions are feasible.

Retrotransposition of the reconstructed L1

To ask whether the reconstructed megabat L1 is capable of supporting retrotransposition, we synthesized it and assessed its activity in a retrotransposition rate assay derived from the work of Moran *et al.* [60]. This assay is routinely used to measure retrotransposition rates of L1s in a tissue culture system [47,61-63]. Reconstruction of fossil sequences can be challenging; even one error in reconstruction could block retrotransposition. Therefore, we synthesized the reconstructed gene in three segments and created all possible chimeric combinations using human L1rp [64-66] as a scaffold (Figure 3). Human L1rp is one of the most active natural human L1s characterized to date, and thus provides a robust background against which to test the effect of each L1 segment on retrotransposition rate. An independent L1rp construct, pWA192 [66], was used as a positive control. An ORF1 mutant of L1rp [67] cloned in the same genetic context as the chimeric L1s was used as a negative control. The chimeric L1s are named by the source of their ORFs and IGR – H for human L1rp or B for the reconstructed megabat L1. For

example, HHH represents the two ORFs and IGR of L1rp (GenBank accession number AF148856), BBB represents the reconstructed megabat L1 (GenBank accession number KF796623) and HBH represents the chimeric L1 that includes human ORF1, megabat IGR and human ORF2.

Both reconstructed megabat ORFs support retrotransposition, but at lower rates than the highly active human L1rp (Figure 4). Comparisons between the human L1 (HHH) and the constructs containing either one or both of the megabat ORFs (HHB, BHH and BHB) show that replacing the human ORFs with a corresponding megabat version reduces the retrotransposition rate ~26-fold. We verified retrotransposition in two positive colonies from each construct by ascertaining splicing of the G418 resistance intron by PCR using primers flanking the *neo* cassette (Figure S2). An alternative start codon for ORF2, located in the IGR, would make ORF2 36 bp longer. We tested the retrotransposition rate of chimeric L1s based on this alternative ORF2 and no change in retrotransposition rate pattern was observed (data not shown).

The megabat IGR is inhibitory to retrotransposition. Replacing the native human L1 IGR with that of the reconstructed megabat (HHH→HBH) reduces the retrotransposition rate ~26-fold (Figure 5A), while introducing the human L1 IGR into the reconstructed megabat L1 (BBB→BHB) increases the retrotransposition rate ~40-fold (Figure 5A). In a mixed ORF context, both HHB→HBB and BHH→BBH result in ~30-fold lower retrotransposition rates. Interestingly, the effect of the megabat IGR on the human construct (HHH→HBH) is similar to that seen when replacing either or both ORFs in the human construct with megabat ORFs (HHH→HHB, BHH or BHB). The retrotransposition rates of the chimeric L1s are drastically lowered with the combination of the reconstructed megabat IGR and any of the reconstructed megabat ORFs (BBH, HBB and BBB). Therefore, we conclude that compared to the HHH

construct, the dampening effect of exchanging the ORFs is non-additive (BHB vs. HHB and BHH) while exchanging either ORF and the IGR at the same time is approximately additive (HHB vs. HBB, BHH vs. BBH and BHB vs. BBB). The hypothesis that retrotransposition rate is dependent on the amount of megabat L1 sequence in the construct is contradicted by the retrotransposition rate of BHB, which is largely made of megabat sequence but has a retrotransposition rate similar to those of constructs with only one bat segment (HHB, BHH and HBH).

Dissecting the inhibitory property of the IGR

To further investigate the inhibitory effect of the reconstructed megabat IGR on retrotransposition and its interaction with the L1 ORFs, we manipulated the megabat IGR and tested variants in the chimeric L1 context. Manipulation of the IGR included truncated versions of the full-length IGR, a shuffled version with the same nucleotide composition (GenBank accession number KF796624) and an IGR with the sense-oriented AUG codons in all three reading frames mutated to AGU. We tested these variant IGRs in all four ORF contexts (HXH, HXB, BXH and BXB, where X indicates the IGR variant). We found that while the absolute level of transposition was affected by whether human or megabats ORFs were framing the IGR, the relative decrease in retrotransposition was comparable in all ORF contexts. Therefore, the effect of the manipulated IGR on retrotransposition is shown only in the human L1rp context, HXH, in Figure 5B; the retrotransposition rates of the manipulated IGRs in all other ORF contexts are shown in Figure S3.

To determine whether the inhibitory property of the megabat IGR is due solely to its length, we truncated one-third or two-thirds of the IGR from either the 5’ end, the 3’ end or both.

All the truncated IGRs increase the retrotransposition rate 0.3- to 0.5-fold compared to the full-length version (Figure 5B; HBH compared to 1-148, 149-297, 298-445 and 149-445) except the truncation of the 3' one-third of the IGR (Figure 5B; HBH compared to 1-297), which decreases the retrotransposition rate ~6.9-fold. Thus, while the length of the IGR accounts for part of its retrotransposition inhibition property, there are also effects from other factors.

Although the megabat L1 IGR is inhibitory to retrotransposition compared to its human counterpart, we would expect to see that at this length, the reconstructed IGR still supports retrotransposition better than a randomized version with the same nucleotide composition. The randomized IGR with the same nucleotide composition reduces the retrotransposition rate ~8.8-fold (Figure 5B; Bat compared to Random), suggesting that there is co-adaptation of the resident IGR with the L1 ORFs.

Since it has been proposed that the translation of ORF2 is dependent on the existence of a close upstream ORF termination [4], we expected to see lowered retrotransposition rates with all the small ORFs within the IGR eliminated, as this makes the stop codon of ORF1 the closest stop upstream of ORF2 and reduces the probability that ORF2 translation will reinitiate before the ribosome is released from the L1 transcript. Mutating the AUG codons in all three possible frames of the IGR into AGUs decreases the retrotransposition rate ~3.3-fold compared to the intact bat IGR (Figure 5B; Bat compared to AUG-).

Discussion

Retrotransposition history of megabat L1s

The acknowledged pattern of L1 evolution is that the active elements within a genome are closely related, giving rise to a single active lineage which dominates the total retrotransposition in the genome for a period of time [38]. Eventually the active elements accumulate debilitating mutations and become less active, but occasionally a new active element derived from an old one will emerge in the L1 population. This element can behave like a 'stealth driver' [68] and remain at low activity in the genome for a long period of time. When evolution drives a new element to high activity, the elements derived from it can eventually dominate the genome and give rise to a new family. Repetition of this lifecycle of L1 families results in the periodic fluctuation of L1 activity.

Prior to L1 extinction, megabat L1s experienced periodic fluctuations in the number of elements fixed in the genome. This pattern is also observed in other mammalian clades, and in most cases each peak in copy number is dominated by a single L1 lineage. However, there are exceptions. For example, the human genome has been dominated by a single L1 lineage, but there was a period in primate evolution beginning about 46 MYA when two lineages were simultaneously active [35]. Similarly, two closely related lineages are currently active in the rodent genus *Peromyscus* [39]. Megabats stand out not only for the extinction of their L1s, but because their genomes have been continuously dominated by multiple active lineages with activity peaks of about the same age. Each peak includes two or three divergent families (Figure 2), a pattern that preceded the mammalian radiation and persisted throughout the history of L1 activity in megabats (Figure 1).

Where multiple lineages are maintained, it is possible that they are specialized on different tissue types (*e.g.*, germ line vs. early embryo), and that there is some difference in the mechanisms of host regulation that control their activity. Thus, one lineage could dominate while the other is relatively quiescent, and eventually the second lineage could escape control and the first lineage be silenced. In other words, there is no reason to expect that lineages would have the same peaks of increased retrotransposition. The fact that distinct lineages experienced fairly synchronized periods of activity and quiescence could suggest global rather than lineage-specific regulation of L1 retrotransposition. Peaks of L1 copy number are generally assumed to indicate transpositional bursts attributable to L1 activity, but other factors might account for peaks of L1 fixation in the genome. For example, host population bottlenecks could account for an increase in the rate of L1 fixation in the genome if there is selection against L1 [69], and such bottlenecks would be expected to affect multiple lineages in a similar manner, accounting for simultaneous peaks of fixation. Another possibility is that these peaks are related to the hypothesized role L1s may play in DNA repair due to their propensity to insert into double-stranded breaks [47,51,70,71]. If a genome undergoes a period of extensive DNA damage due to an environmental or biotic assault, insertion into the resulting double-stranded breaks might lead to simultaneous peaks of retrotransposition of whatever L1 families are active at that time.

Reconstruction of the last active L1 in megabats

To further characterize L1s in megabats at the time of their extinction, we reconstructed the full-length common ancestor of the most active family using a consensus-based method. Because of the unusual mode of L1 evolution [33,36-38], consensus-based reconstruction is the preferred method of ancestral state reconstruction [56,72]. Reconstruction is particularly

challenging for an extinct L1 family because of variation between old L1 insertions that have accumulated private mutations after elements inserted into the genome; this variation eventually dwarfs changes that occur as one family gives rise to the next, and thus to the phylogenetic signal relevant to evolution within active lineages. Since progeny of the most active elements within a family are over-represented in the genome, the resulting reconstructed sequence can best be thought of as representing the most active L1 master sequence at the time of L1 extinction.

The reconstructed L1 sequence of megabat family 1A bears some of the features of a canonical L1 consensus from representative species, but also has some special characteristics to take into consideration. Although we identified and confirmed two amino acid changes in the reconstructed megabat ORF2 at sites conserved in all other species, such private changes at otherwise conserved sites were frequently observed in the L1s used for comparison. The number of private changes in the abovementioned L1s from a set of species varies from zero to seven with a median of two (Table 1 and Text S1and S2), which is in line with the number of private changes in the reconstructed megabat L1. These same two changes were observed in the RepBase reconstruction, providing further confidence that they are not artifacts. It should be noted that mutations in this set of mammalian L1s are not totally saturated, so conserved sites are not necessarily functionally constrained, but functionally constrained sites should be among the conserved sites. Some sites likely appear to be conserved because of the limited number of ORFs available for comparison.

An unusual aspect of L1 sequences is their high adenosine content on the coding strand. This A-bias is prominent in the reconstructed megabat L1, which ranks the fifth among the 31 species surveyed. For comparison, the adenosine content of the megabat genome trace file (30%) is also slightly above the average level (29.5%) of the species surveyed (Table 1). The A-

richness of L1 can cause elongation [61] and post-transcriptional splicing defects [73]. It may also give rise to a codon usage pattern in L1s that is different from the codon usage of host genes. This implies that the high A-content of the reconstructed L1 is a possible contributor to its own retrotransposition rate and likely to have a dampening effect. It has been shown that A-bias correction with codon optimization increases the retrotransposition rate of a native, 'hot' mouse L1 by ~200-fold [61]. Although the same optimization only increases retrotransposition rate of human L1rp ~3-fold, the transcription of the codon-optimized L1rp is increased >40-fold [66].

The most unexpected feature of the reconstructed megabat L1 is its long IGR. Alisch *et al.* [4] and Li *et al.* [5] have shown independently that the IGR is indispensable for the translation of L1 ORF2. The work of Alisch *et al.* [4] also demonstrated that the introduction of a long, structured IGR inhibits the retrotransposition of human L1s. This suggests that the long IGRs in megabat L1 lineage 1 may be inhibitory for retrotransposition. We cannot determine from examination of the megabat genome or from the work of Smit *et al.* [57] whether short or long spacers were ancestral among L1s of the Chiroptera (bats). However, L1s with long IGRs can be found in some marsupials, Laurasiatheria and Afrotheria species. We propose that ancestral mammalian L1s may have had long IGRs and that lineages with short IGRs have arisen independently multiple times during mammalian evolution.

Demonstration that the reconstructed sequences are active

To determine whether the reconstructed megabat lineage 1 element was active, we made chimeric sequences using human L1rp, a highly active *de novo* insertion, as a backbone [64,65]. Ideally, these studies would have been carried out in both human and megabat cell lines.

However, not all cell lines – and not all clones of permissive cell lines – support L1 retrotransposition. Megabat cell lines are not readily available, and we are unaware of an immortalized cell line from any bat that supports L1 activity. Fortunately, HeLa cells are competent hosts of heterologous and chimeric L1 retrotransposition. Mouse L1s readily retrotranspose in HeLa cells [74,75] as do chimeras between human and mouse L1s [62]. However, our studies differ from those of Wagstaff *et al.* [62] in that we did not codon optimize our L1 constructs.

Although exchanging the L1rp ORFs with either or both of the corresponding megabat counterparts lowers the retrotransposition rate considerably, the activity of chimeric L1s is comparable to the majority of full-length human L1s. The retrotransposition rate of chimeric constructs containing megabat ORFs is much lower than the retrotransposition rate of the most active 'hot' L1s, but more active than 82% of full-length L1s in the human reference genome [41]. The retrotransposition rate of BBB is even lower, but still surpasses that of 56% of full-length L1s in the human reference genome.

There are some caveats relevant to this comparison. First, the retrotransposition assays of Brouha *et al.* [41] were conducted in a different genetic background from the one in this study, but both studies use relative numbers normalized by the retrotransposition rate of L1rp, and thus are comparable. Secondly, although the reconstructed megabat L1 (BBB) supported retrotransposition at about the rate of the average active human L1, it would not be expected to generate half the number of insertion events as a 'hot' human L1 because the contribution of individual active L1s to the total retrotransposition activity is unevenly distributed – just six 'hot' elements of the 80-100 full-length human L1s are responsible for more than 80% of the total retrotransposition activity [41]. Since the average human L1 barely contributes to the total L1

retrotransposition in the genome, we conclude that the intact reconstructed megabat L1 is able to retrotranspose, but by this measure transposes at a very low rate. The reconstruction did not include the promoter, as L1 retrotransposition driven by a native promoter is difficult to detect in tissue culture assays [63]. Therefore, interactions with heterologous regulatory sequences are not a factor in this assay. No single component of the reconstructed L1s was responsible for the inhibition of retrotransposition compared to L1rp; replacement of each component had a similar effect. This makes it unlikely that either a rate-limiting megabat L1 protein or an interaction with a specific host factor is responsible for dampening activity. We also note that these assays were conducted in a human cell line (HeLa), which is heterologous to the reconstructed L1, so these estimates must be interpreted with caution.

Conclusion

To our knowledge, the L1 reconstruction presented in this work is the only L1 element to have been reconstructed from a species that does not carry currently active L1s and tested in a tissue culture assay. Wagstaff *et al.* [72] showed that reconstructed ancestral lineages of human L1 are capable of retrotransposition, but their reconstructed human L1s were codon-optimized and thus their level of activity is not directly comparable to our work.

The reconstructed IGR is co-adapted with the ORFs to support retrotransposition. This is most evident in the comparison of the randomized IGR with the intact version (Figure 5B), where retrotransposition with the intact IGR is 8.8-fold higher than the randomized version with the same base composition. Although the length of the IGR has a major effect on retrotransposition rate, other factors such as secondary structure and splicing sites of the L1 transcript can also dramatically change the retrotransposition rate. Li *et al.* [5] demonstrated that

the IGR of a 'hot' mouse L1, L1spa, contains an IRES that enhances the translation of a downstream ORF, and the work of Alisch *et al.* [4] suggests that the termination of another ORF directly upstream of the ORF2 start is the key for its translation. Our data demonstrate that the reconstructed L1 containing an AUG-codon-free IGR has a dramatically lower retrotransposition rate than that of the intact version. This is in line with the evidence found by Alisch *et al.* [4] as well as the original work by Horvath *et al.* [76] that proposes a reinitiation mechanism for the translation of dicistronic structures. Perhaps the most difficult aspect to reconcile about the long IGR in lineage 1 is its evolutionary persistence. An active element that deleted this long IGR would be expected to dramatically increase its retrotransposition rate and thus to dominate future retrotransposition. That is to say, there should have been strong selection favoring the deletion of the IGR. One might expect such a deletion to be 'easy' from an evolutionary perspective since it need not maintain a reading frame, and yet this did not happen.

The tempo of L1 retrotransposition in megabats directly preceding L1 extinction is also noteworthy. A significant burst of retrotransposition occurred just prior to L1 extinction in megabats, contributing 25% of the detectable L1s to the genome. Family 1A accounts for the bulk of this activity – 18% of the total detectable elements in the genome – despite the demonstrated inhibitory effect of the long intergenic spacer on this family. The IGR has a long evolutionary history in this L1 lineage and likely preceded the evolution of megabats. Thus, despite its inhibitory effect on retrotransposition, it is unlikely that it contributed to L1 extinction.

There are some characteristics of bat genomes that make them unique among the mammals. Bats, and especially megabats, have much smaller genomes than other mammals [77]. Data from 43 species of megabats, 62 species of microbats and ~10,000 other mammalian

species suggest that at 2.15 Gbp the megabat average genome size is significantly more constrained than the average of all mammals (3.42 Gbp) and is considerably smaller than even the microbats (2.52 Gbp). It has been proposed that small genome size is related to the ability to fly given the high metabolic rate and small cell size requirements of flight [78-80]. For example, it has been shown that bird genomes are smaller and less variable in size than genomes of mammals and amphibians [77] and that their genome size is inversely correlated with their wing loading, an index of flight ability [81]. Although the long IGR is unlikely to have driven the L1 extinction because of its inhibitory property, L1s with long IGR may have contributed to the length of L1, thus the size of the genome given the prevalence of the L1s with long IGRs. Consequently, this may trigger strong host defense to reduce the genome size and drive the extinction of L1s.

Since transposable elements are the major contributor to mammalian genome size [82], pressure to constrain genome size will likely be reflected by stronger regulation of transposable elements. This regulation could theoretically result in both suppression of transposition and more efficient removal of inserted elements from the genome. Loss of L1 activity would be particularly effective in slowing expansion of the genome since L1s and the SINEs (Short INterspersed Elements), that co-op the L1 replication machinery, together make up approximately a quarter of a typical mammalian genome [40,42]. Compared to other mammals, genome size constraint in bats confers a stronger selective pressure on the host defense mechanisms that control L1 retrotransposition, which could serve as the intrinsic driver for the host to develop anti-transposable element strategies that may increase the likelihood of transposable element quiescence and extinction in this group.

Materials and Methods

Bioinformatic analysis of L1 history in megabats

Since the large majority of L1s are truncated at the 5' end [83], the copy number of 3' ends better represents the history of retrotransposition events. Therefore, we used 575 bp in the 3' end of L1 ORF2 (as constructed below) to get a comprehensive view of L1 retrotransposition. Using the megabat L1 lineage 1 [50] consensus as the query sequence, we ran CENSOR 4.2 [84] against the ~2x genome trace files of *P. vampyrus* (Baylor College of Medicine, ftp.ncbi.nlm.nih.gov/pub/TraceDB/pteropus_vampyrus/) to find detectable sequences with >60% identity and >90% coverage of the query. Using 2000 random sequences from the CENSOR run, subfamilies were identified based on shared sequence variants (co-segregating mutations) with COSEG 0.2.1 (http://www.repeatmasker.org/COSEGDownload.html) [55] following the default parameters. Nine subfamilies were generated and their consensuses used as query sequences for a second round of CENSOR against the *P. vampyrus* genome. All identified L1 sequences from the second CENSOR run were used for a second round of COSEG, which required the additional parameter of at least 250 sequences to form a subfamily. Consensuses of the 64 subfamilies thus generated were used as query sequences to run CENSOR for a third time. Each hit's percent identity to the corresponding query was used to assign it to a L1 subfamily, and the copy numbers in each subfamily were counted. Seven subfamilies containing less than 250 sequences were removed. Consensuses from each of the remaining 57 subfamilies were used as query sequences to run CENSOR for a fourth time and all detected L1s were assigned to their subfamilies by the percent identity of each hit to its query. The 57 subfamily consensuses were aligned with ancestral mammalian L1s from RepBase [56], reconstructed by Smit *et al.*

[57] and Wade *et al.* [58], with the Lasergene software suite (DNASTAR, Madison, WI), and a distance matrix was calculated. Based on the alignment, a maximum likelihood tree was constructed using PhyML [85] with the GTR+I+G model and 100 bootstrap replicates (Figure S1). L1s were then assigned to families based on a <3.5% within-family pairwise distance from their subfamily consensuses. Sequence specificity of L1 families was determined by BLAST [86] against the NCBI whole genome sequencing databases. The consensus sequences of subfamilies 1, 5, 7, 3, 40, 36, 34, 0 and 29 were used as the BLAST queries representing families 1A, 1B, 1C, 2A, 2B, 2C, 2D, 3 and 4, respectively. A subfamily and its corresponding family were considered bat-specific only if <5 of the top 100 BLAST hits were not from bats.

Histograms of L1 age distribution were generated by the R [87] histogram function using a window size of 0.5% (Figures 1 and 2). Percent identities corresponding to retrotransposition peaks of individual families (Figure 2) were determined by R using the kernel smoothing function with 0.2% bandwidth.

Bioinformatic reconstruction of an extinct megabat L1

A full-length consensus sequence of the most recently active L1 from megabat lineage 1 was reconstructed by a series of progressive steps. The seed for the reconstruction was a conserved 575 bp region in the 3' half of ORF2 (Figure 3A). This region was previously amplified by degenerate PCR and a consensus sequence was determined [88]. Walks were performed in the 5' and 3' directions away from the cloned region and continued in both directions until full-length L1s were reconstructed. To aid with the reconstruction, a software pipeline was developed consisting of Perl (http://www.perl.org/), Ruby (https://www.ruby-lang.org/en/) and Bash (http://www.gnu.org/software/bash/) scripts. The pipeline queried,

filtered and extracted data from the genome of *P. vampyrus*. An individual step resulted in the addition of 100-500 bp of sequence to the consensus, depending on the quality of the alignment at the ends, which was then used in the next step of the walk and in the final L1 reconstruction. Candidate sequences were identified in the database using BLAST with default parameters and an e-value of $1x10^{-50}$, parsed through the BioPerl SearchIO module (http://www.bioperl.org) and screened based on their similarity to the input sequence. Only hits with at least 92% identity were retained to assure that the reconstruction did not include older lineages, and then a Ruby script extracted those sequences with overhangs of at least 100 bp. Alignments for each end were created and hand-edited to yield consensuses of clean read which were aligned into a master alignment. A 300-500 bp region from each end was selected to act as the seeds for the next step in the walk. The process was repeated until the entire element was reconstructed. Upon completion of the full-length L1, a 500 bp seed was chosen arbitrarily from the final consensus and the pipeline was run again to verify the reconstruction. Methylated CpG sites evolve rapidly and must be corrected in the final consensus. CpG sites were identified by their high variation and the presence of dinucleotide sequence CG, CA, TG or TA; these were examined, manually edited and designated as CG in the final consensus. This pipeline also reconstructed the most recently active L1 lineage of four additional species listed in Table S1, but required higher percent identities for the walks to reduce the noise introduced by older lineages.

To compare the reconstruction of the extinct L1 to other L1s, sequences from a range of mammalian species were either reconstructed as described above, or selected from the RepBase report of February 2013 [56]. L1 consensuses of all species available in RepBase were aligned except those of dolphin and American opossum which had problematic regions of non-

homology. When multiple L1 consensus sequences for the same species were present in RepBase, the one with highest average percent identity to its genomic sequence was chosen to represent the most recent master L1 in the genome. Some of the RepBase L1 sequences were out of frame at regions containing adenosine runs or contained in-frame stop codons, both resulting in significantly shorter ORFs. The following corrections brought these sequences into the correct reading frame: L1-1_Cpo, ignored an in-frame stop codon at bp 3050-3052 and used the original sequence for the alignment; L1-1_DV, added a N after bp 6015; and, L1A_Mim, deleted an A at bp 1590-1591 and bp 5336-5337.

Synthesis and cloning of the chimeric L1s

The backbone plasmid for chimera constructions used in the retrotransposition assays was based on pL1PA1tag, a gift from Dr. Astrid Roy-Engel. pL1PA1tag contains a codon-optimized consensus of the PA1 family of human L1 in a pBSSK$^-$ (Agilent Technologies, Inc., Santa Clara, CA) backbone. A puromycin resistance gene and its affiliated promoter pPGKpuro (Addgene, Cambridge, MA) were cloned into pL1PA1tag, creating plasmid pLY1004. The L1 insert of pLY1004 was removed by *Nhe*I and *Eco*RI digestion, creating the final plasmid backbone (Figures 3B and 3C).

The reconstructed L1 and manipulated IGR sequences were commercially synthesized by GenScript USA, Inc. (Piscataway, NJ). Reconstructed L1s were synthesized in two blocks consisting of ORF1+IGR and ORF2. The manipulated IGRs were synthesized separately or in combinations containing distinct cloning sites. The synthesized sequences were cloned into pUC57 with flanking ends compatible to the linearized pLY1004 backbone and with *Bsa*I or *Bsm*BI sites to generate compatible overhangs after digestion. ORF1 and IGR were subcloned

into separate pUC57 plasmids. Figure 3B illustrates the principle underlying the construction of the chimeric L1s. L1 ORFs and IGRs were amplified from these plasmids by PCR with Phusion high-fidelity polymerase (ThermoFisher Scientific, Waltham, MA) using primers designed to generate compatible overhangs when the PCR products are digested with *Bsa*I, *Btg*ZI or *Eco*RI. Human L1rp segments were cloned from pWA192 [66], a gift from Dr. Wenfeng An, using the same principle. The L1 ORFs, IGRs and the linearized backbone plasmid pLY1004 were joined together by a multi-way ligation using T4 DNA ligase. All restriction enzymes and DNA modifying enzymes were from New England BioLabs, Inc. (Ipswich, MA) unless otherwise specified. All constructs were confirmed by sequencing the L1 insert.

Retrotransposition assays

Retrotransposition rates were tested in an assay derived from Moran *et al.* [60], in which the number of cell colonies surviving G418 antibiotic selection represents the retrotransposition rate (Figure 3C). Briefly, the transcription and retrotransposition of L1 trigger the splicing of the transcript and excision of the intron of the inverse-oriented *neo* cassette, granting the cell resistance to the antibiotic G418.

The HeLa cell line (ATCC CCL-2) was a gift from Dr. Wenfeng An and maintained in Dulbecco's Modified Eagle Medium with 4500 mg/L glucose and 110 mg/L sodium pyruvate (ThermoFisher Scientific) supplemented by 10% fetal bovine serum (Atlanta Biologicals, Lawrenceville, GA), 2 mM l-alanyl-l-glutamine dipeptide and 100 units/mL Penicillin-Streptomycin (ThermoFisher Scientific). The assay was conducted as described by An *et al.* [66]. The culture medium for antibiotic selection was similar to the cell maintenance medium except 2.5 ug/mL puromycin (CALBIOCHEM, Billerica, MA) or 50 mg/mL G418

(CALBIOCHEM) was added. Plasmids for transfection were prepared with the Promega (Fitchburg, WI) PureYield Plasmid Midiprep System and the cells were transfected with FuGENE HD transfection reagent (Promega) following the manufacturer's protocol. Retrotransposition assays of the chimeric L1s were repeated at least 12 times in three different batches and manipulated IGR assays were repeated at least four times.

To confirm retrotransposition, two retrotransposition-positive colonies of each chimeric L1 construct were isolated with cloning rings, dissociated with trypsin (ThermoFisher Scientific), seeded on T75 flasks and allowed to grow into confluence. Cells were harvested and their genomic DNA was extracted with the QIAamp DNA mini kit (QIAGEN, Germantown, MD). Genotyping PCRs were conducted with primers bracketing the intron of the G418 reporter gene as described by An *et al.* [89]. Briefly, genotyping PCR primers were designed to the *neo* cassette so that cells hosting retrotransposition events, and the corresponding spliced cassettes, yield 653 bp PCR products. pLY1101, a self-ligated version of the linearized pLY1004 without a L1 insertion, was constructed as a positive control; genotyping PCR of pLY1101 yields a 1556 bp construct corresponding to the unspliced *neo* cassette.

Author's Contributions

Conceived and designed the experiments: LY HAW

Performed the experiments: LY JB

Built the bioinformatic pipeline for L1 reconstruction: JB

Performed the L1 reconstructions: JB and LY

Conducted the retrotransposition experiments: LY

Bioinformatics on evolutionary history LY

Analyzed the data: LY LS HAW

Performed the L1 evolution history analysis: LY and HAW

Verified the reconstructions: LS

Contributed reagents/materials/analysis tools: HAW LY JB

Reagents: HAW

Analysis tools: LY JB

Wrote the manuscript: LY LS HAW

Revised the manuscript: HAW and LS

Provided technical assistance: LS

Acknowledgements

We thank Dr. Jerzy Jurka for bioinformatics training, welcome advice and generously allowing us to perform the L1 evolution history analysis on the cluster of the Genetic Information Research Institute. We thank Dr. Celeste Brown and for ideas on how to develop a bioinformatics pipeline for reconstructing ancient L1s; Dr. Wenfeng An for generously providing the L1rp plasmid, advice on cloning, the cell line and technical support that allowed us to establish our tissue culture lab; Dr. Astrid Roy-Engel for providing the backbone plasmid used in our study and her technical support; Drs. Craig Miller and Martina Ederer for helpful discussions.

References

1. Furano AV (2000) The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons. Prog Nucleic Acid Res Mol Biol 64: 255-294.

2. Kulpa DA, Moran JV (2006) Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. Nat Struct Mol Biol 13: 655-660.

3. Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, et al. (2001) Human L1 retrotransposition: cis preference versus trans complementation. Mol Cell Biol 21: 1429-1439.

4. Alisch RS, Garcia-Perez JL, Muotri AR, Gage FH, Moran JV (2006) Unconventional translation of mammalian LINE-1 retrotransposons. Genes Dev 20: 210-224.

5. Li PW, Li J, Timmerman SL, Krushel LA, Martin SL (2006) The dicistronic RNA from the mouse LINE-1 retrotransposon contains an internal ribosome entry site upstream of each ORF: implications for retrotransposition. Nucleic Acids Res 34: 853-864.

6. Belancio VP, Hedges DJ, Deininger P (2008) Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health. Genome Res 18: 343-358.

7. Chen JM, Ferec C, Cooper DN (2006) LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease: mutation detection bias and multiple mechanisms of target gene disruption. J Biomed Biotechnol 2006: 56182.

8. Moran JV, DeBerardinis RJ, Kazazian HH, Jr. (1999) Exon shuffling by L1 retrotransposition. Science 283: 1530-1534.

9. Gilbert N, Lutz-Prigge S, Moran JV (2002) Genomic deletions created upon LINE-1 retrotransposition. Cell 110: 315-325.

10. Symer DE, Connelly C, Szak ST, Caputo EM, Cost GJ, et al. (2002) Human l1 retrotransposition is associated with genetic instability in vivo. Cell 110: 327-338.

11. Garcia-Perez JL, Marchetto MC, Muotri AR, Coufal NG, Gage FH, et al. (2007) LINE-1 retrotransposition in human embryonic stem cells. Hum Mol Genet 16: 1569-1577.

12. Gasior SL, Wakeman TP, Xu B, Deininger PL (2006) The human LINE-1 retrotransposon creates DNA double-strand breaks. J Mol Biol 357: 1383-1393.

13. Feng Q, Moran JV, Kazazian HH, Jr., Boeke JD (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87: 905-916.

14. Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE (2003) Size matters: non-LTR retrotransposable elements and ectopic recombination in Drosophila. Mol Biol Evol 20: 880-892.

15. Deininger PL, Batzer MA (1999) Alu repeats and human disease. Mol Genet Metab 67: 183-193.

16. Han K, Lee J, Meyer TJ, Remedios P, Goodwin L, et al. (2008) L1 recombination-associated deletions generate human genomic variation. Proc Natl Acad Sci U S A 105: 19366-19371.

17. Burwinkel B, Kilimann MW (1998) Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease. J Mol Biol 277: 513-517.

18. Wichman HA, Van den Bussche RA, Hamilton MJ, Baker RJ (1992) Transposable elements and the evolution of genome organization in mammals. Genetica 86: 287-293.

19. Carbone L, Harris RA, Mootnick AR, Milosavljevic A, Martin DI, et al. (2012) Centromere remodeling in Hoolock leuconedys (Hylobatidae) by a new transposable element unique to the gibbons. Genome Biol Evol 4: 648-658.

20. Rebollo R, Farivar S, Mager DL (2012) C-GATE - catalogue of genes affected by transposable elements. Mob DNA 3: 9.

21. Rebollo R, Romanish MT, Mager DL (2012) Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu Rev Genet 46: 21-42.

22. Han JS, Szak ST, Boeke JD (2004) Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature 429: 268-274.

23. Cantrell MA, Carstens BC, Wichman HA (2009) X chromosome inactivation and Xist evolution in a rodent lacking LINE-1 activity. PLoS ONE 4: e6252.

24. Chow JC, Ciaudo C, Fazzari MJ, Mise N, Servant N, et al. (2010) LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. Cell 141: 956-969.

25. Lyon MF (2003) The Lyon and the LINE hypothesis. Semin Cell Dev Biol 14: 313-318.

26. Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, et al. (2005) Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature 435: 903-910.

27. Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, et al. (2009) L1 retrotransposition in human neural progenitor cells. Nature 460: 1127-1131.

28. Muotri AR, Gage FH (2006) Generation of neuronal variability and complexity. Nature 441: 1087-1093.

29. Sasaki T, Nishihara H, Hirakawa M, Fujimura K, Tanaka M, et al. (2008) Possible involvement of SINEs in mammalian-specific brain formation. Proc Natl Acad Sci U S A 105: 4220-4225.

30. Smit AF (1996) The origin of interspersed repeats in the human genome. Curr Opin Genet Dev 6: 743-748.

31. Luo ZX, Yuan CX, Meng QJ, Ji Q (2011) A Jurassic eutherian mammal and divergence of marsupials and placentals. Nature 476: 442-445.

32. Boissinot S, Roos C, Furano AV (2004) Different rates of LINE-1 (L1) retrotransposon amplification and evolution in New World monkeys. J Mol Evol 58: 122-130.

33. Casavant NC, Hardies SC (1994) The dynamics of murine LINE-1 subfamily amplification. J Mol Biol 241: 390-397.

34. Sookdeo A, Hepp CM, McClure MA, Boissinot S (2013) Revisiting the evolution of mouse LINE-1 in the genomic era. Mob DNA 4: 3.

35. Khan H, Smit A, Boissinot S (2006) Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res 16: 78-87.

36. Pascale E, Liu C, Valle E, Usdin K, Furano AV (1993) The evolution of long interspersed repeated DNA (L1, LINE 1) as revealed by the analysis of an ancient rodent L1 DNA family. J Mol Evol 36: 9-20.

37. Adey NB, Schichman SA, Graham DK, Peterson SN, Edgell MH, et al. (1994) Rodent L1 evolution has been driven by a single dominant lineage that has repeatedly acquired new transcriptional regulatory sequences. Mol Biol Evol 11: 778-789.

38. Clough JE, Foster JA, Barnett M, Wichman HA (1996) Computer simulation of transposable element evolution: random template and strict master models. J Mol Evol 42: 52-58.

39. Casavant NC, Lee RN, Sherman AN, Wichman HA (1998) Molecular evolution of two lineages of L1 (LINE-1) retrotransposons in the california mouse, Peromyscus californicus. Genetics 150: 345-357.

40. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.

41. Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, et al. (2003) Hot L1s account for the bulk of retrotransposition in the human population. Proc Natl Acad Sci U S A 100: 5280-5285.

42. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520-562.

43. Yoder JA, Walsh CP, Bestor TH (1997) Cytosine methylation and the ecology of intragenomic parasites. Trends Genet 13: 335-340.

44. Bourc'his D, Bestor TH (2004) Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature 431: 96-99.

45. Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ (2007) Developmentally regulated piRNA clusters implicate MILI in transposon control. Science 316: 744-747.

46. Wissing S, Montano M, Garcia-Perez JL, Moran JV, Greene WC (2011) Endogenous APOBEC3B restricts LINE-1 retrotransposition in transformed cells and human embryonic stem cells. J Biol Chem 286: 36427-36437.

47. Gasior SL, Roy-Engel AM, Deininger PL (2008) ERCC1/XPF limits L1 retrotransposition. DNA Repair (Amst) 7: 983-989.

48. Suzuki J, Yamaguchi K, Kajikawa M, Ichiyanagi K, Adachi N, et al. (2009) Genetic evidence that the non-homologous end-joining repair pathway is involved in LINE retrotransposition. PLoS Genet 5: e1000461.

49. Waters PD, Dobigny G, Pardini AT, Robinson TJ (2004) LINE-1 distribution in Afrotheria and Xenarthra: implications for understanding the evolution of LINE-1 in eutherian genomes. Chromosoma 113: 137-144.

50. Cantrell MA, Scott L, Brown CJ, Martinez AR, Wichman HA (2008) Loss of LINE-1 activity in the megabats. Genetics 178: 393-404.

51. Grahn RA, Rinehart TA, Cantrell MA, Wichman HA (2005) Extinction of LINE-1 activity coincident with a major mammalian radiation in rodents. Cytogenet Genome Res 110: 407-415.

52. Casavant NC, Scott L, Cantrell MA, Wiggins LE, Baker RJ, et al. (2000) The end of the LINE?: lack of recent L1 activity in a group of South American rodents. Genetics 154: 1809-1817.

53. Rinehart TA, Grahn RA, Wichman HA (2005) SINE extinction preceded LINE extinction in sigmodontine rodents: implications for retrotranspositional dynamics and mechanisms. Cytogenet Genome Res 110: 416-425.

54. Platt RN, 2nd, Ray DA (2012) A non-LTR retroelement extinction in Spermophilus tridecemlineatus. Gene 500: 47-53.

55. Smit A, Hubley R (1996-2010) RepeatMasker Open-3.0.

56. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110: 462-467.

57. Smit AF, Toth G, Riggs AD, Jurka J (1995) Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J Mol Biol 246: 401-417.

58. Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, et al. (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326: 865-867.

59. Scott AF, Schmeckpeper BJ, Abdelrazik M, Comey CT, O'Hara B, et al. (1987) Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. Genomics 1: 113-125.

60. Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, et al. (1996) High frequency retrotransposition in cultured mammalian cells. Cell 87: 917-927.

61. Han JS, Boeke JD (2004) A highly active synthetic mammalian retrotransposon. Nature 429: 314-318.

62. Wagstaff BJ, Barnerssoi M, Roy-Engel AM (2011) Evolutionary conservation of the functional modularity of primate and murine LINE-1 elements. PLoS ONE 6: e19672.

63. Naas TP, DeBerardinis RJ, Moran JV, Ostertag EM, Kingsmore SF, et al. (1998) An actively retrotransposing, novel subfamily of mouse L1 elements. EMBO J 17: 590-597.

64. Schwahn U, Lenzner S, Dong J, Feil S, Hinzmann B, et al. (1998) Positional cloning of the gene for X-linked retinitis pigmentosa 2. Nat Genet 19: 327-332.

65. Kimberland ML, Divoky V, Prchal J, Schwahn U, Berger W, et al. (1999) Full-length human L1 insertions retain the capacity for high frequency retrotransposition in cultured cells. Hum Mol Genet 8: 1557-1560.

66. An W, Dai L, Niewiadomska AM, Yetil A, O'Donnell KA, et al. (2011) Characterization of a synthetic human LINE-1 retrotransposon ORFeus-Hs. Mob DNA 2: 2.

67. Ostertag EM, Prak ET, DeBerardinis RJ, Moran JV, Kazazian HH, Jr. (2000) Determination of L1 retrotransposition kinetics in cultured cells. Nucleic Acids Res 28: 1418-1423.

68. Cordaux R, Batzer MA (2009) The impact of retrotransposons on human genome evolution. Nat Rev Genet 10: 691-703.

69. Nei M, Maruyama T, Chakraborty R (1975) The bottleneck effect and genetic variability in populations. Evolution: 1-10.

70. Hutchison CA, III, Hardies SC, Loeb DD, Shehee WR, Edgell MH (1989) LINEs and related retroposons: long interspersed repeated sequences in the eucaryotic genome. In: Berg DE, Howe MM, editors. Mobile DNA. Washington DC: American Society for Microbiology. pp. 593-617.

71. Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD, et al. (2002) DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat Genet 31: 159-165.

72. Wagstaff BJ, Kroutter EN, Derbes RS, Belancio VP, Roy-Engel AM (2013) Molecular reconstruction of extinct LINE-1 elements and their interaction with nonautonomous elements. Mol Biol Evol 30: 88-99.

73. Belancio VP, Hedges DJ, Deininger P (2006) LINE-1 RNA splicing and influences on mammalian gene expression. Nucleic Acids Res 34: 1512-1521.

74. Martin SL, Branciforte D (1993) Synchronous expression of LINE-1 RNA and protein in mouse embryonal carcinoma cells. Mol Cell Biol 13: 5383-5392.

75. Streva VA, Faber ZJ, Deininger PL (2013) LINE-1 and Alu retrotransposition exhibit clonal variation. Mob DNA 4: 16.

76. Horvath CM, Williams MA, Lamb RA (1990) Eukaryotic coupled translation of tandem cistrons: identification of the influenza B virus BM2 polypeptide. EMBO J 9: 2639-2647.

77. Smith JD, Gregory TR (2009) The genome sizes of megabats (Chiroptera: Pteropodidae) are remarkably constrained. Biol Lett 5: 347-351.

78. Gregory TR (2002) A bird's-eye view of the C-value enigma: genome size, cell size, and metabolic rate in the class aves. Evolution 56: 121-130.

79. Tiersch TR, Wachtel SS (1991) On the evolution of genome size of birds. J Hered 82: 363-368.

80. Szarski H (1970) Changes in the amount of DNA in cell nuclei during vertebrate evolution. Nature 226: 651-652.

81. Andrews CB, Mackenzie SA, Gregory TR (2009) Genome size and wing parameters in passerine birds. Proc Biol Sci 276: 55-61.

82. Kidwell MG (2002) Transposable elements and the evolution of genome size in eukaryotes. Genetica 115: 49-63.

83. Fanning TG (1983) Size and structure of the highly repetitive BAM HI element in mice. Nucleic Acids Res 11: 5073-5091.

84. Jurka J, Klonowski P, Dagman V, Pelton P (1996) CENSOR--a program for identification and elimination of repetitive elements from DNA sequences. Comput Chem 20: 119-121.

85. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59: 307-321.

86. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.

87. R Core Team (2013) R: A Language and Environment for Statistical Computing. Vienna, Austria.

88. Cantrell MA, Grahn RA, Scott L, Wichman HA (2000) Isolation of markers from recently transposed LINE-1 retrotransposons. Biotechniques 29: 1310-1316.

89. An W, Han JS, Wheelan SJ, Davis ES, Coombes CE, et al. (2006) Active retrotransposition by a synthetic L1 element in mice. Proc Natl Acad Sci U S A 103: 18662-18667.

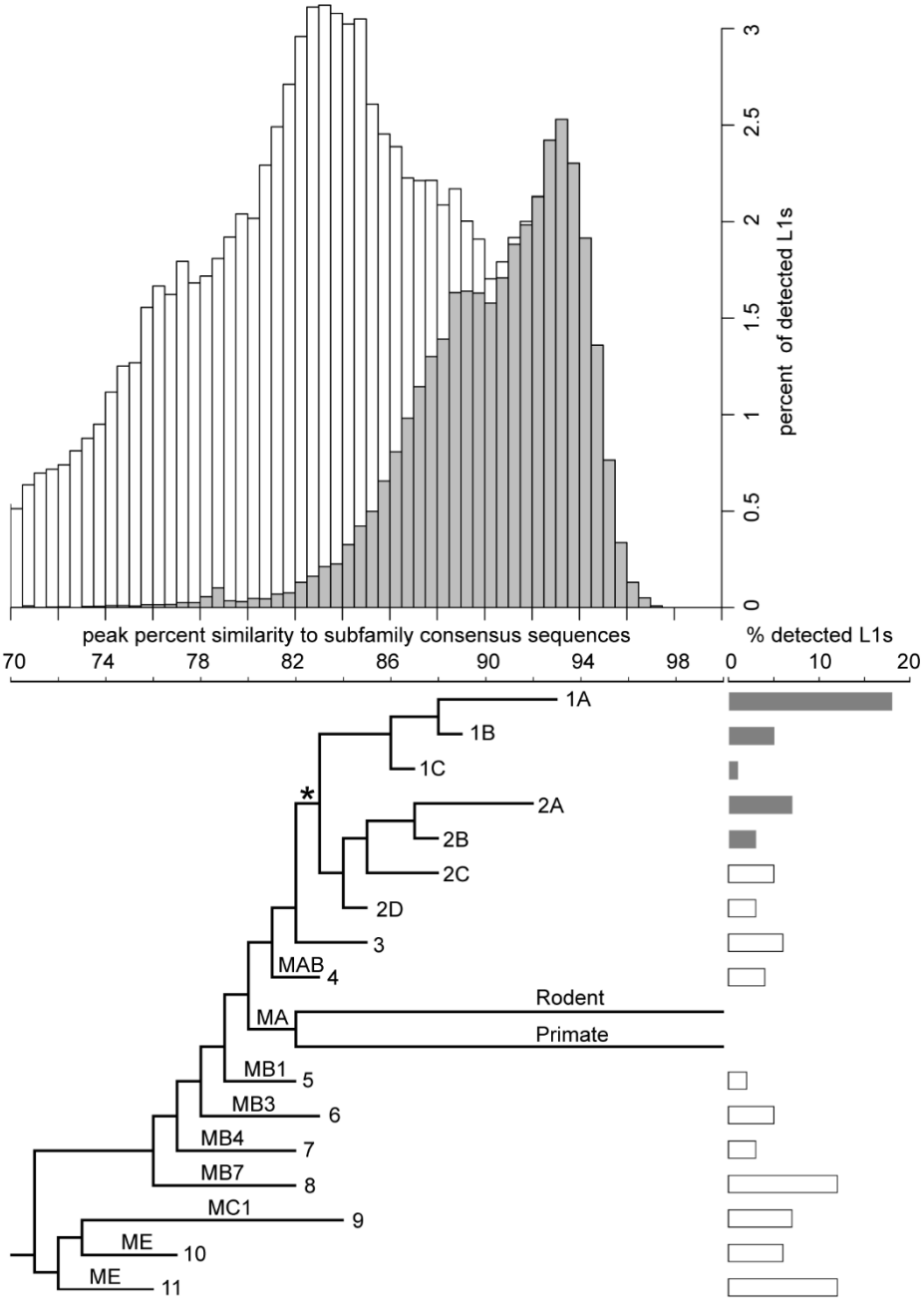


Figure 1. Age distribution and phylogeny of L1s in the megabat genome. The histogram shows the age distribution of megabat L1s as percent of the total 79,978 L1s detected in the megabat genome. Grey bars indicate L1s that are bat-specific. Age of L1s is determined by their percent identity to the corresponding subfamily consensus in 0.5% windows on the horizontal axis – the higher the percent identity, the younger the subfamily. The horizontal axis

is shared with the phylogenetic tree which shows the evolutionary history of L1 families. Taxa names are the numbers assigned to megabat L1 families; names on branches are those given to ancestral mammalian L1 families by Smit *et al.* [57]. Divergence of the human- and rodent-specific L1s and their persistence to present time are indicated by labeled branches. The backbone of the tree is derived from the maximum likelihood tree of all megabat L1 subfamilies and ancestral mammalian L1 families shown in Figure S1, and the branch lengths of the tree were calibrated at the peak of retrotransposition of each family as described in Materials and Methods. * indicates the point after which bat-specific L1s (grey bars) diverged. Lengths of the bars to the right of each terminal branch indicate the percent of all detected L1s contributed by that family.

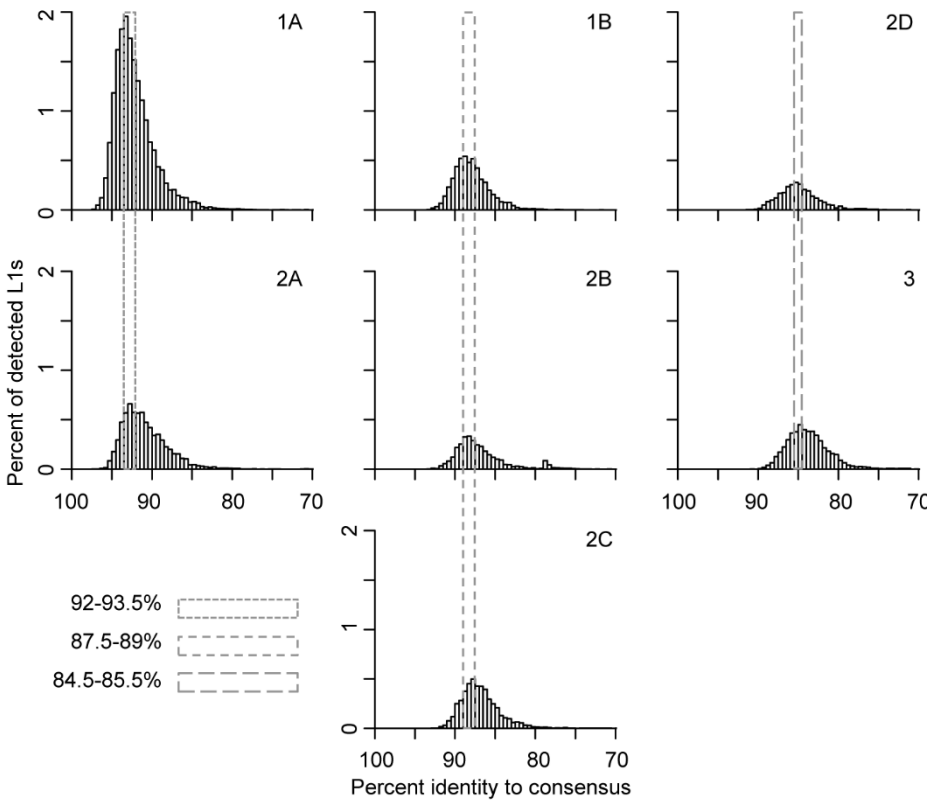


Figure 2. Persistence of concurrently active L1 families. Concurrent L1 families are arranged vertically. Names of families are noted on the top-right corner of each panel. L1 ages are determined by their percent identity to the corresponding subfamily consensus in 0.5% windows – the higher the percent identity, the younger the element. L1 copy numbers are normalized as percent of total detected L1s. The retrotransposition peaks of concurrent families are marked with dashed-line boxes; smaller dashes indicate younger families.

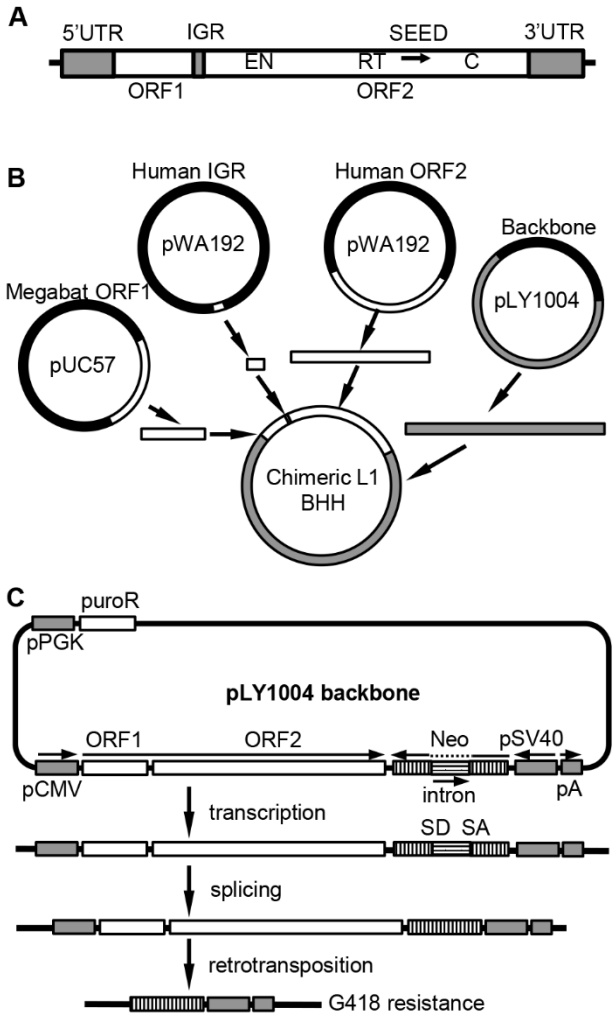


Figure 3. Scheme for assembly of chimeric L1 constructs. (A) Structure of a typical L1. UTR: untranslated region, ORF: open reading frame, IGR: intergenic region, EN: endonuclease motif, RT: reverse transcriptase motif, C: C-terminal domain, SEED: the region amplified by degenerate PCR (arrow) that served as the initial seed for reconstruction of the consensus sequence. (B) Chimeric L1 production. Human and megabat L1 segments were cloned separately into plasmids. L1 segments and the plasmid backbone with compatible overhangs were generated either by PCR or restriction enzyme digestion and joined together by a multi-way ligation. In this example ORF1 and the IGR are from megabat while ORF2 is from human (BBH). All eight combinations were produced in this manner. (C) Retrotransposition rate assay.

The backbone of the constructs, linearized pLY1004, includes the puromycin resistance gene (*puro*R) driven by a constituent promoter (pPGK), and an inverse neomycin resistance gene (*neo*) close to the cloning site for the L1. Puromycin resistance selects for cells that have acquired a L1 construct. Subsequently, neomycin resistance selects for cells that hosted retrotransposition events as follows. Transcription and subsequent retrotransposition of the cloned L1, driven by a pCMV promoter, trigger the splicing between donor (SD) and acceptor (SA) sites, activating the inverse-oriented *neo* cassette which is driven by an SV40 promoter. Thus, a cell will give rise to a colony if it accommodated a retrotransposition event and, thus, excision of the intron in *neo*, allowing it to survive G418 selection.

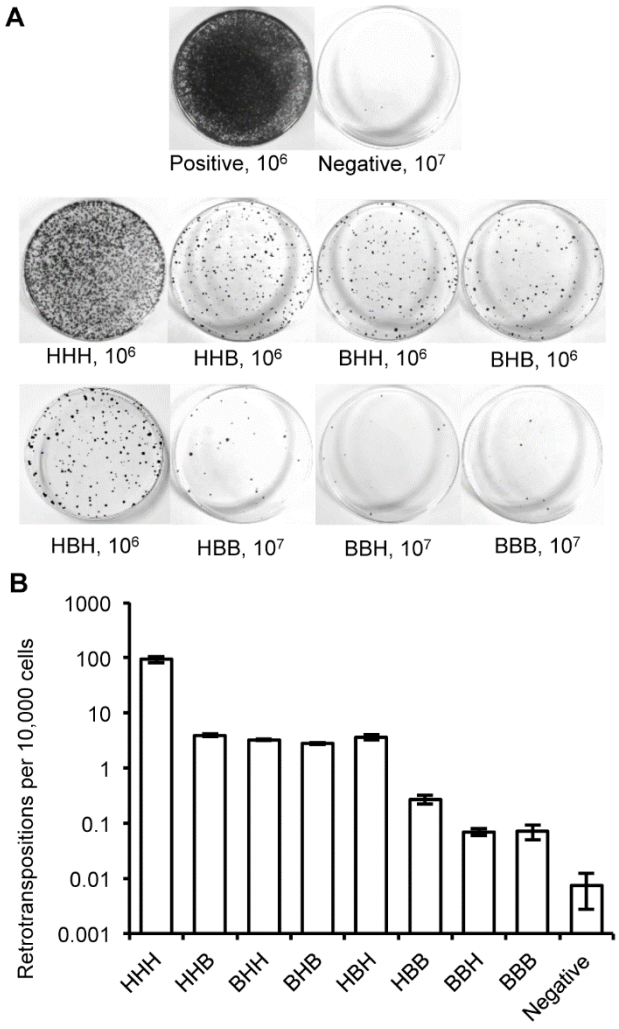


Figure 4. Retrotransposition rate of chimeric L1s. (A) Representative retrotransposition assay plates. Constructs are named with a three letter code based on the origin of their ORF1, IGR and ORF2: **H** for human L1rp; **B** for megabat lineage 1. An independent human L1 construct, pWA192 [66], was used as a positive control and an ORF1 mutant of L1rp [67] that blocks retrotransposition was used as a negative control. The number of cells seeded for G418 selection follows the name; 10-fold more cells were used for the negative control and for constructs with low retrotransposition rates. (B) Comparison of retrotransposition rates (log scale). At least 12 plates were counted for each construct in three independent replicate assays.

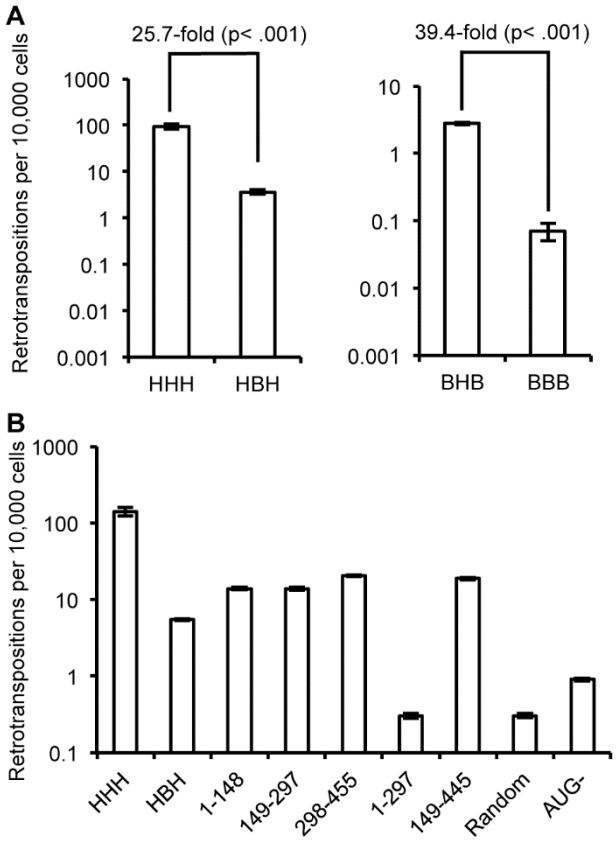


Figure 5. Effect of IGR on retrotransposition rate. (A) Heterologous IGRs: replacing the human L1 IGR with a megabat version reduces the retrotransposition rate ~25.7-fold, while replacing the megabat IGR with a human L1rp IGR increases the retrotransposition ~39.4-fold. (B) Manipulated IGRs were tested in all chimeric L1 backgrounds and the results were qualitatively similar. Data are shown for replacement of the human L1rp IGR (HXH); data for the remaining L1 backgrounds are shown in Figure S3. At least four plates were counted per construct. Numbers below the bars indicate truncation of the megabat IGR. For example, '1-148' indicates a truncated version containing the first third of the IGR (bp 1-148). 'Random' indicates a shuffled version of the megabat IGR of the same length and nucleotide composition. 'AUG-' indicates the megabat IGR with all the AUG start codons (excluding the start at the beginning of ORF2) mutated to AGU.

name / abbreviation	RepBase	latin name	common name	genomic %A	L1 %A ORFs	L1 %A IGR	L1 %A ORFs+IGR	IGR length
family 2A (L1-1_PVa)	x	*Pteropus vampyrus*	large flying fox	30.0	44.1	65.8	44.3	38
family 1A (L1-2_PVa)	x	*Pteropus vampyrus*	large flying fox	29.3	43.4	47.2	43.7	407(445)
L1-Y_CF	x	*Canis lupus familiaris*	dog	29.1	39.4	46.9	39.5	49
L1MAB2_ML	x	*Myotis lucifugus*	little brown bat	28.8	42.6	60.0	42.9	55
L1HS	x	*Homo sapiens*	human	29.2	40.7	44.4	40.7	63
L1-BT	x	*Bos taurus*	cow	29.1	44.4	53.8	44.5	52
L1A_OC	x	*Oryctolagus cuniculus*	rabbit	28.0	40.9	44.4	41.0	81
L1A_Mim	x	*Microcebus murinus*	mouse lemur	29.3	41.4	53.7	41.5	41(86)
L1-2_EC	x	*Equus caballus*	horse	29.2	42.9	47.6	43.2	328
L1-2_Dor	x	*Dipodomys ordii*	kangaroo rat	28.8	40.5	32.5	40.4	40
L1-1B_Cho	x	*Choloepus hoffmanni*	two-toed sloth	30.5	43.1	48.8	43.2	82
L1-1A2_Sar	x	*Sorex araneus*	common shrew	28.6	39.2	21.5	39.3	43
L1-1_Vpa	x	*Vicugna pacos*	alpaca	29.3	43.7	46.2	43.9	442
L1-1_TS	x	*Tarsius syrichta*	Philippine tarsier	30.1	42.1	61.1	42.4	90
L1-1_Tbel	x	*Tupaia belangeri*	tree shrew	29.3	41.0	60.5	41.1	43(55)
L1-1_Str	x	*Ictidomys tridecemlineatus*	13-lined ground squirrel	30.1	42.7	51.2	42.8	82
L1-1_SSc	x	*Sus scrofa*	pig	30.9	43.0	61.8	43.2	68
L1-1_Pca	x	*Procavia capensis*	rock hyrax	29.5	41.0	41.8	41.0	421
L1-1_OP	x	*Ochotona princeps*	American pika	28.4	41.7	47.2	41.7	53
L1-1_MD	x	*Monodelphis domestica*	gray short-tailed opossum	31.0	43.1	40.1	42.7	531
L1-1_LA	x	*Loxodonta africana*	African elephant	29.6	43.4	49.6	43.9	458
L1-1_ET	x	*Echinops telfairi*	lesser hedgehog (tenrec)	28.5	39.1	47.6	39.1	42
L1-1_EE	x	*Erinaceus europaeus*	European hedgehog	29.3	41.0	50.0	41.1	56
L1-1_DN	x	*Dasypus novemcinctus*	armadillo	29.6	43.1	66.7	43.5	75
L1-1_Cpo	x	*Cavia porcellus*	guinea pig	30.1	43.2	66.7	43.3	18(39)
L1-1_Cja	x	*Callithrix jacchus*	marmoset	29.4	41.2	46.0	41.3	63
L1-1_AMe	x	*Ailuropoda melanoleuca*	giant panda	29.2	39.2	37.8	39.0	580
L1_RN	x	*Rattus norvegicus*	brown rat	28.7	41.4	47.5	41.5	59
Fcat		*Felis cattus*	domestic cat	27.4	40.7	48.0	40.6	50
Pham		*Papio hamadryas*	hamadryas baboon	NA	40.8	44.4	40.7	63
Mmul		*Macaca mulatta*	rhesus monkey	29.3	40.8	44.4	40.8	63
Meug		*Macropus eugenii*	tammar wallaby	32.7	41.6	39.3	41.9	565
Opal		*Oryzomys palustris*	marsh rice rat	NA	42.7	52.2	42.7	23
Average				29.4	41.8	49.0	41.9	155(159)

Table 1. (See previous page) L1 consensus sequences used for comparison. Sequences from RepBase are indicated with an X; other sequences were constructed from genomic trace files. Adenosine content is compared between genomic DNA and L1 segments. AT content from the NCBI genome database was divided by two for Genomic %A and does not take into account any strand bias in coding regions. L1 %As were determined from the coding strands. Numbers in parentheses in the IGR length column indicate IGR lengths from alternative ORF2 starts. Average %As and IGR length are in the bottom row.

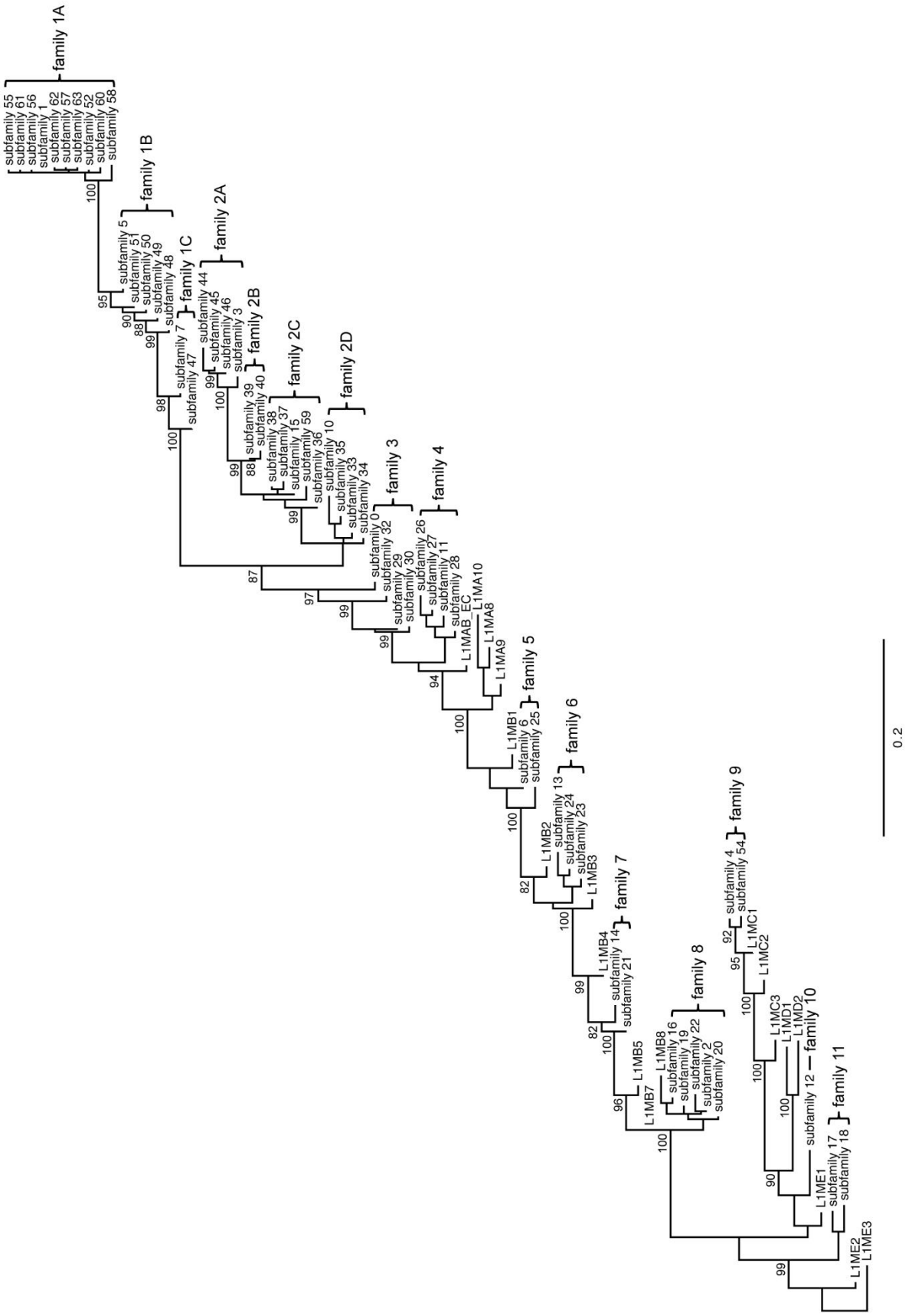

subfamily 55
subfamily 61
subfamily 56
subfamily 1
subfamily 62
subfamily 57
subfamily 63
subfamily 52
subfamily 60
subfamily 58
family 1A
family 1B
subfamily 5
subfamily 51
subfamily 50
subfamily 49
subfamily 48
family 1C
subfamily 7
subfamily 47
family 2A
subfamily 44
subfamily 45
subfamily 46
subfamily 3
family 2B
subfamily 39
subfamily 40
family 2C
subfamily 38
subfamily 37
subfamily 15
subfamily 59
subfamily 36
subfamily 10
family 2D
subfamily 35
subfamily 33
subfamily 34
subfamily 0
subfamily 32
family 3
subfamily 29
subfamily 30
subfamily 26
subfamily 27
subfamily 11
subfamily 28
family 4
L1MAB_EC
L1MA10
L1MA8
L1MA9
L1MB1
subfamily 6
subfamily 25
family 5
subfamily 13
subfamily 24
subfamily 23
L1MB3
family 6
L1MB2
L1MB4
subfamily 14
subfamily 21
family 7
L1MB5
L1MB7
L1MB8
subfamily 16
subfamily 19
subfamily 22
subfamily 2
subfamily 20
family 8
subfamily 4
subfamily 54
family 9
L1MC1
L1MC2
L1MC3
L1MD1
L1MD2
family 10
subfamily 12
L1ME1
subfamily 17
subfamily 18
family 11
L1ME2
L1ME3
0.2

Figure S1. (See previous page) Maximum likelihood tree of the detected megabat L1 subfamilies. Selected ancestral mammalian L1 families, labeled L1MXX, are included to facilitate comparison. The tree was constructed using PhyML [85] with the GTR+I+G model and 100 bootstrap replicates. Bootstrap values >80 are shown. L1 families are designated to the right of the corresponding subfamilies according to Materials and Methods and Table S1.

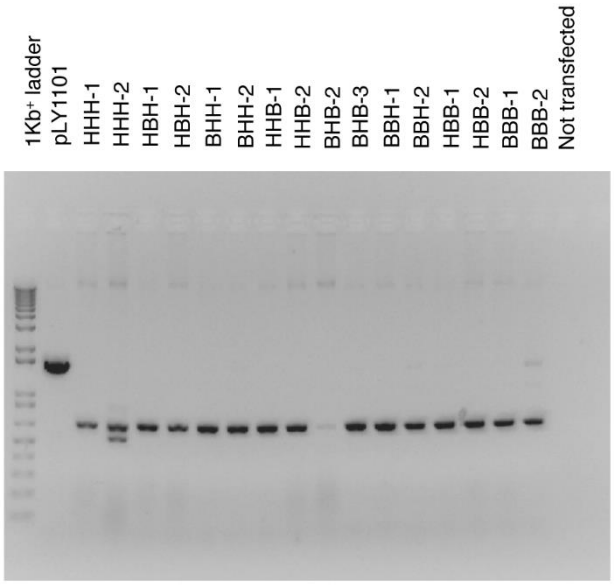


Figure S2. Confirmation of retrotransposition. Retrotransposition was confirmed for each construct by PCR of the *neo* cassette from two surviving colonies. Genomic DNA was extracted and used as template. Genotyping PCR primers were designed to amplify the *neo* cassette so that cells hosting retrotransposition events, and thus the spliced cassette, yield 653 bp PCR products. PCR of positive control construct pLY1101, identical to backbone pLY1004 but with no L1 insertion, yields a 1556 bp product that corresponds to the unspliced *neo* cassette. The 653 bp band was detected from all colonies. Non-specific bands were detected in a few cases; these were not further characterized.

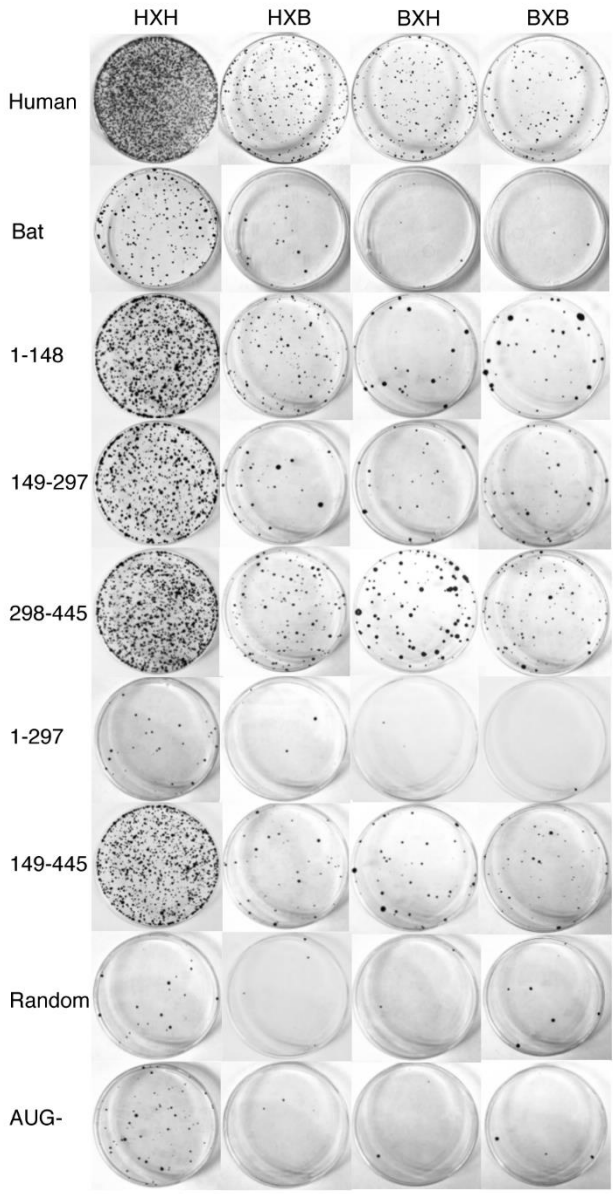


Figure S3. Effect of IGR on retrotransposition rate. Results are shown for all chimeric backgrounds on representative retrotransposition assay plates. Columns represent the various genetic contexts of ORF1/IGR/ORF2; H indicates human L1rp sequence, B indicates reconstructed megabat L1 and X corresponds to the IGR manipulation assayed in each row. Numbers to the left of the rows indicate the truncation of the megabat IGR. For example, '1-148' indicates a truncated version containing the first third of the IGR (bp 1-148). 'Random' indicates a shuffled version of the megabat IGR with the same length and nucleotide

composition. 'AUG-' indicates the megabat IGR with all the AUG start codons (excluding the start at the beginning of ORF2) mutated to AGU.

Family	Subfamilies	Ancestral L1	Fraction (%)	Mean identity (%)	Peak identity (%)
1A	1, 52, 55-58, 60-63	-	18	92	93
1B	5, 48-51	-	5	88	89
1C	7, 47	-	1	86	87
2A	3, 44-46	-	7	91	92
2B	39, 40	-	3	87	88
2C	15, 36-38, 59	-	5	87	88
2D	10, 33-35	-	3	85	85
3	0, 29, 30, 32	-	5	84	85
4	11, 26-28	L1MAB_EC	4	83	83
5	6, 25	L1MB1	2	81	82
6	13, 23, 24	L1MB3	5	82	83
7	14, 21	L1MB4	3	81	82
8	2, 16, 19, 20, 22	L1MB7	12	80	82
9	4, 54	L1MC1	7	82	84
10	12	L1ME	6	76	77
11	17, 18	L1ME	12	74	76

Table S1. Summary of megabat L1 families. Families are based on <3.5% distance among the corresponding subfamilies identified by COSEG and shown in Figure S1. 'Ancestral L1s' are the ancestral mammalian L1 families found in RepBase most closely related to the corresponding megabat families. 'Fraction' indicates the percent of 79,978 total detected megabat L1s in that family. 'Mean identity' refers to the average percent identity of the sequences in each family to their corresponding subfamily consensus, and 'peak identity' refers to the peak of the distribution of the same dataset determined by kernel smoothing as described in Materials and Methods.

Text S1. (See Appendix A) Alignment of L1 ORF1 sequences. Protein alignment of the homologous region of ORF1, amino acids 123-321, bp 1273-1869 of L1rp (GenBank accession number AF148856), including the reconstructed megabat L1 lineage 1 (L1-2_PVa), megabat L1 lineage 2 (L1-1_PVa), 26 RepBase-reconstructed L1 consensuses and four L1s reconstructed by us as described in Materials and Methods. 'Conserved sites' are the conserved amino acid sites among the surveyed species excluding the megabat L1s. L1rp is not shown in the alignment but shares the same nucleotide and amino acid coordinates with L1HS.

Text S2. (See Appendix B) Alignment of L1 ORF2 sequences. Protein alignment of the homologous region of ORF2 spanning the full length L1rp ORF2 (bp 1987-5814, GenBank accession number AF148856), including the reconstructed megabat L1 lineage 1 (L1-2_PVa), megabat L1 lineage 2 (L1-1_PVa), 26 RepBase-reconstructed L1 consensuses and four L1s reconstructed by us as described in Materials and Methods. 'Conserved sites' are the conserved amino acid sites among the surveyed species excluding the megabat L1s. L1rp is not shown in the alignment but shares the same nucleotide and amino acid coordinates with L1HS.

CHAPTER 3

Tracing the History of LINE and SINE Extinction in Sigmodontine Rodents

Lei Yang and Holly A. Wichman

Department of Biological Sciences & Institute for Bioinformatics and Evolutionary Studies,

University of Idaho, Moscow, Idaho, United States of America

Abstract

Background: L1 retrotransposons have co-evolved with their mammalian hosts for the entire history of mammals and currently make up to 20% of a typical mammalian genome. B1 retrotransposons are dependent on L1 for retrotransposition and span the evolutionary history of rodents since their radiation. L1s were found to have lost their activity in a group of South American rodents, the Sigmodontinae, and B1 inactivation preceded the extinction of L1 in the same group. Consequently, a basal group of sigmodontines have active L1s but inactive B1s and a derived clade have both inactive L1s and B1s. It has been suggested that B1s became extinct during a long period of L1 quiescence and that L1s subsequently reemerged in the basal group.

Results: Here we investigate the evolutionary histories of L1 and B1 in the sigmodontine rodents and show that L1 activity continued until after the split of the L1-extinct clade and the basal group. After the split, L1s had a small burst of activity in the former group, followed by extinction. In the basal group activity was initially low but was followed by a dramatic increase in L1 activity. We found the last wave of B1s retrotransposition was large and probably preceded the split between the two rodent clades.

Conclusions: Given that L1s had been steadily retrotransposing during the time corresponding to B1 extinction and that the burst of B1 activity preceding B1 extinction was large, we conclude that B1 extinction was not a result of L1 quiescence. Rather, the burst of B1 activity may have contributed to L1 extinction both by competition with L1 and by putting strong selective pressure on the host to control retrotransposition.

Background

LINEs (Long INterspersed Elements) are autonomous non-LTR (non-long terminal repeat) retrotransposons that move through an RNA intermediate. L1 (LINE-1) is the most successful family of LINEs in eutherian mammals [1] and make up ~20% of a typical mammalian genome [2,3]. A functional full-length L1 is typically 6,000-7,000 bp long and composed of a 5' untranslated region (5'UTR) harboring an RNA polymerase II promoter, two non-overlapping open reading frames (ORFs) known as ORF1 and ORF2 and a 3'UTR followed by a poly-adenosine sequence [4]. The ORF-encoded proteins are strictly required for L1 retrotransposition and are highly *cis*-preferential [5,6]. L1s are adenosine rich (~40%) on their coding strand, which results in biased codon usage compared to host genes [7,8], elongation defects [9], and premature RNA splicing [10]. This A-richness contributes to the inefficiency of L1 retrotransposition and is proposed to regulate the genes in their vicinity [9].

SINEs (Short INterspersed Elements) are relatively short non-autonomous, non-LTR transposable elements. SINEs do not encode proteins for their own retrotransposition and depend on the reverse transcriptase encoded by other transposable elements such as LINEs [11,12]. Although L1s are highly *cis*-preferential [5,6], SINEs can take advantage of L1-encoded proteins for their own retrotransposition [11-13]. Despite their short length, SINEs account for ~10% of a typical mammalian genome [2,3] due to their high copy numbers. Among the ~70 SINE families found in mammals [14], B1 is the most abundant in mouse [3] and possibly most rodent species [15], occupying ~3% of the mouse genome [3]. B1s derived from the RNA component of signal recognition particle 7SL RNA [16,17] and share features with its ancestors – a functional B1 is ~150 bp long and transcribed by RNA polymerase III with the aid

of its two transcription factor binding boxes [18,19]. B1 sequences are rich in CpG sites, which are methylated and thus prone to mutation in mammalian genomes [20], and the elevated mutation rate is pronounced compared to the A-rich L1s. Because the majority of new L1 and B1 inserts are neutrally-evolving pseudogenes, the CpG-rich B1 sequences decay faster than the A-rich L1 sequences.

Both L1 and B1 have long histories of co-evolution with their host genomes. Unlike some transposable elements, there is no known targeted mechanism for L1s excision and thus L1s persist in the genome unless they are removed by non-specific mechanisms. The oldest L1s trace back to the common ancestor of placental mammals and marsupials, ~160 MYA [1,21]. L1s evolve as master lineages so that a single or a few lineages are responsible for the total retrotransposition in a short time window [22-25]. New master elements replace the old ones, eventually dominating retrotransposition, and this replacement process happens recurrently. B1s are younger than L1s, having arisen just before the divergence of the common ancestor of rodents, ~65 MYA [26]. They are rodent-specific, and other SINEs, including B2, B4 and ID elements, are also present in rodent genomes [15]. SINE families have been interacting with L1s for more than 100 MYA, and fossil remnants of extinct SINE families are detectable in well-characterized mammalian genomes [14,27]. Despite being under strict regulation, L1 and B1 make up approximately a quarter of a typical rodent genome [3]. For example, in the mouse genome, there are ~599,000 total copies of L1, responsible for ~19% of the genome [3], of which ~3,000 copies are potentially functional [28], and ~564,000 copies of B1s, responsible for ~3% of the genome [3].

LINEs and SINEs have considerable impact on the mammalian genome, although they were traditionally viewed as “junk DNA”. As LINEs and SINEs, including L1s and B1s,

retrotranspose and recombine, they introduce genome instability [29], cause disease [30] and may occasionally be co-opted by the host to provide host functions, such as their proposed roles in neuro-plasticity [31,32], X chromosome inactivation [33,34], regulatory functions [35,36], DNA break-repair [37] and structural genomic components [38]. Due to the deleterious effects of LINEs and SINEs on the genome, the hosts have evolved many mechanisms to defend against them [39-42]. In addition, the fact that L1 doesn't encode all the enzymatic components required for retrotransposition could result in ongoing competition between L1s and the host for these required host factors [43,44]. Host defense against L1s and B1s are especially strong in germline cells due to germline-specific host defense mechanisms, so that only a limited number of new copies are inserted in each generation [45,46]. L1s and B1s are both epigenetically silenced [47,48] and under the control of small RNAs [49], which are specifically expressed in germline cells.

Since L1 retrotransposition is under strict control by multiple host defenses, it might seem reasonable for the host to occasionally win the evolutionary arms race with L1s, resulting in loss of L1 activity (L1 extinction). L1s do not move horizontally, so such extinctions would affect all derived host species. Two factors are of note here. First, clades with early L1 extinctions could have given rise to large mammalian lineages without L1 activity and be easily detected because of both the number of species affected and the deterioration of the remnant sequences in the genome. Secondly, recent extinctions will be difficult to differentiate from periods of L1 quiescence. To clarify the terms related to loss of L1 activity in this work, we refer to a period of low L1 activity as "quiescence" and complete loss of L1 activity as "extinction". Given the large phylogenetic impact of early extinctions, one might expect L1s to eventually become extinct in most mammalian genomes, and yet L1s have persisted throughout

the entire evolutionary history of their placental mammal and marsupial hosts. Thus, either most L1 extinctions are either recent or rare, or mammalian lineages subject to ancient L1 extinctions do not persist or they give rise to few new species. Understanding the dynamics of L1 extinction will be as important as understanding the dynamics of L1 activity in sorting out the impact of L1s on mammalian genome evolution.

Several cases of L1 extinction have been proposed in the literature [50-56] and two of these are deep extinction events that cover major groups of mammals [50-53]. One of the major L1 extinctions [51-53] occurred in a large group of South American rodents and includes most species in Sigmodontinae. Sigmodontinae is a subfamily of the Cricetidae family, including approximately 377 species classified into 74 genera in nine tribes (Figure 1) [57] and is responsible to 7-8% of the estimated 5,000 mammalian species [58]. Given that B1 retrotransposition is dependent on that of L1, it is expected that B1s should lose their activity simultaneously with L1s. However, the B1 extinction in Sigmodontinae appears to have preceded that of L1s based on samples from 14 genera in five tribes [51-53], where the basal genus *Sigmodon* carries inactive B1 and active L1 and the descendant genera carry both inactive L1 and B1 (Figure 1). It has also been shown that loss of L1 and B1 activity follows the expansion of a group of endogenous retrovirus [59,60].

It was previously hypothesized that the L1 experienced a long-term quiescence as a "stealth driver" [61] before the extinction of L1 and B1, and B1 extinction happened during this period of quiescence [53]. Since B1s are more prone to mutations than the average sequence due to enriched CpG content, Rinehart *et al*. [53] hypothesized that B1 was unable to retrotranspose at a high enough rate during L1 quiescence to replace their active copies, accumulating debilitating mutations more rapidly [20] than L1s. When a more active family of L1 emerged in

the Sigmodontini, B1 was too degenerated to retrotranspose, resulting in B1 extinction even in the presence of high L1 activity.

In this study, we investigate the evolution histories of L1 and B1 spanning the time of their extinctions and the radiation of the extant species in Sigmodontinae (Figure 1). Since the group carrying extinct L1s and B1s (Oryzomyalia, Figure 1) shares a common ancestor, we used the marsh rice rat *Oryzomys palustris* to represent this group, hereafter referred to as the "L1-extinct clade". We used the hispid cotton rat *Sigmodon hispidus* to represent the clade carrying active L1 but inactive B1, hereafter referred to as the "basal group". We used the deer mouse *Peromyscus maniculatus* to represent a closely related clade carrying both active L1 and B1, hereafter referred to as the "outgroup".

Using genome trace files from the species representing the L1-extinct clade and the basal group, we show that the activity of L1 and B1 families that precede the divergence of the clades is comparable in the current genomes of the two groups. L1 families had been steadily replaced before the split of the two groups and maintained activity after the split of the basal group and the L1-extinct clade. Shortly after this split L1 activity ceased in the L1-extinct clade but became highly active in the basal group. B1s, on the other hand, seem to have had a very large increase in activity prior to the split between the L1-extinct clade and the basal group, and there is no strong evidence of activity in the two groups following their divergence. The large burst of B1 activity just prior to extinction suggests that L1 quiescence is unlikely to be responsible for B1 extinction. The last wave of B1 retrotransposition is the largest detectable in the B1 evolutionary history of the group, suggesting that strong competition with L1s or enhanced host defense triggered by radical B1 expansion might have contributed to the extinction of L1.

Results

To investigate the history of L1 retrotransposition in *O. palustris* and *S. hispidus*, we used COSEG [62] to identify closely related L1 groups based on shared, co-segregating sites as described in Methods. We follow the convention of COSEG to designate these groups as *subfamilies*. RepeatMasker [62] was used to initially assign genomic L1 copies to subfamilies, and seven subfamilies with no assigned sequences where removed from further consideration, leaving 47 subfamilies for further analysis.

To examine the activity of L1s in *O. palustris* and *S. hispidus*, we searched the trace files of both genomes separately with the consensus sequences of the abovementioned 47 subfamilies and identified 19,254 sequences in *O. palustris* and 90,526 in *S. hispidus*. The age of each sequence was approximated by its percent divergence from the corresponding subfamily consensus – the higher the percent divergence, the older the sequence. The peak of the distribution was used as an approximation of the age of the subfamily (Table S1). Given the possible changes of evolution rate in the detectable range of L1 evolutionary, a global conversion from percent divergence to time is challenging. However, because of the shared evolutionary history of *O. palustris* and *S. hispidus*, percent divergence is a reasonably good marker to compare the age of L1 subfamilies of the two species.

Subfamily consensus sequences were also subjected to phylogenetic analysis (Figure S1). Subsequently, phylogenetic relationships and sequence similarities between subfamilies were used to assign subfamilies to families with the stipulation that the pairwise distance between subfamilies within a family be no greater than 3.5%. This distance was determined operationally based on the divergences among phylogenetically clustered subfamilies. Clusters of subfamilies

that were similar at the sequence level but differed in age were assigned to different families. This process identified five families specific to *S. hispidus* (S1 to S5), four families shared by *O. palustris* and *S. hispidus* (OS1 to OS4) and two shared by *P. maniculatus*, *O. palustris* and *S. hispidus* (OSP1 and OSP2, Table S1). A distance-based phylogeny reflecting the relationship between L1 families is presented in Figure 2A. Individual sequences were assigned to the families to which their subfamilies belong; the age distribution within a family is based on the distance of each sequence from its subfamily consensus (Figure 3).

As expected, sequences from L1 families shared by *O. palustris* and *S. hispidus* are present in both genomes, and these shared families are fairly synchronized in time and comparable in copy number (Figure 3A). The *Sigmodon*-specific L1 families (Figure 3B, families S1-5) experienced substantial amplification after divergence from the L1-extinct clade, whereas no *Oryzomys*-specific subfamilies were identified by COSEG. The *Sigmodon*-specific subfamilies had a few sequences from the *O. palustris* genome assigned to them, but these assignments appear to be anomalous since the sequences are highly divergent from the subfamily consensus sequences (Table S1). Family OS1, the youngest shared family is of special interest. Family OS1 corresponds to a single L1 subfamily, suggesting that there was little divergence of L1s within the family. It is the last active family prior to the L1 extinction and has ~1.5-fold higher copy numbers per Gbp of sequence in *O. palustris* than in *S. hispidus*. This difference in L1 deposition between *O. palustris* and *S. hispidus* suggests that L1s remained active in the L1-extinct clade after the separation of that group from the basal group. Furthermore, L1s were more active in the lineage leading to Oryzomyalia, in which L1s eventually became extinct, than in the lineage leading to Sigmodontini. A direct comparison of the activity of the L1 families directly preceding this split (OS2), directly following the split (OS1) and at the base of the

Sigmodontini (S5) is presented in Figure 4A. Thus, L1 experienced an expansion (family OS1) in the lineage leading to Oryzomyalia immediately before L1 extinction, while the lineage leading to Sigmodontini experienced a delayed but much larger L1 expansion.

In order to study the B1 dynamics in sigmodontine rodents, we performed the analysis on B1 similar to that done on L1. Because of the short length and CpG-rich nature of B1, we required twice as many sequences to form a subfamily in the second round COSEG as described in Methods. The analysis revealed 30 subfamilies and five families of B1 in both species (Table S2). A distance-based phylogeny reflecting the relationships between B1 families is presented in Figure 2B. One of the families (OS1) is shared by *O. palustris* and *S. hispidus* and the other four (families OSP1-5) are shared by *O. palustris*, *S. hispidus* and *P. maniculatus*. All of the B1 families are shared by *O. palustris* and *S. hispidus* and the representation of these families in both genomes is fairly synchronized in time and comparable in copy number (Figure 5). Since the outgroup, represented by *P. maniculatus*, carries both active L1s and B1s, we know that B1 extinction happened after the split of the outgroup, yet the point at which B1 lost activity in the basal group is to be determined. Here we show that the peak of the most recent B1 family resides at ~11.3% in *O. palustris* and ~10.7% in *S. hispidus* (Table S2). These peaks reside in the same time window as L1 family OS2 (~11.1% in *O. palustris* and ~10.3% in *S. hispidus*, Table S1), suggesting that B1 family OS1 is coincident in time with L1 family OS2. Since L1 family OS2 is the youngest L1 family prior to the separation of the basal group and the L1-extinct clade, the last wave of B1 retrotransposition likely preceded the extinction of L1.

Discussion

In this paper we explore the tempo of L1 and B1 activity surrounding the extinction of both elements that occurred in most species within the rodent subfamily Sigmodontinae. This work is made possible by sequencing methods that allow us to gather large amounts of sequence data and by the availability of a robust species phylogeny for the group (Figure 1). A recent phylogenetic analysis of muroid rodents [63] indicates that the tribe Sigmodontini is basal to the group and sister to the tribe Ichthyomyini. These two tribes are sister to a large, polytomic group (the Oryzomyalia) which includes the remaining five tribes; this group is the result of a rapid radiation of rodents into South America about 5 MYA [64]. Previous work indicated that L1s are extinct in the Oryzomyalia but active in the Sigmodontini, which includes one genus, *Sigmodon*, with 14 species. L1 extinction in the Oryzomyalia has been documented in 14 genera distributed across four tribes spanning this group (Figure 1) [52]. B1s are extinct in Oryzomyalia and Sigmodontini, but the status of both L1s and B1 in the intermediate tribe, Ichthyomyini, is unknown. Thus, L1 extinction from this single event likely affects between 345 and 362 species, or about 7% of all mammalian species.

We reconstructed the shared evolutionary history of L1s and B1s in Sigmodontinae in the period preceding and following extinction of these elements. Our results suggest that L1 master elements have been replaced steadily prior to the extinction of both L1 and B1. This is reflected by the consecutive series of L1 families shared by *O. palustris* and *S. hispidus* after their divergence from *Peromyscus*. B1 elements did not appear to take advantage of every wave of L1 activity, but a wave of L1 retrotransposition (family L1-OS2) corresponds to the B1 retrotransposition peak just prior to B1 extinction (B1-OS1).

There is reasonably strong evidence that L1 extinction occurred after the split between the L1-extinct clade and the basal group. A summary diagram showing the higher level of OS1 activity in *O. palustris* compared to *S. hispidus* (Figure 4A) suggests that the events leading to L1 extinction also happened after the split, rather than that a recovery occurred in *S. hispidus* has been previously suggested [52]. The evolutionary history of B1 in *O. palustris* and *S. hispidus* is comparable. New B1 deposition into the genome was low except for the period directly preceding B1 extinction (Figures 4B and 5). Given the short length of B1s, it is more difficult to identify subfamily clusters, so our estimation of the timing of B1 extinction is weaker than for L1. However, two lines of evidence suggest that the last burst of B1 activity occurred prior to the split between the L1-extinct and basal groups. First, the peak activity of B1OSP1 corresponds most closely to the peak activity of L1OS2, which appears to precede the split of these two rodent clades. Secondly, there is no indication of large differences of activity for any of the B1 subfamilies, as was the case for L1. We suggest that finding the status of both L1s and B1s in the Ichthyomyini lineage might be critical to resolving the timing of B1 extinction.

The most challenging part of studying transposable element evolution history in rodents is the limitation of time windows reflected by detectable sequences. The sequences detectable by RepeatMasker decrease drastically beyond 40% divergence. Since the mutation rate in the rodent lineage is one of the highest in all mammals, 40% divergence in L1 and B1 traces back to the common ancestor of sigmodontine rodents and *P. maniculatus*, while similar studies on bats (Chapter 2) and primates [65,66] trace back to the common ancestor of mammals. Fortunately, *P. maniculatus* carries both active L1s and B1s and is close enough to serve as an outgroup in this study. We were able to identify an L1 family shared by *O. palustris*, *S. hispidus* and *P. maniculatus*, family OSP1.

However, there is an advantage of studying rodents in this type of evolutionary study. Since the mutation rate in the rodent lineage is higher than that of primates and bats due to shorter generation time, evolution in L1 and B1 families reflected by a given span of divergence covers a wider window of time compared to more slowly evolving species. This gives the age distributions of L1s and B1s higher resolution and allows us to discern subtle differences between subfamily ages.

This study is fully bioinformatics-based, but several points are important if one is to consider the underlying molecular events relevant to transpositional bursts and extinctions. L1 and B1 retrotransposition is regulated by a plethora of cellular factors [39-41,49] and reliant on others [43,44]. For evolutionary studies, especially the ones related to L1 and B1 extinction, the historical state of host cellular factors could dramatically change the retrotransposition landscape. Given that not all cellular factors that affect L1 and B1 retrotransposition are known and that coevolution between the elements and these cellular factors is expected, it is not currently possible to fully deduce the molecular events surrounding L1 extinction. However, from an evolutionary perspective, fixed retrotransposition events are recorded in the genome and evolve neutrally as pseudogenes unless excised or too old to be recognized. Therefore, the fossil record of L1s and B1s in the genome is a good temporal record of retrotransposition over time. However, one should keep in mind that estimation of retrotransposition rate based on historical L1 copy numbers could be affected by the excision rate of the host genome. It has been shown that the mammalian genomes have been constantly expelling sequences by various mechanisms and the excision rate varies in different clades of mammals [67]. As old insertions are not actively making new copies, they are exposed to the excision mechanisms for longer time, thus fewer copies of the older families are represented on the histogram. Old L1 and B1 copies also

suffer from the recognition limitation of alignment algorithms. Detectable L1 and B1 copies are drastically reduced beyond 40% divergence.

Methods

O. palustris and *S. hispidus* genomic DNA was sequenced in two separate batches using MiSeq (Illumina, Inc., San Diego, CA) at the IBEST Genomic Resources Core (University of Idaho, Moscow, ID). Paired-end libraries were generated with an insert size of 450-550 bp; ~13 and 14 million total reads were generated for *O. palustris* and *S. hispidus*, respectively. Sequences were processed with SeqyClean (https://bitbucket.org/izhbannikov/seqyclean) and the paired-ends were joined with FLASH [68]. Genome coverage was equivalent to approximately 1.5X; 5.47 Gbp of sequence were generated for *O. palustris* and 6.06 Gbp for *S. hispidus*, but we note that genome size within the sigmodontine rodents varies. Although the genome size of *O. palustris* is not documented to our knowledge, the genome size of sister species in *Oryzomys* suggests that *Sigmodon* genomes are 11-16% larger than those of *Oryzomys* [69].

L1 reconstruction for both species was generated based on partial genomic sequences generated by 454 Pyrosequencing (Roche Applied Science, Penzberg, Germany) at the IBEST Genomic Resources Core, 203 Mbp of sequence for *O. palustris* and 214 Mbp for *S. hispidus*. *P. maniculatus* genome trace files were obtained from NCBI. Reconstruction of the 3' ends of *O. palustris* and *S. hispidus* L1s started with a 575 bp consensus seed in the 3' half of L1 ORF2 generated following Cantrell *et al*. [70]. A bioinformatic pipeline for reconstructing a full length L1 is described by Yang *et al*. (Chapter 2). Briefly, sequences were acquired from the genome trace files based on percent identity. The overhangs of the found sequences allowed the creation

of new seeds at both ends of the L1 fragment and were used to initiate another round of query. In this case, the reconstruction walk was repeated in the 3' direction until the 3' end of ORF2 was reached. Percent identity cutoff was set at 92% for *O. palustris* and higher percent identity (97 to 99%) was used for *S. hispidus* to assure a satisfactory consensus for each walk and the exclusion of older L1 elements. The 3' 300 bp of the reconstructed L1s were then used as the reference sequences for COSEG analysis described below.

B1 sequences from Rinehart *et al*. [53] were used as starting seeds for B1 analysis. The PCR-amplified B1s from *O. palustris* and *S. hispidus* were aligned with Lasergene MegAlign (DNASTAR, Madison, WI) and the consensus sequence (146 bp) was used as the reference sequence for COSEG analysis.

L1 and B1 subfamilies in *O. palustris* and *S. hispidus* were identified and characterized in similar fashion as described below and are summarized in Table S1 and S2.

The reconstructed 300 bp sequences from the 3' end of *O. palustris* and *S. hispidus* L1 ORF2 were each used as the initial L1 query sequences, and the full length B1 consensuses from each species, based on Rinehart *et al*. [53], were used as the initial B1 query sequences. *O. palustris* and *S. hispidus* MiSeq genomic DNA libraries were queried to identify homologous sequences using RepeatMasker [62] with default parameters. Hits from each search were filtered for >90% coverage of the query sequence and subsequently used for the first COSEG [62] (http://www.repeatmasker.org/COSEGDownload.html) run to identify subfamilies base on shared, co-segregating sequence variants. All COSEG runs were conducted under default parameter except as noted. Parameters were set such that at least 250 sequences were required to form an L1 subfamily and 1,000 were required to form a B1 subfamily. In order to identify older subfamilies, the consensus sequences of the subfamilies identified by the first COSEG run were

used as queries to again search the *O. palustris* and *S. hispidus* MiSeq libraries using RepeatMasker. The identified sequences from the second RepeatMasker run were filtered for >90% coverage and extracted. *O. palustris* and *S. hispidus* sequences are combined and a second COSEG run was carried out on the combined sequences. To avoid the possible formation of random subfamilies due to the short length of B1 and the high copy number of the detected sequences, the sequences required to form a subfamily was increased from 1,000 (for the former separate run) to 2,000, whereas this number for L1 remained unchanged at 250. The consensus sequences of the resulting COSEG subfamilies were trimmed to exclude ends that were not common to all subfamilies and the CpG sites were removed and, thus, treated as gaps by RepeatMasker and not counted for the divergence calculation. These modified subfamily consensus sequences were used for a final query of the individual *O. palustris* and *S. hispidus* MiSeq libraries using RepeatMasker. Sequences from this third run were assigned to subfamilies based on percent divergence and this information was stored for further analysis.

P. maniculatus genome trace files were data-mined in a similar fashion through a single round of RepeatMasker and COSEG. The *O. palustris* L1 and B1 sequences described above were used as the initial query seeds for this run. Selected *P. maniculatus* subfamilies were used to demarcate the ages of the subfamilies identified in the *O. palustris* and *S. hispidus* genomes (Figure 3).

Subfamily consensus sequences generated by the second COSEG run of the *O. palustris* and *S. hispidus* libraries were combined and aligned with MegAlign using the Clustal W method for L1 or Clustal V method for B1 and a distance matrix was calculated based on the alignment. Based on the alignment, a maximum likelihood tree was constructed using PhyML [71] with the GTR+I+G model and 100 bootstrap replicates (Figure S1). L1 and B1 sequences were then

assigned to families based on the topology of the tree and a no more than 3.5% within-family pairwise distance from their subfamily consensuses for L1 and 4.4% for B1. Given that the L1 and B1 masters are constantly being replaced during evolution, perfect designation of large families is not possible. The 3.5% threshold was chosen so as to cluster closely related subfamilies without inflating the number of families. Families are named according to their species-specificity and age: "S" indicates *Sigmodon*-specific families, "OS" for families shared by *Sigmodon* and *Oryzomys* and "OSP" for families shared by *Sigmodon*, *Oryzomys* and *Peromyscus*; numbers in family names indicates the age of a family within the family group with "1" being the youngest. Histograms of L1 and B1 age distributions were generated by R [72] histogram function using a window size of 1% (Figure 3). Percent divergence corresponding to retrotransposition peaks of individual families and subfamilies were determined by R using the kernel smoothing function with 0.4% bandwidth (Table S1 and S2).

Author's Contributions

Perceived and designed the experiment: LY and HAW

Performed the bioinformatics analysis: LY

Analyzed the data: LY and HAW

Wrote the manuscript: LY and HAW

Acknowledgements

We thank LuAnn Scott for helpful discussions, editing and proofreading of the manuscript. We thank Dr. Jerzy Jurka at the Genetic Information Research Institute for offering the bioinformatics training. We thank John Brunsfeld and Dr. Celeste Brown on helpful ideas of the L1 reconstruction pipeline design. We thank Drs. Wenfeng An, Celeste Brown and James Foster for helpful comments and discussions. We thank the IBEST Genomics Resources Core for helping us to generate the high-throughput sequencing data used and the IBEST Computer Resources Core for hosting the clusters used for the bioinformatics analysis. This work is funded by National Institute of Health R01-GM38737 to Holly Wichman and National Science Foundation DDIG-1210694 to Holly Wichman and Lei Yang.

References

1. Smit AF (1996) The origin of interspersed repeats in the human genome. Curr Opin Genet Dev 6: 743-748.
2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.
3. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520-562.
4. Furano AV (2000) The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons. Prog Nucleic Acid Res Mol Biol 64: 255-294.

5. Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, et al. (2001) Human L1 retrotransposition: cis preference versus trans complementation. Mol Cell Biol 21: 1429-1439.

6. Kulpa DA, Moran JV (2006) Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. Nat Struct Mol Biol 13: 655-660.

7. Han JS, Boeke JD (2004) A highly active synthetic mammalian retrotransposon. Nature 429: 314-318.

8. An W, Dai L, Niewiadomska AM, Yetil A, O'Donnell KA, et al. (2011) Characterization of a synthetic human LINE-1 retrotransposon ORFeus-Hs. Mob DNA 2: 2.

9. Han JS, Szak ST, Boeke JD (2004) Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature 429: 268-274.

10. Belancio VP, Hedges DJ, Deininger P (2006) LINE-1 RNA splicing and influences on mammalian gene expression. Nucleic Acids Res 34: 1512-1521.

11. Dewannieux M, Esnault C, Heidmann T (2003) LINE-mediated retrotransposition of marked Alu sequences. Nat Genet 35: 41-48.

12. Dewannieux M, Heidmann T (2005) L1-mediated retrotransposition of murine B1 and B2 SINEs recapitulated in cultured cells. J Mol Biol 349: 241-247.

13. Wallace N, Wagstaff BJ, Deininger PL, Roy-Engel AM (2008) LINE-1 ORF1 protein enhances Alu SINE retrotransposition. Gene 419: 1-6.

14. Vassetzky NS, Kramerov DA (2013) SINEBase: a database and tool for SINE analysis. Nucleic Acids Res 41: D83-89.

15. Deininger PL, Tiedge H, Kim J, Brosius J (1996) Evolution, expression, and possible function of a master gene for amplification of an interspersed repeated DNA family in rodents. Prog Nucleic Acid Res Mol Biol 52: 67-88.

16. Weiner AM (1980) An abundant cytoplasmic 7S RNA is complementary to the dominant interspersed middle repetitive DNA sequence family in the human genome. Cell 22: 209-218.

17. Ullu E, Tschudi C (1984) Alu sequences are processed 7SL RNA genes. Nature 312: 171-172.

18. Geiduschek EP, Kassavetis GA (2001) The RNA polymerase III transcription apparatus. J Mol Biol 310: 1-26.

19. Schramm L, Hernandez N (2002) Recruitment of RNA polymerase III to its target promoters. Genes Dev 16: 2593-2620.

20. Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res 8: 1499-1504.

21. Luo ZX, Yuan CX, Meng QJ, Ji Q (2011) A Jurassic eutherian mammal and divergence of marsupials and placentals. Nature 476: 442-445.

22. Casavant NC, Hardies SC (1994) The dynamics of murine LINE-1 subfamily amplification. J Mol Biol 241: 390-397.

23. Pascale E, Liu C, Valle E, Usdin K, Furano AV (1993) The evolution of long interspersed repeated DNA (L1, LINE 1) as revealed by the analysis of an ancient rodent L1 DNA family. J Mol Evol 36: 9-20.

24. Adey NB, Schichman SA, Graham DK, Peterson SN, Edgell MH, et al. (1994) Rodent L1 evolution has been driven by a single dominant lineage that has repeatedly acquired new transcriptional regulatory sequences. Mol Biol Evol 11: 778-789.

25. Clough JE, Foster JA, Barnett M, Wichman HA (1996) Computer simulation of transposable element evolution: random template and strict master models. J Mol Evol 42: 52-58.

26. Kramerov DA, Vassetzky NS (2005) Short retroposons in eukaryotic genomes. Int Rev Cytol 247: 165-221.

27. Ogiwara I, Miya M, Ohshima K, Okada N (1999) Retropositional parasitism of SINEs on LINEs: identification of SINEs and LINEs in elasmobranchs. Mol Biol Evol 16: 1238-1250.

28. Goodier JL, Ostertag EM, Du K, Kazazian HH, Jr. (2001) A novel active L1 retrotransposon subfamily in the mouse. Genome Res 11: 1677-1685.

29. Hedges DJ, Deininger PL (2007) Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. Mutat Res 616: 46-59.

30. Belancio VP, Hedges DJ, Deininger P (2008) Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health. Genome Res 18: 343-358.

31. Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, et al. (2005) Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature 435: 903-910.

32. Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, et al. (2009) L1 retrotransposition in human neural progenitor cells. Nature 460: 1127-1131.

33. Chow JC, Ciaudo C, Fazzari MJ, Mise N, Servant N, et al. (2010) LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. Cell 141: 956-969.

34. Cantrell MA, Carstens BC, Wichman HA (2009) X chromosome inactivation and Xist evolution in a rodent lacking LINE-1 activity. PLoS ONE 4: e6252.

35. Sasaki T, Nishihara H, Hirakawa M, Fujimura K, Tanaka M, et al. (2008) Possible involvement of SINEs in mammalian-specific brain formation. Proc Natl Acad Sci U S A 105: 4220-4225.

36. Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, et al. (2010) Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet 42: 631-634.

37. Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD, et al. (2002) DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat Genet 31: 159-165.

38. Carbone L, Harris RA, Mootnick AR, Milosavljevic A, Martin DI, et al. (2012) Centromere remodeling in Hoolock leuconedys (Hylobatidae) by a new transposable element unique to the gibbons. Genome Biol Evol 4: 648-658.

39. Wissing S, Montano M, Garcia-Perez JL, Moran JV, Greene WC (2011) Endogenous APOBEC3B restricts LINE-1 retrotransposition in transformed cells and human embryonic stem cells. J Biol Chem 286: 36427-36437.

40. Suzuki J, Yamaguchi K, Kajikawa M, Ichiyanagi K, Adachi N, et al. (2009) Genetic evidence that the non-homologous end-joining repair pathway is involved in LINE retrotransposition. PLoS Genet 5: e1000461.

41. Gasior SL, Roy-Engel AM, Deininger PL (2008) ERCC1/XPF limits L1 retrotransposition. DNA Repair (Amst) 7: 983-989.

42. Goodier JL, Cheung LE, Kazazian HH, Jr. (2012) MOV10 RNA helicase is a potent inhibitor of retrotransposition in cells. PLoS Genet 8: e1002941.

43. Dai L, Taylor MS, O'Donnell KA, Boeke JD (2012) Poly(A) binding protein C1 is essential for efficient L1 retrotransposition and affects L1 RNP formation. Mol Cell Biol 32: 4323-4336.

44. Taylor MS, Lacava J, Mita P, Molloy KR, Huang CR, et al. (2013) Affinity Proteomics Reveals Human Host Factors Implicated in Discrete Stages of LINE-1 Retrotransposition. Cell 155: 1034-1048.

45. Cordaux R, Hedges DJ, Herke SW, Batzer MA (2006) Estimating the retrotransposition rate of human Alu elements. Gene 373: 134-137.

46. Huang CR, Schneider AM, Lu Y, Niranjan T, Shen P, et al. (2010) Mobile interspersed repeats are major structural variants in the human genome. Cell 141: 1171-1182.

47. Yoder JA, Walsh CP, Bestor TH (1997) Cytosine methylation and the ecology of intragenomic parasites. Trends Genet 13: 335-340.

48. Bourc'his D, Bestor TH (2004) Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature 431: 96-99.

49. Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ (2007) Developmentally regulated piRNA clusters implicate MILI in transposon control. Science 316: 744-747.

50. Cantrell MA, Scott L, Brown CJ, Martinez AR, Wichman HA (2008) Loss of LINE-1 activity in the megabats. Genetics 178: 393-404.

51. Casavant NC, Scott L, Cantrell MA, Wiggins LE, Baker RJ, et al. (2000) The end of the LINE?: lack of recent L1 activity in a group of South American rodents. Genetics 154: 1809-1817.

52. Grahn RA, Rinehart TA, Cantrell MA, Wichman HA (2005) Extinction of LINE-1 activity coincident with a major mammalian radiation in rodents. Cytogenet Genome Res 110: 407-415.

53. Rinehart TA, Grahn RA, Wichman HA (2005) SINE extinction preceded LINE extinction in sigmodontine rodents: implications for retrotranspositional dynamics and mechanisms. Cytogenet Genome Res 110: 416-425.

54. Platt RN, 2nd, Ray DA (2012) A non-LTR retroelement extinction in Spermophilus tridecemlineatus. Gene 500: 47-53.

55. Boissinot S, Roos C, Furano AV (2004) Different rates of LINE-1 (L1) retrotransposon amplification and evolution in New World monkeys. J Mol Evol 58: 122-130.

56. Waters PD, Dobigny G, Pardini AT, Robinson TJ (2004) LINE-1 distribution in Afrotheria and Xenarthra: implications for understanding the evolution of LINE-1 in eutherian genomes. Chromosoma 113: 137-144.

57. Smith MF, Patton JL (1999) Phylogenetic relationships and the radiation of sigmodontine rodents in South America: evidence from cytochrome b. Journal of mammalian evolution 6: 89-128.

58. Wilson DE (2005) Mammal Species of the World: A Taxonomic and Geographic Reference: JHU Press.

59. Cantrell MA, Ederer MM, Erickson IK, Swier VJ, Baker RJ, et al. (2005) MysTR: an endogenous retrovirus family in mammals that is undergoing recent amplifications to unprecedented copy numbers. J Virol 79: 14698-14707.

60. Erickson IK, Cantrell MA, Scott L, Wichman HA (2011) Retrofitting the genome: L1 extinction follows endogenous retroviral expansion in a group of muroid rodents. J Virol 85: 12315-12323.

61. Cordaux R, Batzer MA (2009) The impact of retrotransposons on human genome evolution. Nat Rev Genet 10: 691-703.

62. Smit A, Hubley R (1996-2010) RepeatMasker Open-3.0.

63. Schenk JJ, Rowe KC, Steppan SJ (2013) Ecological opportunity and incumbency in the diversification of repeated continental colonizations by muroid rodents. Syst Biol 62: 837-864.

64. Marshall LG, Butler RF, Drake RE, Curtis GH, Tedford RH (1979) Calibration of the great american interchange. Science 204: 272-279.

65. Smit AF, Toth G, Riggs AD, Jurka J (1995) Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J Mol Biol 246: 401-417.

66. Khan H, Smit A, Boissinot S (2006) Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res 16: 78-87.

67. Gregory TR (2004) Insertion-deletion biases and the evolution of genome size. Gene 324: 15-34.

68. Magoc T, Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27: 2957-2963.

69. Gregory TR (2014) Animal Genome Size Database.

70. Cantrell MA, Grahn RA, Scott L, Wichman HA (2000) Isolation of markers from recently transposed LINE-1 retrotransposons. Biotechniques 29: 1310-1316.

71. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59: 307-321.

72. R Core Team (2013) R: A Language and Environment for Statistical Computing. Vienna, Austria.

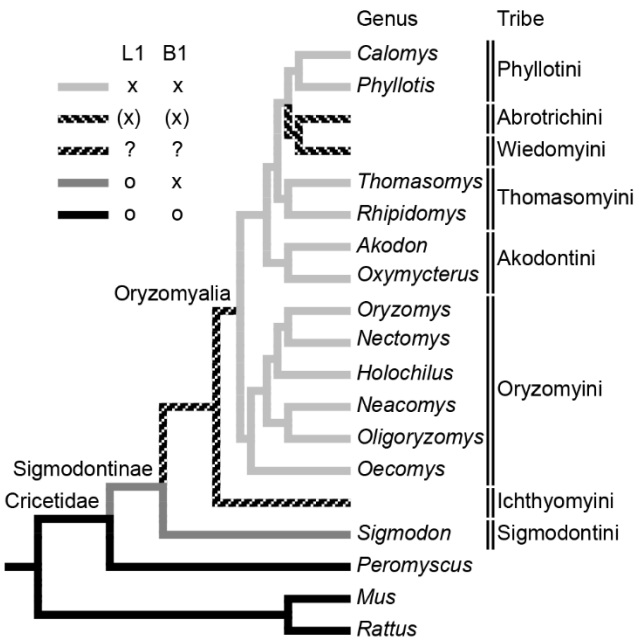


Figure 1. The phylogeny of the sigmodontine rodents. The tree is based on Schenck *et al.* [63]. Taxa are the sampled genera in the group; tribes are indicated on the right side of the taxa. Eight of the nine tribes and 12 of the 14 sampled genera by Rinehart *et al.* [53] are shown. L1 and B1 activity of each taxon is demonstrated by gray scale and: black indicates active L1 and B1, dark gray indicates active L1 and inactive B1 and forward hatching indicates the taxa where L1 activity cannot be inferred and back hatching indicates the taxa where L1 can be inferred to be active. "o" corresponds to active L1 and B1 and "x" corresponds to inactive L1 and B1.

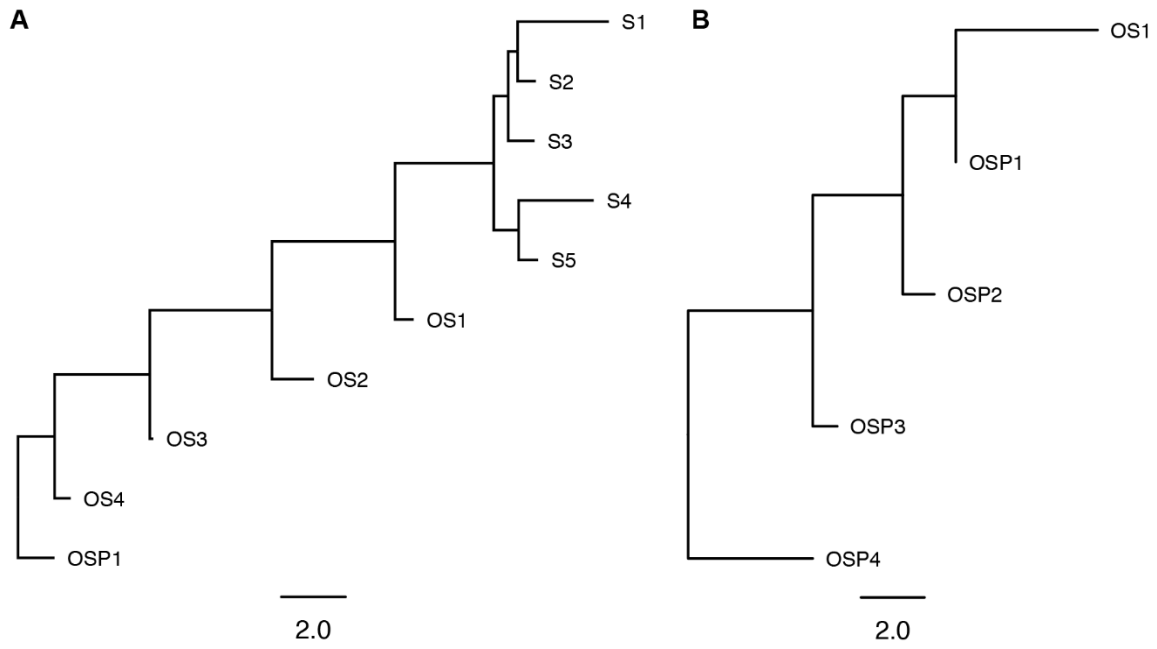


Figure 2. The phylogenies of L1 and B1 families. Panel A shows the L1 tree and B shows the B1 tree. To reflect ages of the families, the trees were based on the distance between families. The distance between any two families was calculated by taking the average pairwise distance of the consensus sequences of subfamilies that belong to each family.

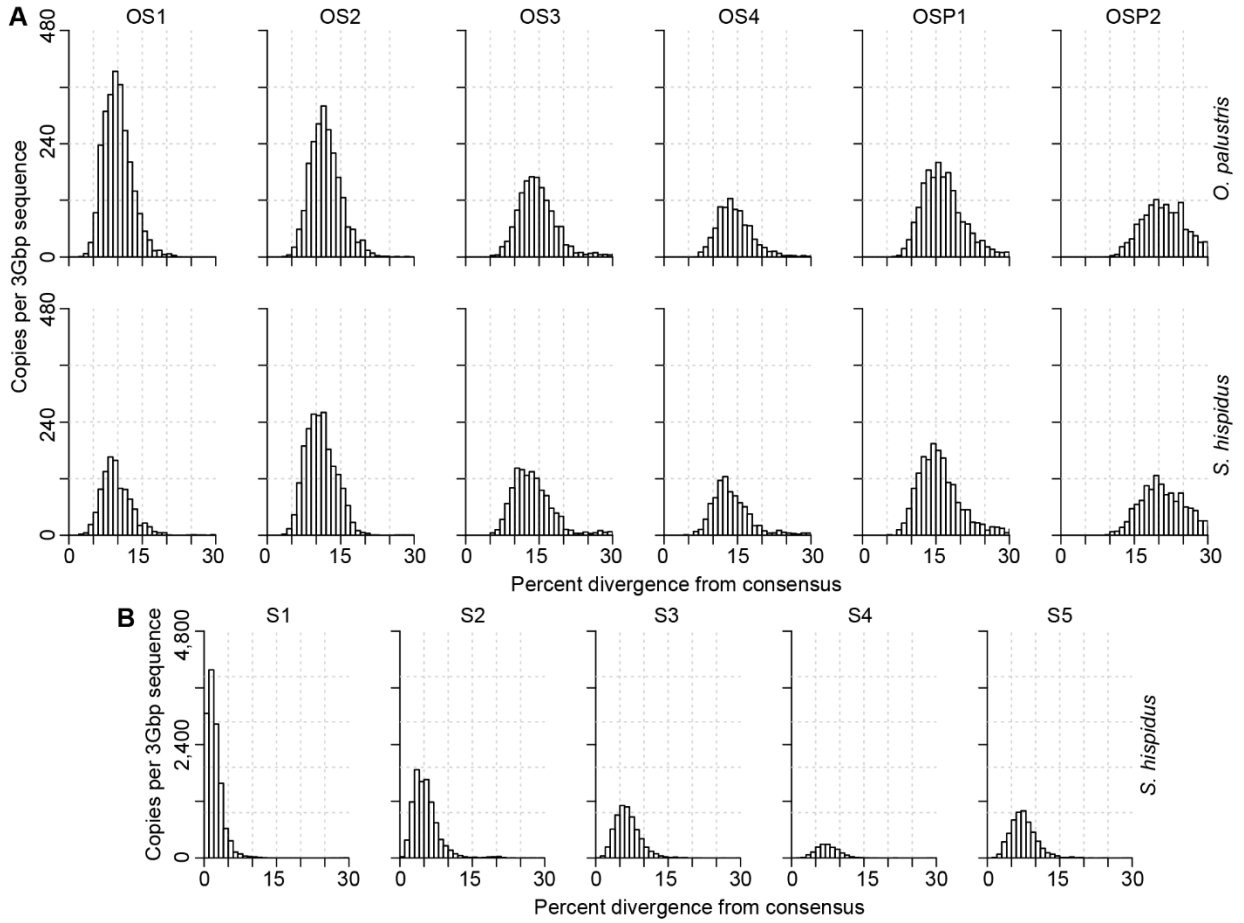


Figure 3. The age distribution of L1 families. L1 families in each row are arranged in chronological order with the youngest families on the left. The species analyzed in each row is indicated at the right. Names of families are noted on the top of each panel. L1 copy number is plotted by percent divergence from the corresponding subfamily consensus in 1% windows. The age of each family is approximated by the peak of the distribution. L1 copy numbers are normalized as copies per three Gbp of MiSeq sequence which approximates the copy number per haploid genome. Panel A shows the shared families and panel B shows the *Sigmodon*-specific families.

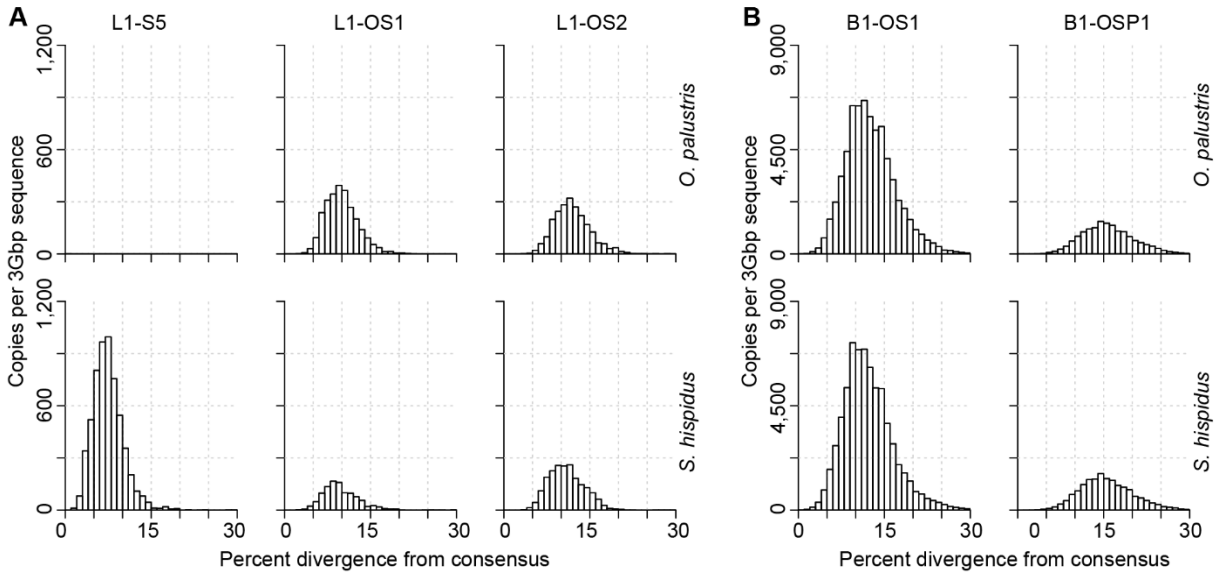


Figure 4. Comparison of L1 and B1 families spanning their extinction. Panel A presents L1 families S5, OS1 and OS2 arranged in a chronological order with the youngest families on the left, and panel B presents B1 families OS1 and OSP1. The species analyzed in each row is indicated at the right. Names of families are noted at the top. Copy number of L1 OS2 is comparable in *O. palustris* and *S. hispidus*, but more OS1 copies were detected in *O. palustris*. Subsequently, there was a new wave of L1 retrotransposition in *S. hispidus* (family S5), but no younger waves of L1 retrotransposition events were identified in *O. palustris*. B1 OS1 corresponds to L1 OS2 in terms of age.

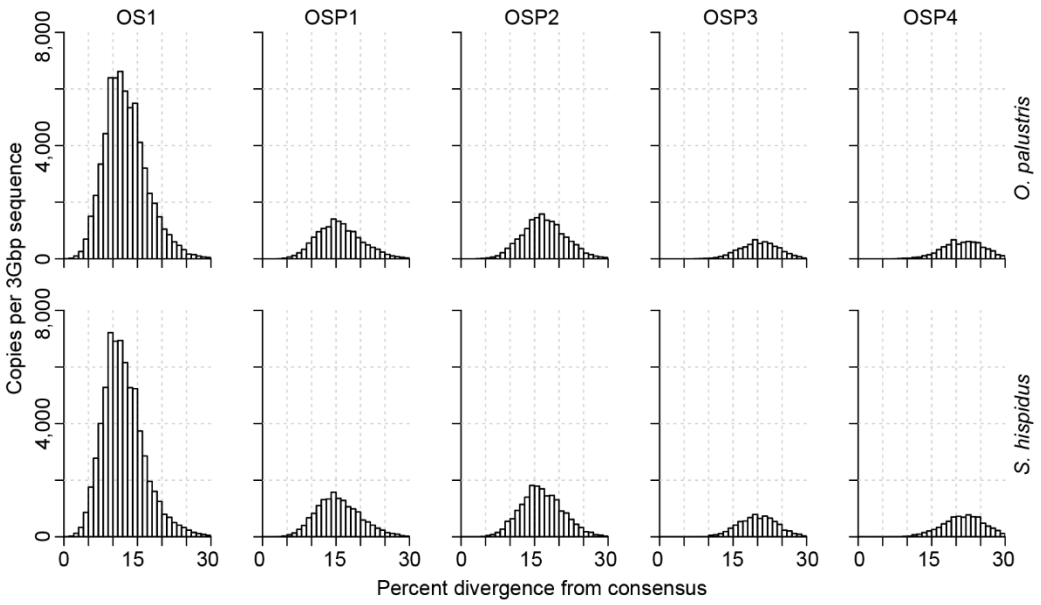


Figure 5. The age distribution of B1 families. B1 families in each row are arranged in chronological order with the youngest families on the left. The species analyzed in each row is indicated at the right. Names of families are noted on the top of each panel. B1 copy number is plotted by percent divergence from the corresponding subfamily consensus in 1% windows. The age of each family is approximated by the peak of the distribution. B1 copy numbers are normalized as copies per three Gbp of MiSeq sequence which approximates the copy number per haploid genome.

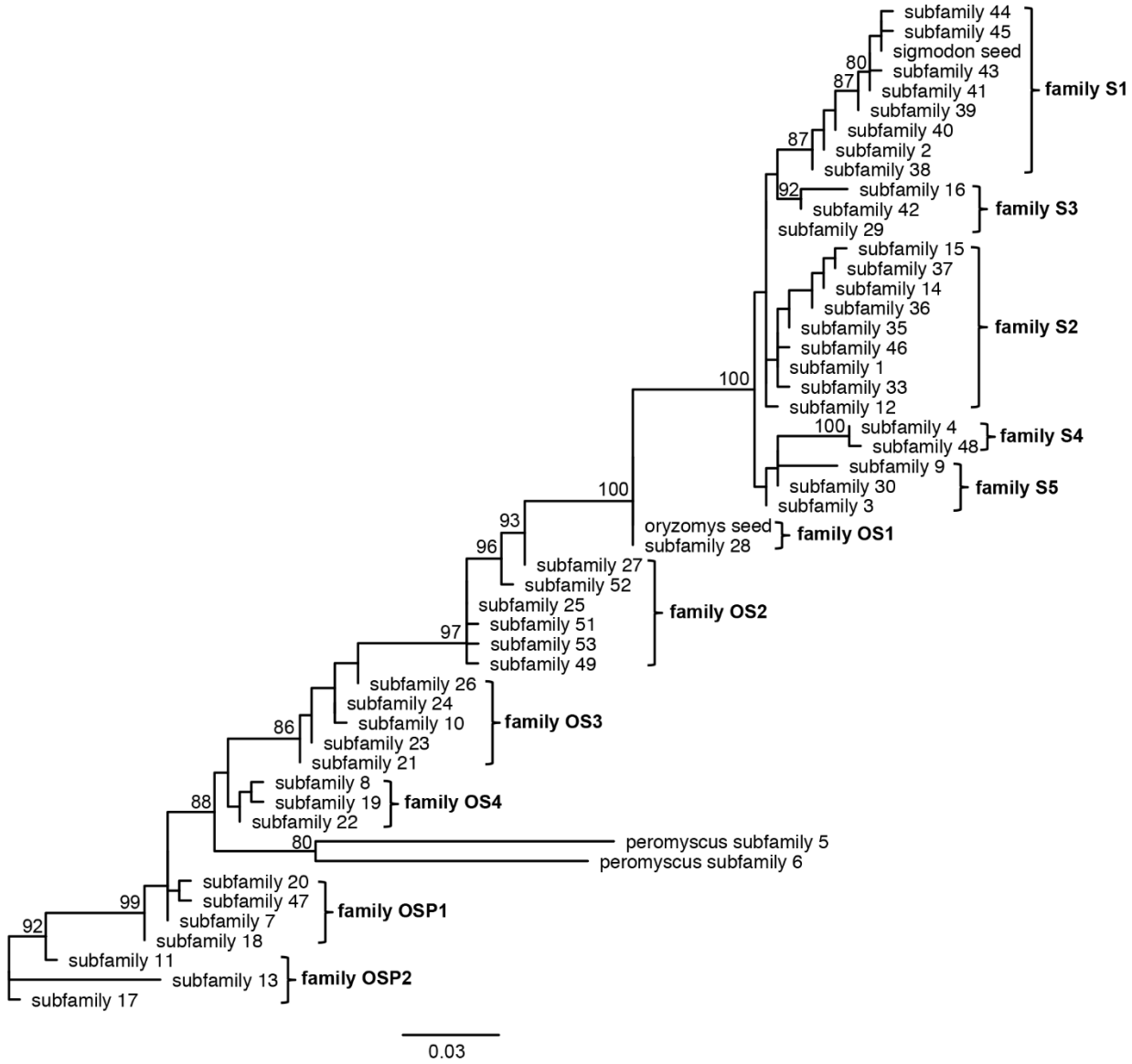


Figure S1. The maximum likelihood phylogeny of detected L1 subfamilies. Reconstructed *O. palustris* and *S. hispidus* L1s, labeled 'seed', and *P. maniculatus* subfamilies 5 and 6 are included as markers. The tree was reconstructed using PhyML [71] with the GTR+I+G model and 100 bootstrap replicates. Bootstrap values > 80% are shown.

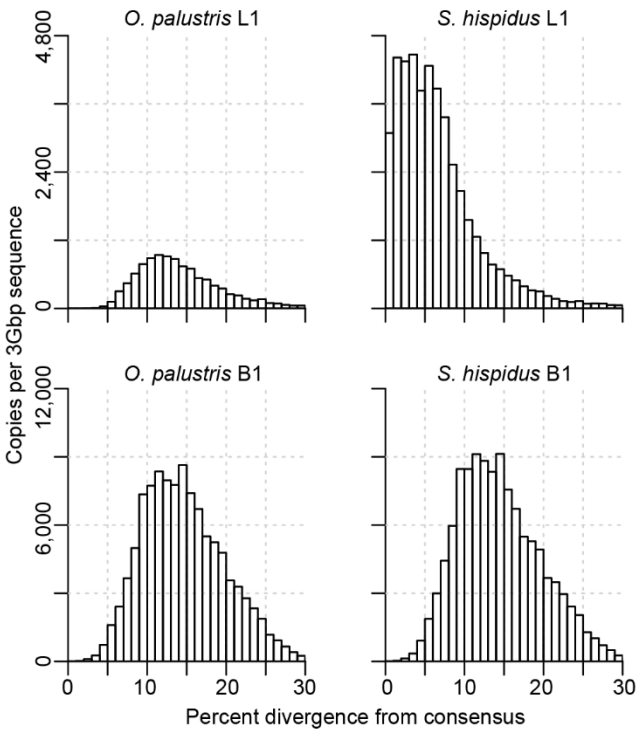


Figure S2. The age distribution of all detected L1 and B1 sequences. Ages of sequences are approximated by their percent divergence from the corresponding subfamily consensus sequences and plotted in 1% windows. Species and retrotransposon names are indicated at the top of each panel.

subfamily	ory copy	sig copy	ory peak	sig peak	family	ory copy	sig copy	ory % total	sig % total	ory fam peak	sig fam peak
44	0	1307	NA	0.5	S1	11	12770	0%	28%	NA	0.5
45	0	576	NA	0.5							
43	1	1592	0.9	0.5							
41	1	850	0.5	0.5							
39	2	3156	1.7	0.9							
40	2	560	4.4	2.1							
2	3	3743	17.4	2.5							
38	3	987	13.9	2.9							
16	1	470	34.4	6.4	S3	4	6437	0%	14%	NA	4.4
42	1	532	13.5	6.4							
29	3	5436	15.9	6.0							
15	1	345	22.6	2.9	S2	24	9770	0%	22%	NA	6.0
37	1	396	26.5	3.3							
14	2	138	15.5	3.3							
36	1	843	26.1	3.7							
35	3	2621	18.6	4.0							
46	4	1776	14.3	5.2							
1	3	1747	19.8	4.8							
33	1	621	23.7	5.6							
12	7	1284	18.2	5.6							
4	2	1193	17.4	7.2	S4	2	1636	0%	4%	NA	7.2
48	1	443	27.7	6.8							
9	5	1153	26.1	7.6	S5	9	5874	0%	13%	NA	6.8
30	4	993	14.7	6.8							
3	10	3728	20.6	6.8							
28	2619	1052	9.6	8.8	OS1	2619	1052	25%	2%	9.6	8.8
27	424	356	11.1	9.2	OS2	2239	2030	21%	5%	11.1	10.3
52	365	364	11.5	10.0							
25	375	350	11.9	10.7							
51	384	374	10.3	10.3							
53	207	244	11.9	14.3							
49	485	343	11.5	10.7							
26	291	233	14.3	11.1	OS3	1474	1287	14%	3%	13.9	12.3
24	283	251	13.5	13.9							
10	262	260	14.3	11.5							
23	127	113	14.3	13.5							
21	512	430	13.5	12.7							
8	180	171	13.5	12.3	OS4	944	901	9%	2%	13.5	12.3
19	438	429	13.5	12.7							
22	325	300	13.5	12.3							
20	149	118	14.7	13.9	OSP1	1872	1753	18%	4%	15.5	14.7
47	354	357	13.5	13.1							
7	387	353	13.9	13.9							
18	982	925	16.7	15.1							
11	34	43	23.3	24.5	OSP2	1352	1304	13%	3%	20.2	19.4
13	1202	1168	20.2	19.8							
17	116	93	19.8	19.0							
Total	10560	44815									

Table S1. The statistics and designation of L1 subfamilies and families. "Ory" stands for *O. palustris* and "Sig" stands for *S. hispidus*. "Peak" indicates the peak of the L1 divergence distribution of the subfamily or family identified by kernel smoothing. Copy numbers are normalized as copies per three Gbp of MiSeq sequence used for the search, which approximates the copy number per haploid genome. Designation of families is only shown after the first subfamily that belongs to it; all subsequent subfamilies belong to this family until the

demarcation of the next family. Characters in family names: "S" represents *S. hispidus*-specific, "OS" for shared by *O. palustris* and *S. hispidus* and "OSP" for shared by *O. palustris*, *S. hispidus* and *P. maniculatus*. Numbers in the family names reflect their ages among the family group with "1" being the youngest. Copy numbers of families are rounded sums of subfamily copy numbers per three Gbp of sequences and, thus, are occasionally off by one.

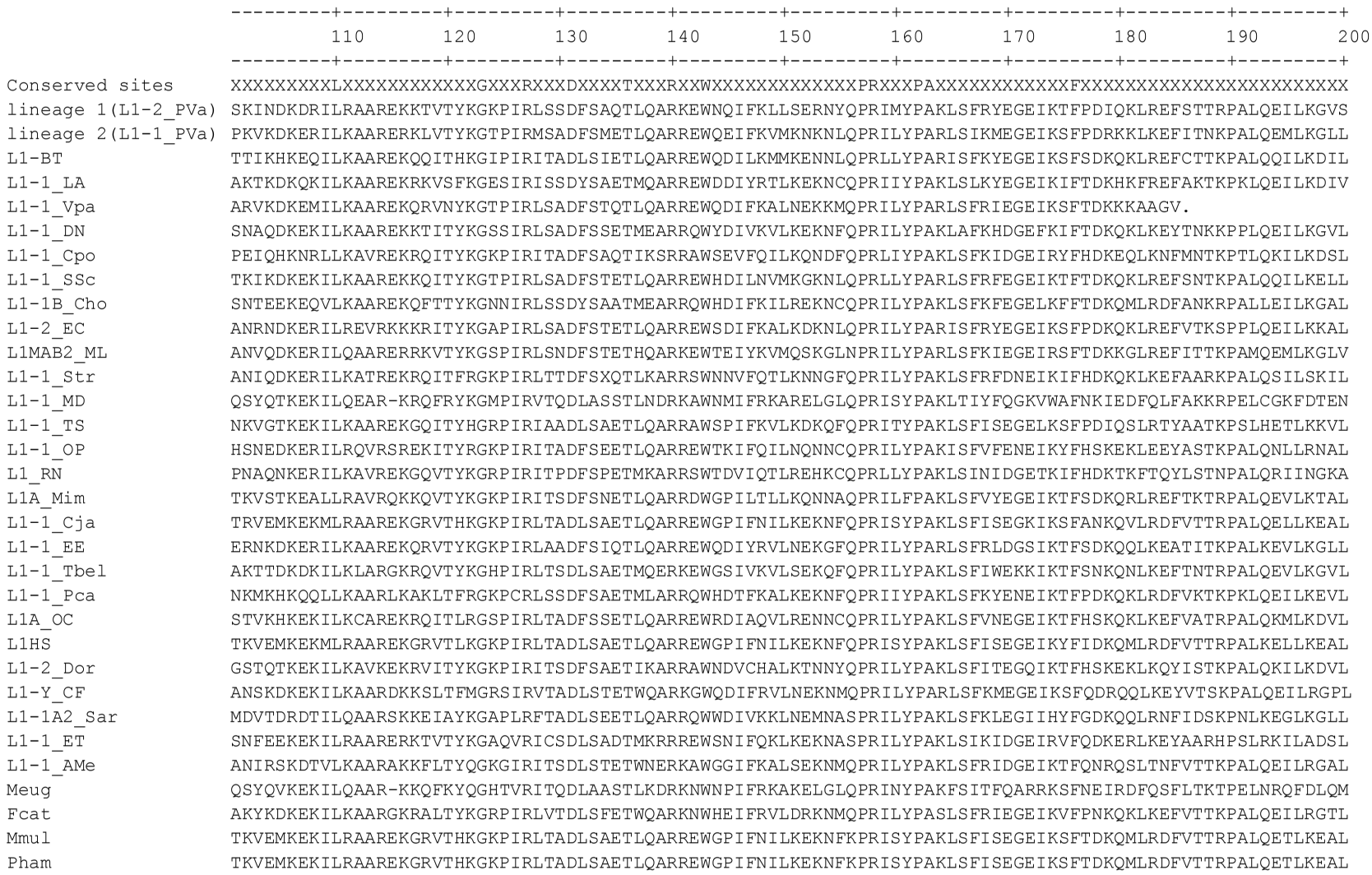
subfamily	ory copy	sig copy	ory peak	sig peak	family	ory copy	sig copy	ory % total	sig % total	ory fam peak	sig fam peak
3	6501	6781	10.3	10.3	OS1	65656	67732	60.4%	57.6%	11.1	10.7
16	4482	4808	10.3	10.7							
32	3036	3075	10.7	10.7							
5	3822	4445	11.1	10.3							
12	2068	2228	11.1	10.3							
2	2136	2324	11.1	11.1							
33	2566	2602	11.1	11.1							
20	3664	2487	11.5	10.0							
18	15471	16688	11.5	10.7							
35	5688	6243	11.5	10.7							
21	2725	2770	11.9	10.7							
25	3113	2930	11.9	11.1							
24	2207	2333	11.9	11.5							
28	3662	3443	11.9	11.5							
7	2539	2352	12.3	13.5							
8	1973	2221	12.7	11.9							
26	2353	2582	14.7	14.3	OSP1	14361	15841	13.2%	13.5%	15.1	14.7
14	2090	2047	14.7	13.5							
1	4684	5122	14.7	14.3							
11	1967	2285	18.2	19.8							
13	3268	3804	14.7	14.3							
9	3593	4072	15.1	14.7	OSP2	15950	18431	14.7%	15.7%	16.3	15.5
4	6435	7568	16.3	15.5							
36	2667	2990	17.0	17.0							
27	1516	1773	16.7	16.3							
31	1739	2028	17.0	16.7							
17	6766	8107	21.8	22.2	OSP4	6766	8107	6.2%	6.9%	21.8	22.2
19	2188	2722	20.2	20.2	OSP3	5978	7456	5.5%	6.3%	20.6	20.2
6	969	1201	20.2	18.6							
10	2821	3533	20.6	20.2							
Total	108711	117567									

Table S2. The statistics and designation of B1 subfamilies and families. "Ory" stands for *O. palustris* and "Sig" stands for *S. hispidus*. "Peak" indicates the peak of the B1 divergence distribution of the subfamily or family identified by kernel smoothing. Copy numbers are normalized by per three Gbp of MiSeq sequence used for the search. Designation of families is only shown after the first subfamily that belongs to it; all subsequent subfamilies belong to this family until the demarcation of the next family. Characters in family names: "OS" represents families shared by *O. palustris* and *S. hispidus* and "OSP" for families shared by *O. palustris*, *S. hispidus* and *P. maniculatus*. Numbers in the family names reflect their ages within the family group with "1" being the youngest. Copy numbers of families are rounded sums of subfamily copy numbers per three Gbp of sequences.

APPENDIX A

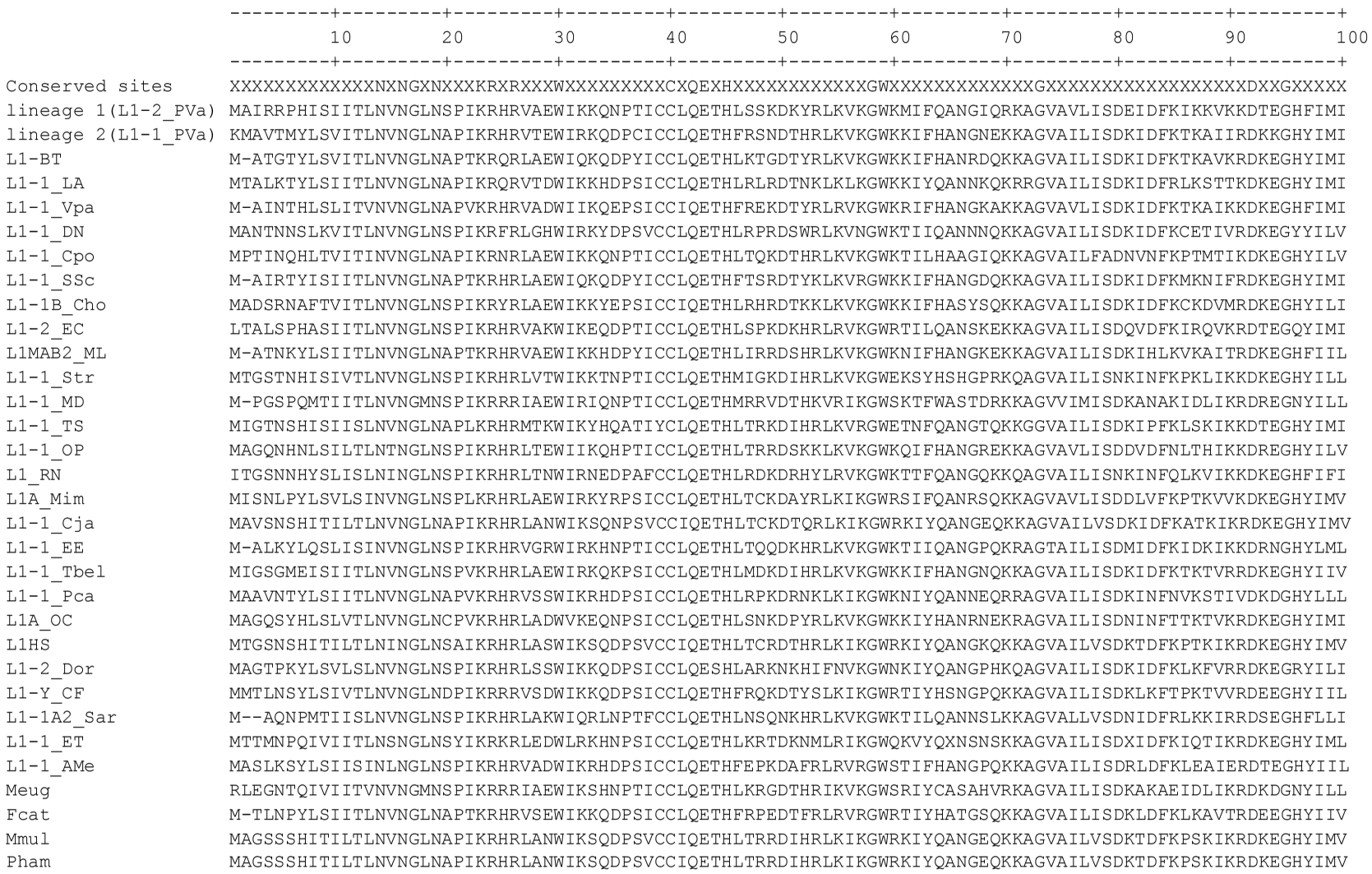
Text 2.S1: (See Chapter 2, Text S1) Alignment of L1 ORF1 sequences.

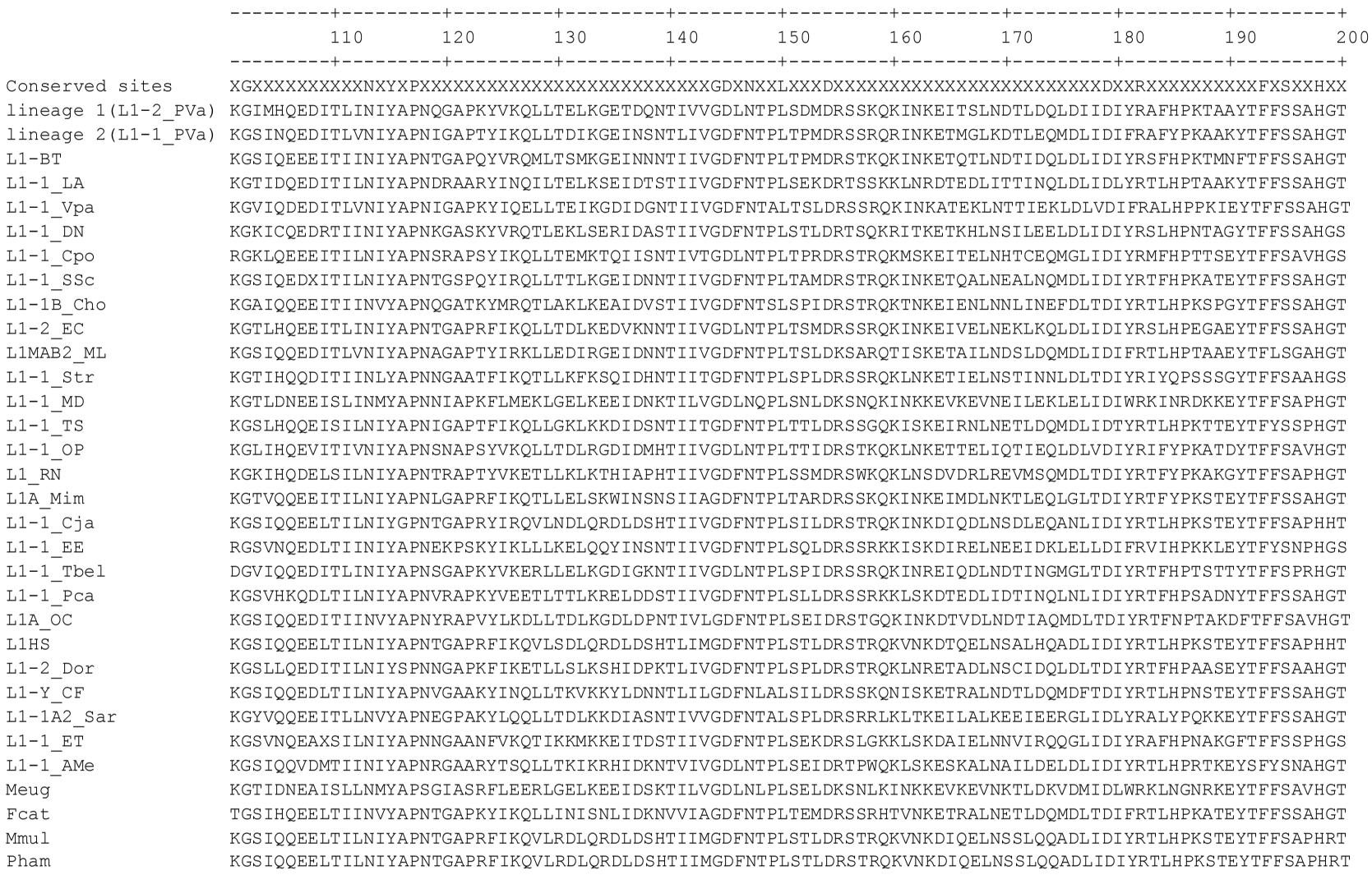
```
                 ---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
                          10        20        30        40        50        60        70        80        90        100
                 ---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
Conserved sites  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXNXRXXGXXEXXXXXXXXXXXXXXXXXXXXEXXXXXXXXXXXXXXXXXXXRXPXXXXXXXXXXRXXXXXX
lineage 1(L1-2_PVa) DRNLEINQKEEERNRRMKNNEREIQELADTIRRGNIRIMGIIE-GEEKEQGLESIFRQIVDENFPNLRNEL-ELGIQEVNRTPNYLNPKRPSPRHIVLKL
lineage 2(L1-1_PVa) DKVAEYTQSEQQKEKRIXKNEDSLRDLWDNMKRKNIRIIGVPE-EEEREQGIENLFEEIMTENFPNLAKEL-DIQVQEAQRVPNKRNSNRPTPRHIIIKM
L1-BT           DRMVEINESERIKEKRIKRNEDNLRDLQDNIKRYNIRIIGVPE-EEDKKKDHEKILEEIIVENFPKMGKEI-ITQVQETQRVPNRINPRRNTPRHILIKL
L1-1_LA         DQGINTNIAEKKSDKRIKKNEETLRIMWDSIKKDNLRVIGVPE-QGGGTENTEKIVEELLTENFPDIMKDE-RISIQDAHRTPFKIDPKRKTPRHIIIKL
L1-1_Vpa        DRATESTHSEELQDKQIKNNDNSIRDLWDNIKRPNLRIIGVPE-GEEGSKGIEKVFEEIMTENFPNLKKES-DIQVQEAQRVPNRKNPNRPTPRHIIIKM
L1-1_DN         DSTSEIKQIVEXVDKKIEKIQLGLRDLNDNAKRSNIRIIGIPE-GEEKGKGSEGVLQEIMAENFPNLLKET-DVHIQEAQRTPLVINPNRPTPRHILVKL
L1-1_Cpo        DRIAASEQERKDLLKITRNQETTIQQLQDDAKKNNIRMIGINEKEGDNIKDVKRIFREVIAENFPSMRSET-DIRISEAYRTPNSHNQNKTTPRHIIITI
L1-1_SSc        DRLVEITDAEQKREKRLKTNEESLRELWDNVKRTNIRIIGVPE-GEEREKETEKIFQEIIAENFPNMGKES-LTQIQEAQRVPYKINPRRNTPRHILIKL
L1-1B_Cho       DDRMENERTKERMGKKIEKIEMDLRDMIDKIKRPNIRLIGVPE-GEEKGKGLERVFKEIVGENFPNLLHNI-NTQSINAQRTPNRINPNKPTPRHILIRL
L1-2_EC         DRQAEWRQTEEERELRIKKNEENLREIMDSMRSKNIRIIGIPE-NMEKENGAESVLNEIIEENFPNLGING-EMCVEEGFRSPRFVNVKRPTARHIIVKL
L1MAB2_ML       DKEAKHTQTVLQMEKKIKRQEESLRELWDNMKRNNIRIIGVPE-QQKDEHGLENLFEEIISENFPEVGKK----KVTQAQRVPNKVNPKRPTPRHIIITM
L1-1_Str        DKVYQLEKSIVNTEKMLKSHEQSIQEIWDVXKKPNLRVIGIEEGTEIQTKGMDSILNEIIVENFPEMKDGM-DCQILEAYRTPNIQNHNRPTPRHIIMKI
L1-1_MD         QELIKQSQNTKKLEENIKYLTDKVIDLENRGRRENLRIIGLPE-KPEINTKLDMVIQDIIKENCPEILEQGGNTSTDRAHRTPSTLNPQKTTPRNVIAKF
L1-1_TS         DQNIEITQTLKNTENKLKKTEQNLQEMSDYLKRPNLRIIGLPEAERETETTLEQTFHEIIQENFPYLISDA-KIQTQEIQRTPARQQMRRPTPRHIIIRL
L1-1_OP         DTQNEHTQFIKQLETSLNKANKTIQEMKDNLRKSNIRIIGLPE-GAEKESGMQMVLDEIIQENFQNTWNMN-PAQIQDGQRTPSRYDPKRSSPRHMVLKF
L1_RN           DSIEIIDSTVKDNVKRKKLLVQNIQEIQDSMRRSNLRIIGIEESEDSQLKGPVNIFNKIIEENFPNLKKEI-PIDIQEAYRTPNRLDQKRNTSRHIIVKT
L1A_Mim         DNTLQLNKSVTEIEQRNKRKEQSLQELWDYVKKPNVRVIGLPEGEEDNTQGLDKLFEDIIEENFPGLAQNL-DIQVQEAQRTPGRFNANRKTSRHAVIRL
L1-1_Cja        DQLNEIKREGKIREKSAKRNEQSLQEMWDYVKRPNLRLIGVPECDEENESKLENTLQDIIQENFPNLARQA-NIQVQEIQRTPQRYSSRRATPRHIIVRF
L1-1_EE         DELETTKKEVRDLKKRLRDAENNNRVLWDDFKRNNIRIIGLPE-EEREG-EEESILQAIIAENFSSLDNTK-DIKIQEAQRVPNRINPDLKTPRHVILRL
L1-1_Tbel       DKTSDLDKSIKKIEKTTMKNEDNIRGILDTIKRPNIRIIGIPE-EEENNKGLENLFHEILEENFPNLERHS-NIQTQEIQRTPSRINPRRSSPRHIIARL
L1-1_Pca        DRATDASVYEQKSEKRNKKNEETLRSMWDSIKRNNLRLIGVPE-QGETSENTESIVAELLKENFPEIMKEE-NIDISDAYRTPPNIDLKRKTPRHIIIKL
L1A_OC          DREQERIQSDQRKEEEIRNLKHIVGNLQDTIKKPNIRVLGVPE-GMEREKGLEGLFSEILAENFPGLEKDR-EILVQEAHRTPNKHDQKRSSPRHVVIKL
L1HS            DEMNEMKREGKFREKRIKRNEQSLQEIWDYVKRPNLRLIGVPESDVENGTKLENTLQDIIQENFPNLARQA-NVQIQEIQRTPQRYSSRRATPRHIIVRF
L1-2_Dor        DSLDIIEKEHTTIQDKSKKYDRSIQHLEDTIRRPNLRIMGVEEHLETEVNGLGNLFNRILAENFPNIQKDR-PIQIQEAFRTPNRPDQNRTSHRHIIIHT
L1-Y_CF         DKLIAKRETEEKRDKQLKDHEDRLREINDSLRKKNLRLIGVPE-GAERDRGPEYVFEQILAENFPNLGRET-GIQIQEIERSPPKINKNRSTPRHLIVKL
L1-1A2_Sar      DELQAAYRQQQXMGRDLKIALGRIRVLGDXFKRNNIRIIGLPE-GQGTNPNEKATVKKIIAEKFPELDN-A-GIQIQGARRVPAKRDPNRKTPRHIIVTM
L1-1_ET         DTQTDLSKRERQSDRKIKETEDSLRSMTDAMKRNNIRIIGLPE-HNTTHKSTAKIAKEFLEENFPTLTREN-QALIQEAERTPARLDPKKNTPRHIIVKI
L1-1_AMe        DGLVEEKTKIEAGLKKIHAHECRLREITDSMKRSNVRIIGIPE-GVEKNRGLEEIFEQIVAENFPNLARET-SIRVQEAERTPSKLNQDKPTPRHVIVQF
Meug            QESVKQNLKNEKIEENLKYLIGKTTDLENRSRRENLRIIGLPE-THDEEKSLDNIFQEIIKENCPEVLDSEGKIVIERIHRSPPERDPKLKTPRNIVAKF
Fcat            DKVMEKEEAEKKRDKKIQEYEGKIRELSDTLKRNNIRIIGIPE-EEERGKGAEGVLEEIIAENFPELGKEK-GIEIQEAQRTPFRRNLNRSSARHIIVKL
Mmul            DQMNEMKREEKPKEKRRKRNEQSLQEVWDYVKRPNLRLIGVPESEGENGTKLENTLQDIIQENFPNLVGQA-NIQIQEIQRTPQRYSSRRATPRHIIARF
Pham            DQMNEMKREEKPKEKRRKRNEQSLQEVWDYVKRPNLRLIGVPESEGENGTKLENTLQDIIQENFPNLVGQA-NIQIQEIQRTPQRYSSRRATXRHIIARF
```

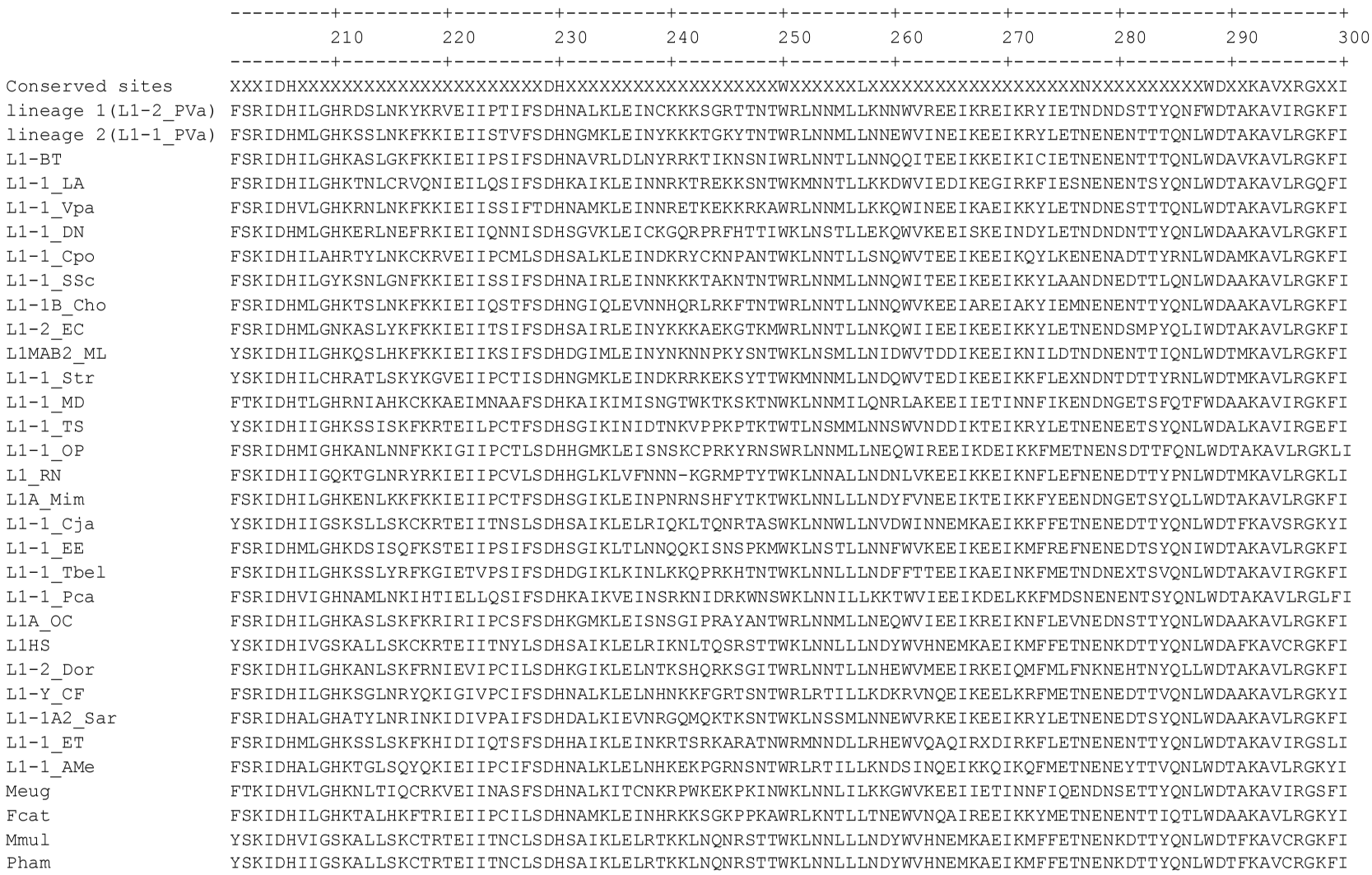

110 120 130 140 150 160 170 180 190 200
Conserved sites XXXXXXXXXLXXXXXXXXXXXXGXXXRXXXDXXXXTXXXRXXWXXXXXXXXXXXXXXPRXXXPAXXXXXXXXXXXXXFXXXXXXXXXXXXXXXXXXXXXXX
lineage 1(L1-2_PVa) SKINDKDRILRAAREKKTVTYKGKPIRLSSDFSAQTLQARKEWNQIFKLLSERNYQPRIMYPAKLSFRYEGEIKTFPDIQKLREFSTTRPALQEILKGVS
lineage 2(L1-1_PVa) PKVKDKERILKAARERKLVTYKGTPIRMSADFSMETLQARREWQEIFKVMKNKNLQPRILYPARLSIKMEGEIKSFPDRKKLKEFITNKPALQEMLKGLL
L1-BT TTIKHKEQILKAAREKQQITHKGIPIRITADLSIETLQARREWQDILKMMKENNLQPRLLYPARISFKYEGEIKSFSDKQKLREFCTTKPALQQILKDIL
L1-1_LA AKTKDKQKILKAAREKRKVSFKGESIRISSDYSAETMQARREWDDIYRTLKEKNCQPRIIYPAKLSLKYEGEIKIFTDKHKFREFAKTKPKLQEILKDIV
L1-1_Vpa ARVKDKEMILKAAREKQRVNYKGTPIRLSADFSTQTLQARREWQDIFKALNEKKMQPRILYPARLSFRIEGEIKSFTDKKKAAGV.
L1-1_DN SNAQDKEKILKAAREKKTITYKGSSIRLSADFSSETMEARRQWYDIVKVLKEKNFQPRILYPAKLAFKHDGEFKIFTDKQKLKEYTNKKPPLQEILKGVL
L1-1_Cpo PEIQHKNRLLKAVREKRQITYKGKPIRITADFSAQTIKSRRAWSEVFQILKQNDFQPRLIYPAKLSFKIDGEIRYFHDKEQLKNFMNTKPTLQKILKDSL
L1-1_SSc TKIKDKEKILKAAREKKQITYKGTPIRLSADFSTETLQARREWHDILNVMKGKNLQPRLLYPARLSFRFEGEIKTFTDKQKLREFSNTKPALQQILKELL
L1-1B_Cho SNTEEKEQVLKAAREKQFTTYKGNNIRLSSDYSAATMEARRQWHDIFKILREKNCQPRILYPAKLSFKFEGELKFFTDKQMLRDFANKRPALLEILKGAL
L1-2_EC ANRNDKERILREVRKKKRITYKGAPIRLSADFSTETLQARREWSDIFKALKDKNLQPRILYPARISFRYEGEIKSFPDKQKLREFVTKSPPLQEILKKAL
L1MAB2_ML ANVQDKERILQAARERRKVTYKGSPIRLSNDFSTETHQARKEWTEIYKVMQSKGLNPRILYPARLSFKIEGEIRSFTDKKGLREFITTKPAMQEMLKGLV
L1-1_Str ANIQDKERILKATREKRQITFRGKPIRLTTDFSXQTLKARRSWNNVFQTLKNNGFQPRILYPAKLSFRFDNEIKIFHDKQKLKEFAARKPALQSILSKIL
L1-1_MD QSYQTKEKILQEAR-KRQFRYKGMPIRVTQDLASSTLNDRKAWNMIFRKARELGLQPRISYPAKLTIYFQGKVWAFNKIEDFQLFAKKRPELCGKFDTEN
L1-1_TS NKVGTKEKILKAAREKGQITYHGRPIRIAADLSAETLQARRAWSPIFKVLKDKQFQPRITYPAKLSFISEGELKSFPDIQSLRTYAATKPSLHETLKKVL
L1-1_OP HSNEDKERILRQVRSREKITYRGKPIRITADFSEETLQARREWTKIFQILNQNNCQPRILYPAKISFVFENEIKYFHSKEKLEEYASTKPALQNLLRNAL
L1_RN PNAQNKERILKAVREKGQVTYKGRPIRITPDFSPETMKARRSWTDVIQTLREHKCQPRLLYPAKLSINIDGETKIFHDKTKFTQYLSTNPALQRIINGKA
L1A_Mim TKVSTKEALLRAVRQKKQVTYKGKPIRITSDFSNETLQARRDWGPILTLLKQNNAQPRILFPAKLSFVYEGEIKTFSDKQRLREFTKTRPALQEVLKTAL
L1-1_Cja TRVEMKEKMLRAAREKGRVTHKGKPIRLTADLSAETLQARREWGPIFNILKEKNFQPRISYPAKLSFISEGKIKSFANKQVLRDFVTTRPALQELLKEAL
L1-1_EE ERNKDKERILKAAREKQRVTYKGKPIRLAADFSIQTLQARREWQDIYRVLNEKGFQPRILYPARLSFRLDGSIKTFSDKQQLKEATITKPALKEVLKGLL
L1-1_Tbel AKTTDKDKILKLARGKRQVTYKGHPIRLTSDLSAETMQERKEWGSIVKVLSEKQFQPRILYPAKLSFIWEKKIKTFSNKQNLKEFTNTRPALQEVLKGVL
L1-1_Pca NKMKHKQQLLKAARLKAKLTFRGKPCRLSSDFSAETMLARRQWHDTFKALKEKNFQPRIIYPAKLSFKYENEIKTFPDKQKLRDFVKTKPKLQEILKEVL
L1A_OC STVKHKEKILKCAREKRQITLRGSPIRLTADFSSETLQARREWRDIAQVLRENNCQPRILYPAKLSFVNEGEIKTFHSKQKLKEFVATRPALQKMLKDVL
L1HS TKVEMKEKMLRAAREKGRVTLKGKPIRLTADLSAETLQARREWGPIFNILKEKNFQPRISYPAKLSFISEGEIKYFIDKQMLRDFVTTRPALKELLKEAL
L1-2_Dor GSTQTKEKILKAVKEKRVITYKGKPIRITSDFSAETIKARRAWNDVCHALKTNNYQPRILYPAKLSFITEGQIKTFHSKEKLKQYISTKPALQKILKDVL
L1-Y_CF ANSKDKEKILKAARDKKSLTFMGRSIRVTADLSTETWQARKGWQDIFRVLNEKNMQPRILYPARLSFKMEGEIKSFQDRQQLKEYVTSKPALQEILRGPL
L1-1A2_Sar MDVTDRDTILQAARSKKEIAYKGAPLRFTADLSEETLQARRQWWDIVKKLNEMNASPRILYPAKLSFKLEGIIHYFGDKQQLRNFIDSKPNLKEGLKGLL
L1-1_ET SNFEEKEKILRAARERKTVTYKGAQVRICSDLSADTMKRRREWSNIFQKLKEKNASPRILYPAKLSIKIDGEIRVFQDKERLKEYAARHPSLRKILADSL
L1-1_AMe ANIRSKDTVLKAARAKKFLTYQGKGIRITSDLSTETWNERKAWGGIFKALSEKNMQPRILYPAKLSFRIDGEIKTFQNRQSLTNFVTTKPALQEILRGAL
Meug QSYQVKEKILQAAR-KKQFKYQGHTVRITQDLAASTLKDRKNWNPIFRKAKELGLQPRINYPAKFSITFQARRKSFNEIRDFQSFLTKTPELNRQFDLQM
Fcat AKYKDKEKILKAARGKRALTYKGRPIRLVTDLSFETWQARKNWHEIFRVLDRKNMQPRILYPASLSFRIEGEIKVFPNKQKLKEFVTTKPALQEILRGTL
Mmul TKVEMKEKILRAAREKGRVTHKGKPIRLTADLSAETLQARREWGPIFNILKEKNFKPRISYPAKLSFISEGEIKSFTDKQMLRDFVTTRPALQETLKEAL
Pham TKVEMKEKILRAAREKGRVTHKGKPIRLTADLSAETLQARREWGPIFNILKEKNFKPRISYPAKLSFISEGEIKSFTDKQMLRDFVTTRPALQETLKEAL

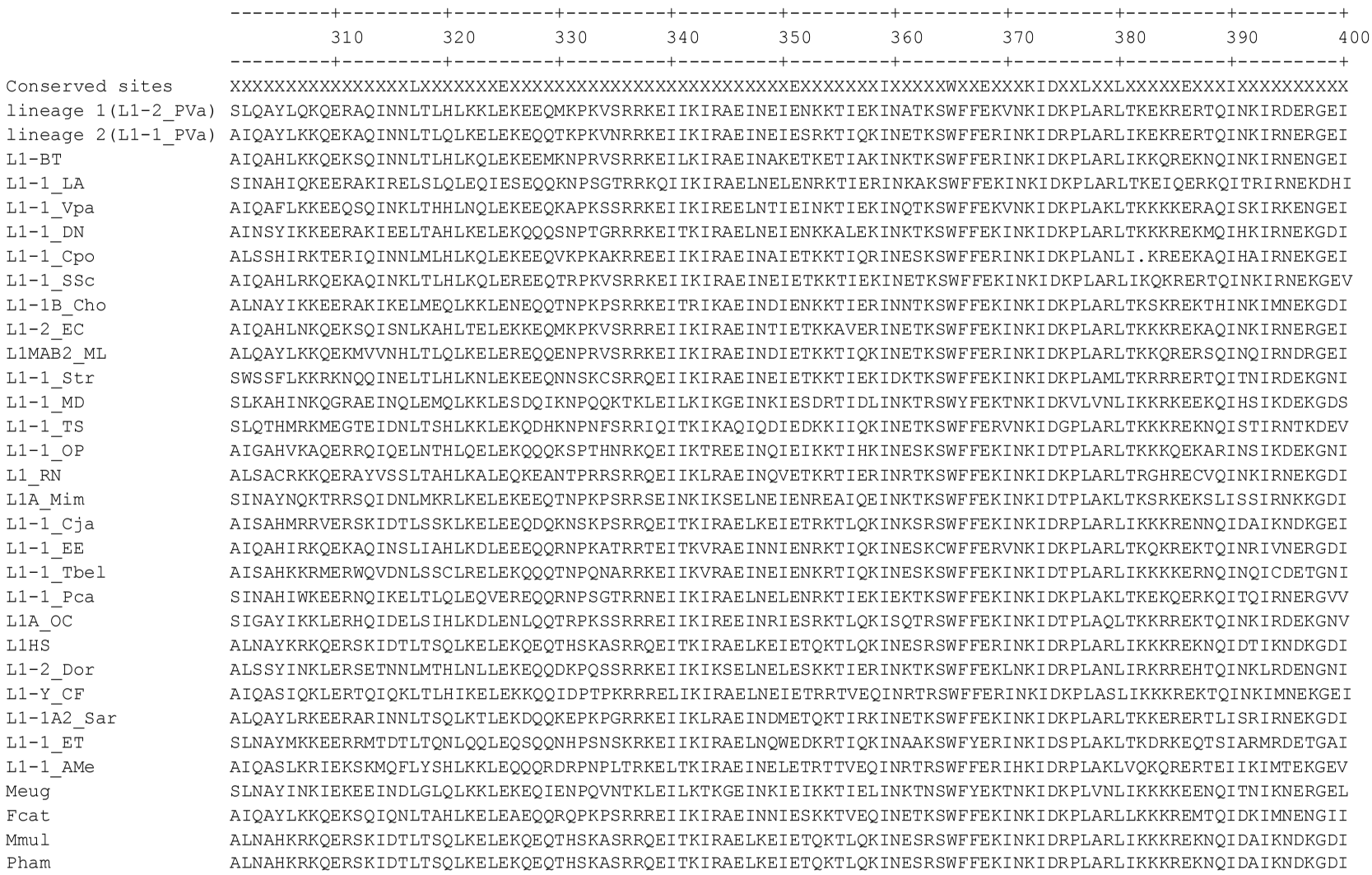
APPENDIX B

Text 2.S2: (See Chapter 2, Text S2) Alignment of L1 ORF2 sequences.


10 20 30 40 50 60 70 80 90 100
Conserved sites XXXXXXXXXXXXXNXNGXNXXXKRXRXXXWXXXXXXXXXCXQEXHXXXXXXXXXXXXXGWXXXXXXXXXXXXXGXXXXXXXXXXXXXXXXXXDXXGXXXXX
lineage 1(L1-2_PVa) MAIRRPHISIITLNVNGLNSPIKRHRVAEWIKKQNPTICCLQETHLSSKDKYRLKVKGWKMIFQANGIQRKAGVAVLISDEIDFKIKKVKKDTEGHFIMI
lineage 2(L1-1_PVa) KMAVTMYLSVITLNVNGLNAPIKRHRVTEWIRKQDPCICCLQETHFRSNDTHRLKVKGWKKIFHANGNEKKAGVAILISDKIDFKTKAIIRDKKGHYIMI
L1-BT M-ATGTYLSVITLNVNGLNAPTKRQRLAEWIQKQDPYICCLQETHLKTGDTYRLKVKGWKKIFHANRDQKKAGVAILISDKIDFKTKAVKRDKEGHYIMI
L1-1_LA MTALKTYLSIITLNVNGLNAPIKRQRVTDWIKKHDPSICCLQETHLRLRDTNKLKLKGWKKIYQANNKQKRRGVAILISDKIDFRLKSTTKDKEGHYIMI
L1-1_Vpa M-AINTHLSLITVNVNGLNAPVKRHRVADWIIKQEPSICCIQETHFREKDTYRLRVKGWKRIFHANGKAKKAGVAVLISDKIDFKTKAIKKDKEGHFIMI
L1-1_DN MANTNNSLKVITLNVNGLNSPIKRFRLGHWIRKYDPSVCCLQETHLRPRDSWRLKVNGWKTIIQANNNQKKAGVAILISDKIDFKCETIVRDKEGYYILV
L1-1_Cpo MPTINQHLTVITINVNGLNAPIKRNRLAEWIKKQNPTICCLQETHLTQKDTHRLKVKGWKTILHAAGIQKKAGVAILFADNVNFKPTMTIKDKEGHYILV
L1-1_SSc M-AIRTYISIITLNVNGLNAPTKRHRLAEWIQKQDPYICCLQETHFTSRDTYKLKVRGWKKIFHANGDQKKAGVAILISDKIDFKMKNIFRDKEGHYIMI
L1-1B_Cho MADSRNAFTVITLNVNGLNSPIKRYRLAEWIKKYEPSICCIQETHLRHRDTKKLKVKGWKKIFHASYSQKKAGVAILISDKIDFKCKDVMRDKEGHYILI
L1-2_EC LTALSPHASIITLNVNGLNSPIKRHRVAKWIKEQDPTICCLQETHLSPKDKHRLRVKGWRTILQANSKEKKAGVAILISDQVDFKIRQVKRDTEGQYIMI
L1MAB2_ML M-ATNKYLSIITLNVNGLNAPTKRHRVAEWIKKHDPYICCLQETHLIRRDSHRLKVKGWKNIFHANGKEKKAGVAILISDKIHLKVKAITRDKEGHFIIL
L1-1_Str MTGSTNHISIVTLNVNGLNSPIKRHRLVTWIKKTNPTICCLQETHMIGKDIHRLKVKGWEKSYHSHGPRKQAGVAILISNKINFKPKLIKKDKEGHYILL
L1-1_MD M-PGSPQMTIITLNVNGMNSPIKRRRIAEWIRIQNPTICCLQETHMRRVDTHKVRIKGWSKTFWASTDRKKAGVVIMISDKANAKIDLIKRDREGNYILL
L1-1_TS MIGTNSHISIISLNVNGLNAPLKRHRMTKWIKYHQATIYCLQETHLTRKDIHRLKVRGWETNFQANGTQKKGGVAILISDKIPFKLSKIKKDTEGHYIMI
L1-1_OP MAGQNHNLSILTLNTNGLNSPIKRHRLTEWIIKQHPTICCLQETHLTRRDSKKLKVKGWKQIFHANGREKKAGVAVLISDDVDFNLTHIKKDREGHYILV
L1_RN ITGSNNHYSLISLNINGLNSPIKRHRLTNWIRNEDPAFCCLQETHLRDKDRHYLRVKGWKTTFQANGQKKQAGVAILISNKINFQLKVIKKDKEGHFIFI
L1A_Mim MISNLPYLSVLSINVNGLNSPLKRHRLAEWIRKYRPSICCLQETHLTCKDAYRLKIKGWRSIFQANRSQKKAGVAVLISDDLVFKPTKVVKDKEGHYIMV
L1-1_Cja MAVSNSHITILTLNVNGLNAPIKRHRLANWIKSQNPSVCCIQETHLTCKDTQRLKIKGWRKIYQANGEQKKAGVAILVSDKIDFKATKIKRDKEGHYIMV
L1-1_EE M-ALKYLQSLISINVNGLNSPIKRHRVGRWIRKHNPTICCLQETHLTQQDKHRLKVKGWKTIIQANGPQKRAGTAILISDMIDFKIDKIKKDRNGHYLML
L1-1_Tbel MIGSGMEISIITLNVNGLNSPVKRHRLAEWIRKQKPSICCLQETHLMDKDIHRLKVKGWKKIFHANGNQKKAGVAILISDKIDFKTKTVRRDKEGHYIIV
L1-1_Pca MAAVNTYLSIITLNVNGLNAPVKRHRVSSWIKRHDPSICCLQETHLRPKDRNKLKIKGWKNIYQANNEQRRAGVAILISDKINFNVKSTIVDKDGHYLLL
L1A_OC MAGQSYHLSLVTLNVNGLNCPVKRHRLADWVKEQNPSICCLQETHLSNKDPYRLKVKGWKKIYHANRNEKRAGVAILISDNINFTTKTVKRDKEGHYIMI
L1HS MTGSNSHITILTLNINGLNSAIKRHRLASWIKSQDPSVCCIQETHLTCRDTHRLKIKGWRKIYQANGKQKKAGVAILVSDKTDFKPTKIKRDKEGHYIMV
L1-2_Dor MAGTPKYLSVLSLNVNGLNSPIKRHRLSSWIKKQDPSICCLQESHLARKNKHIFNVKGWNKIYQANGPHKQAGVAILISDKIDFKLKFVRRDKEGRYILI
L1-Y_CF MMTLNSYLSIVTLNVNGLNDPIKRRRVSDWIKKQDPSICCLQETHFRQKDTYSLKIKGWRTIYHSNGPQKKAGVAILISDKLKFTPKTVVRDEEGHYIIL
L1-1A2_Sar M--AQNPMTIISLNVNGLNSPIKRHRLAKWIQRLNPTFCCLQETHLNSQNKHRLKVKGWKTILQANNSLKKAGVALLVSDNIDFRLKKIRRDSEGHFLLI
L1-1_ET MTTMNPQIVIITLNSNGLNSYIKRKRLEDWLRKHNPSICCLQETHLKRTDKNMLRIKGWQKVYQXNSNSKKAGVAILISDXIDFKIQTIKRDKEGHYIML
L1-1_AMe MASLKSYLSIISINLNGLNSPIKRHRVADWIKRHDPSICCLQETHFEPKDAFRLRVRGWSTIFHANGPQKKAGVAILISDRLDFKLEAIERDTEGHYIIL
Meug RLEGNTQIVIITVNVNGMNSPIKRRRIAEWIKSHNPTICCLQETHLKRGDTHRIKVKGWSRIYCASAHVRKAGVAILISDKAKAEIDLIKRDKDGNYILL
Fcat M-TLNPYLSIITLNVNGLNAPTKRHRVSEWIKKQDPSICCLQETHFRPEDTFRLRVRGWRTIYHATGSQKKAGVAILISDKLDFKLKAVTRDEEGHYIIV
Mmul MAGSSSHITILTLNVNGLNAPIKRHRLANWIKSQDPSVCCIQETHLTRRDIHRLKIKGWRKIYQANGEQKKAGVAILVSDKTDFKPSKIKRDKEGHYIMV
Pham MAGSSSHITILTLNVNGLNAPIKRHRLANWIKSQDPSVCCIQETHLTRRDIHRLKIKGWRKIYQANGEQKKAGVAILVSDKTDFKPSKIKRDKEGHYIMV


110 120 130 140 150 160 170 180 190 200
Conserved sites XGXXXXXXXXXXNXYXPXXXXXXXXXXXXXXXXXXXXXXXXXXGDXNXXLXXXDXXXXXXXXXXXXXXXXXXXXXXXXDXXRXXXXXXXXXXFXSXXHXX
lineage 1(L1-2_PVa) KGIMHQEDITLINIYAPNQGAPKYVKQLLTELKGETDQNTIVVGDLNTPLSDMDRSSKQKINKEITSLNDTLDQLDIIDIYRAFHPKTAAYTFFSSAHGT
lineage 2(L1-1_PVa) KGSINQEDITLVNIYAPNIGAPTYIKQLLTDIKGEINSNTLIVGDFNTPLTPMDRSSRQRINKETMGLKDTLEQMDLIDIFRAFYPKAAKYTFFSSAHGT
L1-BT KGSIQEEEITIINIYAPNTGAPQYVRQMLTSMKGEINNNTIIVGDFNTPLTPMDRSTKQKINKETQTLNDTIDQLDLIDIYRSFHPKTMNFTFFSSAHGT
L1-1_LA KGTIDQEDITILNIYAPNDRAARYINQILTELKSEIDTSTIIVGDFNTPLSEKDRTSSKKLNRDTEDLITTINQLDLIDLYRTLHPTAAKYTFFSSAHGT
L1-1_Vpa KGVIQDEDITLVNIYAPNIGAPKYIQELLTEIKGDIDGNTIIVGDFNTALTSLDRSSRQKINKATEKLNTTIEKLDLVDIFRALHPPKIEYTFFSSAHGT
L1-1_DN KGKICQEDRTIINIYAPNKGASKYVRQTLEKLSERIDASTIIVGDFNTPLSTLDRTSQKRITKETKHLNSILEELDLIDIYRSLHPNTAGYTFFSSAHGS
L1-1_Cpo RGKLQEEEITILNIYAPNSRAPSYIKQLLTEMKTQIISNTIVTGDLNTPLTPRDRSTRQKMSKEITELNHTCEQMGLIDIYRMFHPTTSEYTFFSAVHGS
L1-1_SSc KGSIQEDXITILNIYAPNTGSPQYIRQLLTTLKGEIDNNTIIVGDFNTPLTAMDRSTRQKINKETQALNEALNQMDLIDIYRTFHPKATEYTFFSSAHGT
L1-1B_Cho KGAIQQEEITIINVYAPNQGATKYMRQTLAKLKEAIDVSTIIVGDFNTSLSPIDRSTRQKTNKEIENLNNLINEFDLTDIYRTLHPKSPGYTFFSSAHGT
L1-2_EC KGTLHQEEITLINIYAPNTGAPRFIKQLLTDLKEDVKNNTIIVGDLNTPLTSMDRSSRQKINKEIVELNEKLKQLDLIDIYRSLHPEGAEYTFFSSAHGT
L1MAB2_ML KGSIQQEDITLVNIYAPNAGAPTYIRKLLEDIRGEIDNNTIIVGDFNTPLTSLDKSARQTISKETAILNDSLDQMDLIDIFRTLHPTAAEYTFLSGAHGT
L1-1_Str KGTIHQQDITIINLYAPNNGAATFIKQTLLKFKSQIDHNTIITGDFNTPLSPLDRSSRQKLNKETIELNSTINNLDLTDIYRIYQPSSSGYTFFSAAHGS
L1-1_MD KGTLDNEEISLINMYAPNNIAPKFLMEKLGELKEEIDNKTILVGDLNQPLSNLDKSNQKINKKEVKEVNEILEKLELIDIWRKINRDKKEYTFFSAPHGT
L1-1_TS KGSLHQQEISILNIYAPNIGAPTFIKQLLGKLKKDIDSNTIITGDFNTPLTTLDRSSGQKISKEIRNLNETLDQMDLIDTYRTLHPKTTEYTFYSSPHGT
L1-1_OP KGLIHQEVITIVNIYAPNSNAPSYVKQLLTDLRGDIDMHTIIVGDLNTPLTTIDRSTKQKLNKETTELIQTIEQLDLVDIYRIFYPKATDYTFFSAVHGT
L1_RN KGKIHQDELSILNIYAPNTRAPTYVKETLLKLKTHIAPHTIIVGDFNTPLSSMDRSWKQKLNSDVDRLREVMSQMDLTDIYRTFYPKAKGYTFFSAPHGT
L1A_Mim KGTVQQEEITILNIYAPNLGAPRFIKQTLLELSKWINSNSIIAGDFNTPLTARDRSSKQKINKEIMDLNKTLEQLGLTDIYRTFYPKSTEYTFFSSAHGT
L1-1_Cja KGSIQQEELTILNIYGPNTGAPRYIRQVLNDLQRDLDSHTIIVGDFNTPLSILDRSTRQKINKDIQDLNSDLEQANLIDIYRTLHPKSTEYTFFSAPHHT
L1-1_EE RGSVNQEDLTIINIYAPNEKPSKYIKLLLKELQQYINSNTIIVGDFNTPLSQLDRSSRKKISKDIRELNEEIDKLELLDIFRVIHPKKLEYTFYSNPHGS
L1-1_Tbel DGVIQQEDITLINIYAPNSGAPKYVKERLLELKGDIGKNTIIVGDLNTPLSPIDRSSRQKINREIQDLNDTINGMGLTDIYRTFHPTSTTYTFFSPRHGT
L1-1_Pca KGSVHKQDLTILNIYAPNVRAPKYVEETLTTLKRELDDSTIIVGDFNTPLSLLDRSSRKKLSKDTEDLIDTINQLNLIDIYRTFHPSADNYTFFSSAHGT
L1A_OC KGSIQQEDITIINVYAPNYRAPVYLKDLLTDLKGDLDPNTIVLGDFNTPLSEIDRSTGQKINKDTVDLNDTIAQMDLTDIYRTFNPTAKDFTFFSAVHGT
L1HS KGSIQQEELTILNIYAPNTGAPRFIKQVLSDLQRDLDSHTLIMGDFNTPLSTLDRSTRQKVNKDTQELNSALHQADLIDIYRTLHPKSTEYTFFSAPHHT
L1-2_Dor KGSLLQEDITILNIYSPNNGAPKFIKETLLSLKSHIDPKTLIVGDFNTPLSPLDRSTRQKLNRETADLNSCIDQLDLTDIYRTFHPAASEYTFFSAAHGT
L1-Y_CF KGSIQQEDLTILNIYAPNVGAAKYINQLLTKVKKYLDNNTLILGDFNLALSILDRSSKQNISKETRALNDTLDQMDFTDIYRTLHPNSTEYTFFSSAHGT
L1-1A2_Sar KGYVQQEEITLLNVYAPNEGPAKYLQQLLTDLKKDIASNTIVVGDFNTALSPLDRSRRLKLTKEILALKEEIEERGLIDLYRALYPQKKEYTFFSSAHGT
L1-1_ET KGSVNQEAXSILNIYAPNNGAANFVKQTIKKMKKEITDSTIIVGDFNTPLSEKDRSLGKKLSKDAIELNNVIRQQGLIDIYRAFHPNAKGFTFFSSPHGS
L1-1_AMe KGSIQQVDMTIINIYAPNRGAARYTSQLLTKIKRHIDKNTVIVGDLNTPLSEIDRTPWQKLSKESKALNAILDELDLIDIYRTLHPRTKEYSFYSNAHGT
Meug KGTIDNEAISLLNMYAPSGIASRFLEERLGELKEEIDSKTILVGDLNLPLSELDKSNLKINKKEVKEVNKTLDKVDMIDLWRKLNGNRKEYTFFSAVHGT
Fcat TGSIHQEELTIINVYAPNTGAPKYIKQLLINISNLIDKNVVIAGDFNTPLTEMDRSSRHTVNKETRALNETLDQMDLTDIFRTLHPKATEYTFFSSAHGT
Mmul KGSIQQEELTILNIYAPNTGAPRFIKQVLRDLQRDLDSHTIIMGDFNTPLSTLDRSTRQKVNKDIQELNSSLQQADLIDIYRTLHPKSTEYTFFSAPHRT
Pham KGSIQQEELTILNIYAPNTGAPRFIKQVLRDLQRDLDSHTIIMGDFNTPLSTLDRSTRQKVNKDIQELNSSLQQADLIDIYRTLHPKSTEYTFFSAPHRT

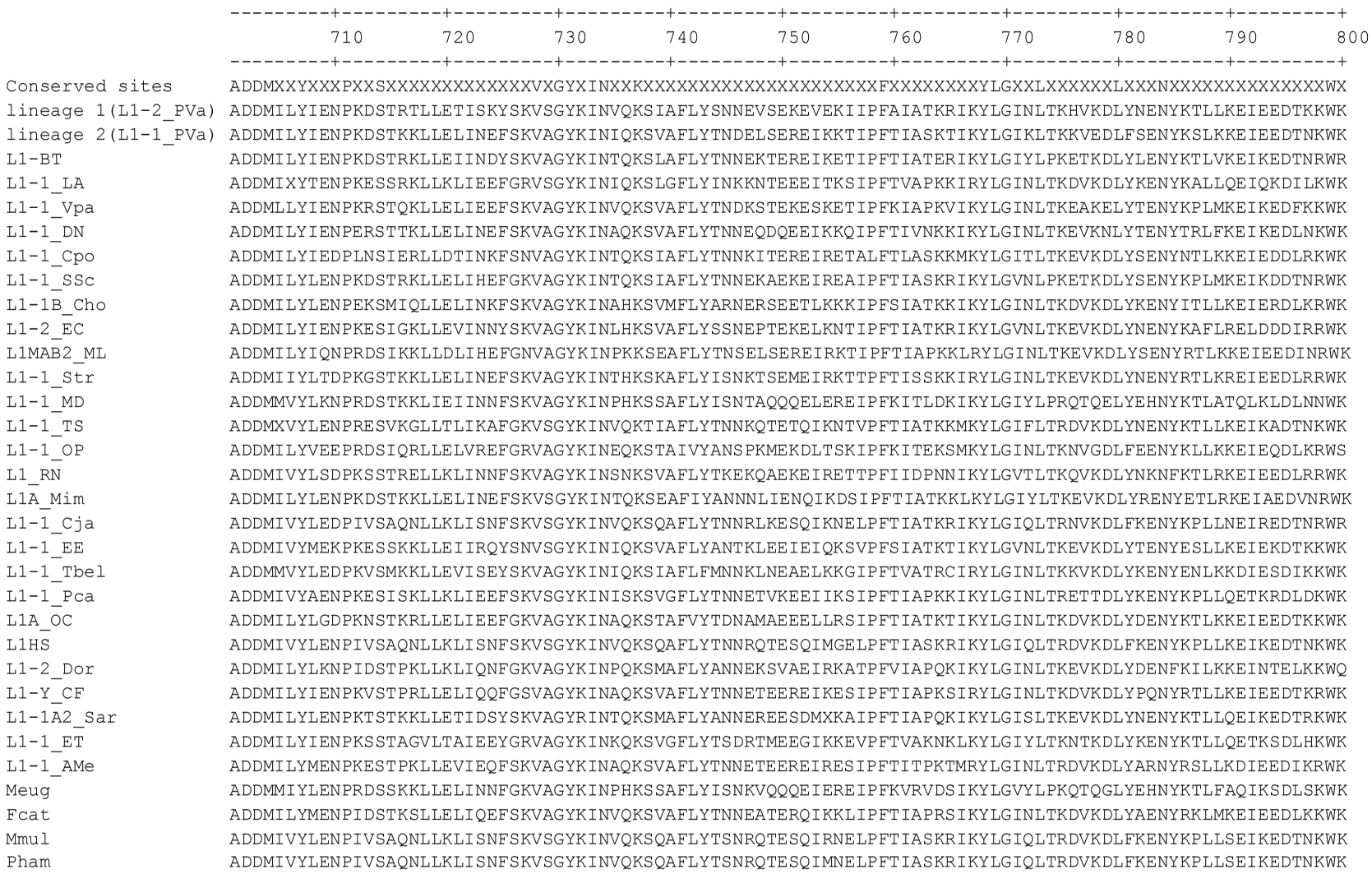

210 220 230 240 250 260 270 280 290 300
Conserved sites XXXIDHXXXXXXXXXXXXXXXXXXXXXDHXXXXXXXXXXXXXXXXXXXWXXXXXXLXXXXXXXXXXXXXXXXXXXNXXXXXXXXXXWDXXKAVXRGXXI
lineage 1(L1-2_PVa) FSRIDHILGHRDSLNKYKRVEIIPTIFSDHNALKLEINCKKKSGRTTNTWRLNNMLLKNNWVREEIKREIKRYIETNDNDSTTYQNFWDTAKAVIRGKFI
lineage 2(L1-1_PVa) FSRIDHMLGHKSSLNKFKKIEIISTVFSDHNGMKLEINYKKKTGKYTNTWRLNNMLLNNEWVINEIKEEIKRYLETNENENTTTQNLWDTAKAVLRGKFI
L1-BT FSRIDHILGHKASLGKFKKIEIIPSIFSDHNAVRLDLNYRRKTIKNSNIWRLNNTLLNNQQITEEIKKEIKICIETNENENTTTQNLWDAVKAVLRGKFI
L1-1_LA FSRIDHILGHKTNLCRVQNIEILQSIFSDHKAIKLEINNRKTREKKSNTWKMNNTLLKKDWVIEDIKEGIRKFIESNENENTSYQNLWDTAKAVLRGQFI
L1-1_Vpa FSRIDHVLGHKRNLNKFKKIEIISSIFTDHNAMKLEINNRETKEKKRKAWRLNNMLLKKQWINEEIKAEIKKYLETNDNESTTTQNLWDTAKAVLRGKFI
L1-1_DN FSKIDHMLGHKERLNEFRKIEIIQNNISDHSGVKLEICKGQRPRFHTTIWKLNSTLLEKQWVKEEISKEINDYLETNDNDNTTYQNLWDAAKAVLRGKFI
L1-1_Cpo FSKIDHILAHRTYLNKCKRVEIIPCMLSDHSALKLEINDKRYCKNPANTWKLNNTLLSNQWVTEEIKEEIKQYLKENENADTTYRNLWDAMKAVLRGKFI
L1-1_SSc FSKIDHILGYKSNLGNFKKIEIISSIFSDHNAIRLEINNKKKTAKNTNTWRLNNMLLNNQWITEEIKEEIKKYLAANDNEDTTLQNLWDAAKAVLRGKFI
L1-1B_Cho FSRIDHMLGHKTSLNKFKKIEIIQSTFSDHNGIQLEVNNHQRLRKFTNTWRLNNTLLNNQWVKEEIAREIAKYIEMNENENTTYQNLWDAAKAVLRGKFI
L1-2_EC FSRIDHMLGNKASLYKFKKIEIITSIFSDHSAIRLEINYKKKAEKGTKMWRLNNTLLNKQWIIEEIKEEIKKYLETNENDSMPYQLIWDTAKAVLRGKFI
L1MAB2_ML YSKIDHILGHKQSLHKFKKIEIIKSIFSDHDGIMLEINYNKNNPKYSNTWKLNSMLLNIDWVTDDIKEEIKNILDTNDNENTTIQNLWDTMKAVLRGKFI
L1-1_Str YSKIDHILCHRATLSKYKGVEIIPCTISDHNGMKLEINDKRRKEKSYTTWKMNNMLLNDQWVTEDIKEEIKKFLEXNDNTDTTYRNLWDTMKAVLRGKFI
L1-1_MD FTKIDHTLGHRNIAHKCKKAEIMNAAFSDHKAIKIMISNGTWKTKSKTNWKLNNMILQNRLAKEEIIETINNFIKENDNGETSFQTFWDAAKAVIRGKFI
L1-1_TS YSKIDHIIGHKSSISKFKRTEILPCTFSDHSGIKINIDTNKVPPKPTKTWTLNSMMLNNSWVNDDIKTEIKRYLETNENEETSYQNLWDALKAVIRGEFI
L1-1_OP FSRIDHMIGHKANLNNFKKIGIIPCTLSDHHGMKLEISNSKCPRKYRNSWRLNNMLLNEQWIREEIKDEIKKFMETNENSDTTFQNLWDTAKAVLRGKLI
L1_RN FSKIDHIIGQKTGLNRYRKIEIIPCVLSDHHGLKLVFNNN-KGRMPTYTWKLNNALLNDNLVKEEIKKEIKNFLEFNENEDTTYPNLWDTMKAVLRGKLI
L1A_Mim FSKIDHILGHKENLKKFKKIEIIPCTFSDHSGIKLEINPNRNSHFYTKTWKLNNLLLNDYFVNEEIKTEIKKFYEENDNGETSYQLLWDTAKAVLRGKFI
L1-1_Cja YSKIDHIIGSKSLLSKCKRTEIITNSLSDHSAIKLELRIQKLTQNRTASWKLNNWLLNVDWINNEMKAEIKKFFETNENEDTTYQNLWDTFKAVSRGKYI
L1-1_EE FSRIDHMLGHKDSISQFKSTEIIPSIFSDHSGIKLTLNNQQKISNSPKMWKLNSTLLNNFWVKEEIKEEIKMFREFNENEDTSYQNIWDTAKAVLRGKFI
L1-1_Tbel FSKIDHILGHKSSLYRFKGIETVPSIFSDHDGIKLKINLKKQPRKHTNTWKLNNLLLNDFFTTEEIKAEINKFMETNDNEXTSVQNLWDTAKAVIRGKFI
L1-1_Pca FSRIDHVIGHNAMLNKIHTIELLQSIFSDHKAIKVEINSRKNIDRKWNSWKLNNILLKKTWVIEEIKDELKKFMDSNENENTSYQNLWDTAKAVLRGLFI
L1A_OC FSRIDHILGHKASLSKFKRIRIIPCSFSDHKGMKLEISNSGIPRAYANTWRLNNMLLNEQWVIEEIKREIKNFLEVNEDNSTTYQNLWDAAKAVLRGKFI
L1HS YSKIDHIVGSKALLSKCKRTEIITNYLSDHSAIKLELRIKNLTQSRSTTWKLNNLLLNDYWVHNEMKAEIKMFFETNENKDTTYQNLWDAFKAVCRGKFI
L1-2_Dor FSKIDHILGHKANLSKFRNIEVIPCILSDHKGIKLELNTKSHQRKSGITWRLNNTLLNHEWVMEEIRKEIQMFMLFNKNEHTNYQLLWDTAKAVLRGKFI
L1-Y_CF FSRIDHILGHKSGLNRYQKIGIVPCIFSDHNALKLELNHNKKFGRTSNTWRLRTILLKDKRVNQEIKEELKRFMETNENEDTTVQNLWDAAKAVLRGKYI
L1-1A2_Sar FSRIDHALGHATYLNRINKIDIVPAIFSDHDALKIEVNRGQMQKTKSNTWKLNSSMLNNEWVRKEIKEEIKRYLETNENEDTSYQNLWDAAKAVLRGKFI
L1-1_ET FSRIDHMLGHKSSLSKFKHIDIIQTSFSDHHAIKLEINKRTSRKARATNWRMNNDLLRHEWVQAQIRXDIRKFLETNENENTTYQNLWDTAKAVIRGSLI
L1-1_AMe FSRIDHALGHKTGLSQYQKIEIIPCIFSDHNALKLELNHKEKPGRNSNTWRLRTILLKNDSINQEIKKQIKQFMETNENEYTTVQNLWDTAKAVLRGKYI
Meug FTKIDHVLGHKNLTIQCRKVEIINASFSDHNALKITCNKRPWKEKPKINWKLNNLILKKGWVKEEIIETINNFIQENDNSETTYQNLWDTAKAVIRGSFI
Fcat FSKIDHILGHKTALHKFTRIEIIPCILSDHNAMKLEINHRKKSGKPPKAWRLKNTLLTNEWVNQAIREEIKKYMETNENENTTIQTLWDAAKAVLRGKYI
Mmul YSKIDHVIGSKALLSKCTRTEIITNCLSDHSAIKLELRTKKLNQNRSTTWKLNNLLLNDYWVHNEMKAEIKMFFETNENKDTTYQNLWDTFKAVCRGKFI
Pham YSKIDHIIGSKALLSKCTRTEIITNCLSDHSAIKLELRTKKLNQNRSTTWKLNNLLLNDYWVHNEMKAEIKMFFETNENKDTTYQNLWDTFKAVCRGKFI


310 320 330 340 350 360 370 380 390 400
Conserved sites XXXXXXXXXXXXXXXXLXXXXXXXEXXXXXXXXXXXXXXXXXXXXXXXXEXXXXXXIXXXXWXXEXXXKIDXXLXXLXXXXXEXXXIXXXXXXXXXX
lineage 1(L1-2_PVa) SLQAYLQKQERAQINNLTLHLKKLEKEEQMKPKVSRRKEIIKIRAEINEIENKKTIEKINATKSWFFEKVNKIDKPLARLTKEKRERTQINKIRDERGEI
lineage 2(L1-1_PVa) AIQAYLKKQEKAQINNLTLQLKELEKEQQTKPKVNRRKEIIKIRAEINEIESRKTIQKINETKSWFFERINKIDRPLARLIKEKRERTQINKIRNERGEI
L1-BT AIQAHLKKQEKSQINNLTLHLKQLEKEEMKNPRVSRRKEILKIRAEINAKETKETIAKINKTKSWFFERINKIDKPLARLIKKQREKNQINKIRNENGEI
L1-1_LA SINAHIQKEERAKIRELSLQLEQIESEQQKNPSGTRRKQIIKIRAELNELENRKTIERINKAKSWFFEKINKIDKPLARLTKEIQERKQITRIRNEKDHI
L1-1_Vpa AIQAFLKKEEQSQINKLTHHLNQLEKEEQKAPKSSRRKEIIKIREELNTIEINKTIEKINQTKSWFFEKVNKIDKPLAKLTKKKKERAQISKIRKENGEI
L1-1_DN AINSYIKKEERAKIEELTAHLKELEKQQQSNPTGRRRKEITKIRAELNEIENKKALEKINKTKSWFFEKINKIDKPLARLTKKKREKMQIHKIRNEKGDI
L1-1_Cpo ALSSHIRKTERIQINNLMLHLKQLEKEEQVKPKAKRREEIIKIRAEINAIETKKTIQRINESKSWFFERINKIDKPLANLI.KREEKAQIHAIRNEKGEI
L1-1_SSc AIQAHLRKQEKAQINKLTLHLKQLEREEQTRPKVSRRKEIIKIRAEINEIETKKTIEKINETKSWFFEKINKIDKPLARLIKQKRERTQINKIRNEKGEV
L1-1B_Cho ALNAYIKKEERAKIKELMEQLKKLENEQQTNPKPSRRKEITRIKAEINDIENKKTIERINNTKSWFFEKINKIDKPLARLTKSKREKTHINKIMNEKGDI
L1-2_EC AIQAHLNKQEKSQISNLKAHLTELEKKEQMKPKVSRRREIIKIRAEINTIETKKAVERINETKSWFFEKINKIDKPLARLTKKKREKAQINKIRNERGEI
L1MAB2_ML ALQAYLKKQEKMVVNHLTLQLKELEREQQENPRVSRRKEIIKIRAEINDIETKKTIQKINETKSWFFERINKIDKPLARLTKKQRERSQINQIRNDRGEI
L1-1_Str SWSSFLKKRKNQQINELTLHLKNLEKEEQNNSKCSRRQEIIKIRAEINEIETKKTIEKIDKTKSWFFEKINKIDKPLAMLTKRRRERTQITNIRDEKGNI
L1-1_MD SLKAHINKQGRAEINQLEMQLKKLESDQIKNPQQKTKLEILKIKGEINKIESDRTIDLINKTRSWYFEKTNKIDKVLVNLIKKRKEEKQIHSIKDEKGDS
L1-1_TS SLQTHMRKMEGTEIDNLTSHLKKLEKQDHKNPNFSRRIQITKIKAQIQDIEDKKIIQKINETKSWFFERVNKIDGPLARLTKKKREKNQISTIRNTKDEV
L1-1_OP AIGAHVKAQERRQIQELNTHLQELEKQQQKSPTHNRKQEIIKTREEINQIEIKKTIHKINESKSWFFEKINKIDTPLARLTKKKQEKARINSIKDEKGNI
L1_RN ALSACRKKQERAYVSSLTAHLKALEQKEANTPRRSRRQEIIKLRAEINQVETKRTIERINRTKSWFFEKINKIDKPLARLTRGHRECVQINKIRNEKGDI
L1A_Mim SINAYNQKTRRSQIDNLMKRLKELEKEEQTNPKPSRRSEINKIKSELNEIENREAIQEINKTKSWFFEKINKIDTPLAKLTKSRKEKSLISSIRNKKGDI
L1-1_Cja AISAHMRRVERSKIDTLSSKLKELEEQDQKNSKPSRRQEITKIRAELKEIETRKTLQKINKSRSWFFEKINKIDRPLARLIKKKRENNQIDAIKNDKGEI
L1-1_EE AIQAHIRKQEKAQINSLIAHLKDLEEEQQRNPKATRRTEITKVRAEINNIENRKTIQKINESKCWFFERVNKIDKPLARLTKQKREKTQINRIVNERGDI
L1-1_Tbel AISAHKKRMERWQVDNLSSCLRELEKQQQTNPQNARRKEIIKVRAEINEIENKRTIQKINESKSWFFEKINKIDTPLARLIKKKKERNQINQICDETGNI
L1-1_Pca SINAHIWKEERNQIKELTLQLEQVEREQQRNPSGTRRNEIIKIRAELNELENRKTIEKIEKTKSWFFEKINKIDKPLAKLTKEKQERKQITQIRNERGVV
L1A_OC SIGAYIKKLERHQIDELSIHLKDLENLQQTRPKSSRRREIIKIREEINRIESRKTLQKISQTRSWFFEKINKIDTPLAQLTKKRREKTQINKIRDEKGNV
L1HS ALNAYKRKQERSKIDTLTSQLKELEKQEQTHSKASRRQEITKIRAELKEIETQKTLQKINESRSWFFERINKIDRPLARLIKKKREKNQIDTIKNDKGDI
L1-2_Dor ALSSYINKLERSETNNLMTHLNLLEKEQQDKPQSSRRKEIIKIKSELNELESKKTIERINKTKSWFFEKLNKIDRPLANLIRKRREHTQINKLRDENGNI
L1-Y_CF AIQASIQKLERTQIQKLTLHIKELEKKQQIDPTPKRRRELIKIRAELNEIETRRTVEQINRTRSWFFERINKIDKPLASLIKKKREKTQINKIMNEKGEI
L1-1A2_Sar ALQAYLRKEERARINNLTSQLKTLEKDQQKEPKPGRRKEIIKLRAEINDMETQKTIRKINETKSWFFEKINKIDKPLARLTKKERERTLISRIRNEKGDI
L1-1_ET SLNAYMKKEERRMTDTLTQNLQQLEQSQQNHPSNSKRKEIIKIRAELNQWEDKRTIQKINAAKSWFYERINKIDSPLAKLTKDRKEQTSIARMRDETGAI
L1-1_AMe AIQASLKRIEKSKMQFLYSHLKKLEQQQRDRPNPLTRKELTKIRAEINELETRTTVEQINRTRSWFFERIHKIDRPLAKLVQKQRERTEIIKIMTEKGEV
Meug SLNAYINKIEKEEINDLGLQLKKLEKEQIENPQVNTKLEILKTKGEINKIEIKKTIELINKTNSWFYEKTNKIDKPLVNLIKKKKEENQITNIKNERGEL
Fcat AIQAYLKKQEKSQIQNLTAHLKELEAEQQRQPKPSRRREIIKIRAEINNIESKKTVEQINETKSWFFEKINKIDKPLARLLKKKREMTQIDKIMNENGII
Mmul ALNAHKRKQERSKIDTLTSQLKELEKQEQTHSKASRRQEITKIRAELKEIETQKTLQKINESRSWFFEKINKIDRPLARLIKKKREKNQIDAIKNDKGDI
Pham ALNAHKRKQERSKIDTLTSQLKELEKQEQTHSKASRRQEITKIRAELKEIETQKTLQKINESRSWFFEKINKIDRPLARLIKKKREKNQIDAIKNDKGDI

```
                 ---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
                          410       420       430       440       450       460       470       480       490       500
                 ---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
Conserved sites  XXXXXXIXXXXXXXXXXXLYXXXXXXXXXXXXXXXXXXXXXXXLXXXXXXXXNXPIXXXEXXXXIXXXXXXKXPGPDGXXXEFXXXXXXXXXPXLXXXXXXIX
lineage 1(L1-2_PVa) TTDIAEIQRIIQEYYEKVYNTKFNNLEEMDQYLEKYNLPRLNQEELENLNRPISSMEIETIIKNLPKSKSPGPDGFTSEFYQTFKEDLIPILLKLFQKIE
lineage 2(L1-1_PVa) TTDSTEIQWIIRKYYEQLYANKLDNLEEMDTFLETYNLPRLSQEETENLNRPITTNEIESVIKNLPKNKSPGPDGFIGEFYQTFKEELSPILLKLFQKIQ
L1-BT            TTDNTEIQRIIRDYYQQLYANKMDNVEEMDKFLEKYNFPKLDQEEIENLNRPITTMEIETVIKNLSANKSPGPDGFTAEFYRKFREELTPILLKLFQKIA
L1-1_LA          TTEPNEIKRIISDYYEKLYSNKFXNLEEMDEFLEKHYLPKLTHSEVEQLNRPITKKEIETVIKKLPTKKSPGPDGFTAEFYQTFREELTPLLLKVFQSIE
L1-1_Vpa         TTNKIEIQNIIREYYEKLYGTKLDNLEEMDKFLETYCPPKLNQEETEHLNNPITRKEIEIAIKNLPTNKSPGPDGFTGEFYQTYKEELIPVLLKLFQTIE
L1-1_DN          TTDPTEIKTIIRGYFEKLYSNKNDNLEEMDKFLETHKQPILTKEEIDDLNKPITSREIESVIKNLPTKKSPGPDGFTGEFYKTFRKELTPILLKLFQKIE
L1-1_Cpo         TTDPIEIQKIINTYFENLYSQKFDNTEEIDRFLETYEVPKLDQEDVKLLNNPISVNEIENVIKSLPTKKSPGPDGFTAEFYKKYKEDLMPTLLKLFNEIE
L1-1_SSc         TTDTTEIQRIIRDYYMQLYANKMENLEEMDKFLEKYNLPRLNQDEIEKMNGPITRTEIETVIKKLPTNKSPGPDGFTGEFYQTFREELTPLLLKLFQKIA
L1-1B_Cho        TADPEEIXKIIRGYYEQLYANKLDNVEEMDNFLETYEQPRLTREEIEDLNQPITSKEIQSVIKKLPTNKCPGPDGFTGEFYQTFQKELTPILLKLFQNIE
L1-2_EC          TTDSAEIQQIIREYYKKLYANRMDNLEEMDKFLDSYNLPKLTQEEADNLNRPITRKEIETAIKNIPKNKTPGPDGFPGEFYQTFREDLIPILFKLFQKIR
L1MAB2_ML        TTDPTEIQMIVKQYYGQLYSNKLDNLEEMDKFLEKYNIPKLNQEESKNLNRPITMEEIEAVIRKLPANKSPGPDGFTGEFYQTFKEELKPTLLRLLQKIQ
L1-1_Str         TTDTTEIQKIIRKYFETLYSNKIEDSEDIDKFLKAYDLPRLRQEDTHNLNRPITKEEIEEAIKRLPTKKSPGPDGYTAEFYKTFKEELIPILFKLFQEIE
L1-1_MD          TSNEEEIKAIIRNYFAQLYGNKYTNLGEMDEYIQKYKLPRLTEEEIEFLNNPISEIEIHQAIKELPKKKSPGPDGFTCEFYQTFREQLTPILYKLFDIIS
L1-1_TS          TSDPEEIQKIIRDYYVHLYGNKLENQKEMEDFLTSHNLPRLEQEEIETLNRPITIKEIDHVIRKLPTKKSPGPDGFPAEFYKTFKEELIPILLKVFQAIE
L1-1_OP          TTDTASIKAIIRNYYKALYSNKSEDHQEMEKFLDFYHLPKLSPEATNDLNKPITEAEIESVIKDLPTKKSPGPDGFTTEFYKTFRTELTPILYKLFKTIE
L1_RN            TTDSEEIQKIIRSYYKNLYSTKLENLQEMDNFLDRYQVSKLNQEQINQLNNPITPKEIEAVIKGLPTKKSPGPDGFSAEFYQTFIEDLIPILSKLFHKIE
L1A_Mim          TTDPKEIQDTIYEYYKNLYAHKLENVEEMDKFLETHSLPRLNQEEIDSLNRPISTAEIETAIKNLPKKKSPGPDGFTPEFYHTYKEELVPILQKLFHNIE
L1-1_Cja         TTDPTEIQTIIREYYKQLYAHKLVNLEEMDKFLDTCVLPSLNQEEVETMNRPITRSEVEAAIKSLPHKKSPGPDGFTAEFYQTHKEELLPFLLKLFQIIQ
L1-1_EE          TTDTAEIQHIMRGFYEQLYATKLENLEEMNDFLDTYQLPKLSKEEVDNMNRPITANEIETVIKNLPKNKSPGPDGFTNEFYKTFKEELIPLLLKVFQKIE
L1-1_Tbel        TTDIAEIHNIIRGYYERLYANKQDNLEEMDKFLDAYKLPRLDQEDIENLNRPITSQEIESVINNLPTKKSPGPDGFTCEFYKTFKEELTPILLKLFLKIQ
L1-1_Pca         TTDATEIKNIITEYYERLYSNKFENLGEMDXFLDSYHLPKLSQREAEQLNKPITKEEIEKVIKTLPTKKSPGPDGFTAEFYQAFREELSPILLRLFHSIE
L1A_OC           TTDTTEIKRIIRNYYKDLYASKQGNLSEMDRFLDTCNLPKLNQEDIENLNRPITETEIETVIKALPTKKSPGPDGFTAEFYQTFKEELTPILLKLFRTIE
L1HS             TTDPTEIQTTIREYYKHLYANKLENLEEMDTFLDTYTLPRLNQEEVESLNRPITGSEIVAIINSLPTKKSPGPDGFTAEFYQRYKEELVPFLLKLFQSIE
L1-2_Dor         TTETTKIQNIIREYFSKLYANKFENLTEMDLFLANIDMPKLNQDELNYLNRPISSIEIETAIRDLPAKKSPGPDGFTAEFYKAFKAELTPILLQLFNEIE
L1-Y_CF          TTNTKEIQTILKTYYEQLYANKLGNLEEMDAFLESHKLPKLEQEEIENLNRPITREEIEAVIKNLPRHKSPGPDGFPGEFYQTFKEEIIPILLKLFGKIE
L1-1A2_Sar       ATETNEIQKIIRDYFENLYATKQENLEEMDKFLDSYNLPRLNQEDLEXLNSPINIKEIETVIKSLPKNKSPGPDGFTSEFFQTFKEDLLPVLLKLFQEIE
L1-1_ET          TTDPNEIKRIITKYYEGLYSNEFRNMEDMDKYLEKQSLPRLSQTEIKNLNKPIAKEEIERVIKNLPTKKAPGPDGFTAEFYQAFREELTPILHKVFHNIE
L1-1_AMe         TTSTIEIARIIRNFYQQLYAKKLNNLEEMEAFLETYKLPRLKQEEIDFLNRPINYEEIESVINNLPNNKTPGPDGFPGEFYQTFKEEIIPILLKLFQKIE
Meug             TSNEEEIKTIIRNYFAQLYAHKFDNLNEMDEYFKKYKLPRLTEEEVEYLNNPISEKEIEQAINELPRKKSPGPDGFTSEFYQTFKEQLIPILHTLFLKIG
Fcat             TTNPSEIQTIIREYYEKLYANKLDNLEEMDKFLNTHTLPKLNQEEIESLNRPITSEEIESVIKNLPTNKSPGPDGFPGEFYQTFKAEIIPILLKLFQEIE
Mmul             TTDPTEIQTTIREYYKHLYANKLENLEEMDNFLDTYTLPRLNQEEVESLNRPIAGSEIEAIINSLPTKKSPGPDGFTAEFYQRYKEELVPFLLKLFQSIE
Pham             TTDPTEIQTTIREYYKHLYANKLENLEEMDNFLDTYTLPRLNQEEVESLNRPIXGSEIEAIINSLPTKKSPGPDGFTAEFYQRYKEELVPFLLKLFQSIE
```

```
                 ---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
                          510       520       530       540       550       560       570       580       590       600
                 ---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
Conserved sites  XXXXLPXXFXXXXXXLXPKXXXXXXXXXENXRPXSXMNXDAKXXXXILAXRXQXXIXXXXHXDQXGFIXXXQXXXNXRKXIXXXXXIXXXXXXXXXXSXD
lineage 1(L1-2_PVa) EEAILPNSFYEANITLIPKPGKDNTKKENYRPISLMNTDAKILNKILANRIQQHIKKIIHHDQVGFIPGAQGWFNIRKSINVIHHINKLKNKNHMIISID
lineage 2(L1-1_PVa) EEGRLPNSFYEATITLIPKPGKDITKKENYRPISLMNIDAKILNKILANRIQQYIKNIIHHDQVGFIPGMQGWXNIRKSINVIHHINKSKDKNHMIISID
L1-BT            EEGKLPNSFYEATITLIPKPDKDPTKKENYRPISLMNIDAKILNKILAIRIQQHIKKIIHHDQVGFIPGMQGFFNIRKSINVIHHINKLKNKNHMIISID
L1-1_LA          NDGILPNSFYEATISLIPKPGKDITKKENYRPISLMNIDAKILNKILANRIQQHIKKIIHPDQVGFIPGMQGWFNIRKTINVIHHINKTKDKNHMILSID
L1-1_Vpa         KEGILPNSFYEATITLIPKPGKDTTKKENYRPISLMNIDAKILTKILANRIQQHIKKIIHHDQVGFIPGTQGWFNIRKSINVIHHINKRKDKNHMIISID
L1-1_DN          TEGTLPNSFYDANITLVPKPNKDTTRKENYRPISLMNLDAKILNKILANRIQQHIKRIIHHDQVGFIPGMQGWFNIRKSINVIHHINRLKEKNHMIISID
L1-1_Cpo         REAILPKSFLEANITLVPKPEKDPTKKENYRPISLMNTDAKILNKILANRMQQIIKKIIHHDQVGFIPGMQGWFNIRKSINVIHHINKAKNKNHMIISID
L1-1_SSc         EEGILPNSFYEATVTLVPKPDKDSTKKENYRPXSLMNIDAKILNKILANRIQQYIKRIVHHDQVGFIPGMQGFFNIRKSISVIHHINKLKNKNHMILSID
L1-1B_Cho        ENGTLPNTFYEANINLIPKPGKDATKKENYRPISLMNIDAKILNKILANRIQRHIKKIIHHDQVGFIPGMQGWFNIRKSINVLQHINKSKEKNQMIISID
L1-2_EC          EDGTLPNTFYEANITLIPKPDKDTTKKENYRPISLMNIDAKILNKILATRIQQFIKRIIHQDQVGFIPGTQGWFNIRKSINVIHHINKLRNKNHMIISID
L1MAB2_ML        EEGTLPSSFYEASITLIPKPGKDITMKENYRPISLMNIDAKILNKILANRIQQYIRRIIHHDQVGFIPEMQGWYNIRKSINVIHHINKLKEKNHMVISID
L1-1_Str         KEGALPNSFYEANITLIPKPDKDTSKKENYRPISLMNLDAKILNKILANRIQKHIRKLVHHDQVGFIPGMQGWFNIRKSINAIHHINRLKDKNHMIISID
L1-1_MD          KEGVLPNSFYDTNMVLIPKPGRSKTEKENYRPISLMNIDAKILNRILAKRLQQVIRRIIHHDQVGFIPGMQGWFNIRKTIHIIDHINKQTSKNHMIISID
L1-1_TS          KDGTLPKSFYEANITLIPKPGKDPTKKENYRPISLMNIDAKILNKILANRIQQYISKIIHHDQVGFIPGMQGWFNIRKTINVIKYINRCQNKNHMIISLD
L1-1_OP          KEATLPNSFYEANITLIPKPDRELTEKENYRPISLMNIDAKILNKILANRIQKHIRQIVHPDQVGFIPGMQGWFNIRKSINVIHHIQKLKNKNHMIVSID
L1_RN            TDGALPNSFYEATITLIPKPHKDTTKKENFRPISLMNIDAKILNKILANRIQEHIKTIIHHDQVGFIPGMQGWFNIRKTINVIHYINKLKEQNHMIISLD
L1A_Mim          KNGNLPDTFYEANITLIPKPGKDATKKENYRPISLMNIDAKIFNKILANRIQTLIKKIIHHDQVGFIPGMQGWFNIRKSINAIHHINRSKNKDHMILSID
L1-1_Cja         KEGILPKSFYETNIILIPKPGRDSTRKENFRPISMMNIDAKIFNKILASRLQQHIKKLIHHDQVGFIPGMQGWFNIRKSINVIHHINRTKNKNHMIISID
L1-1_EE          DTGILPASFYEANITLIPKADRDTTKKENYRPISLMNIDAKILNKILANRIQQYIKKIVHHDQVGFIPGMQGWFNIRKSINVIHHINKSKTKNHMVISID
L1-1_Tbel        KEGNLPNSFYEASIILIPKTDKDMTKKENYRPISLMNIDAKILNKILANRIQQHITKIIHHDQVGFIPGMQGWFNIRKTINVIHHINRKKDKNHMIISID
L1-1_Pca         RDGVLPNSFYEASITLIPKPGKDPTKKENYRPISLMNLDAKIFNKILANRIQQHIKKIIHHDQVGFIPGMQGWFNIRKTINVINHINNTKDKDHLILSID
L1A_OC           EEGILPNSFYEASITLIPKPEKDAALKENYRPISLMNIDAKILNKILANRMQQHIRKIIHPDQVGFIPGMQGWFNVRKTINVIHHINRLQKKNHMIISID
L1HS             KEGILPNSFYEASIILIPKPGRDTTKKENFRPISLMNIDAKILNKILANRIQQHIKKLIHHDQVGFIPGMQGWFNIRKSINVIQHINRAKDKNHMIISID
L1-2_Dor         RESSLPNTFYEAIITLIPKPGRDSSKKENYRPISLMNIDAKLLNKILANRLQQFIQKIIHHDQTGFISGMQSWFNIRKSINIIHHINRSKNKNHMVISID
L1-Y_CF          RDGVLPNSFYEASITLIPKPDKDPAKKENYRPISLMNMDAKILNKILANRIQQYIKKIIHHDQVGFIPGTQGWFNTRKTINVIHHISKRKTKNHMILSLD
L1-1A2_Sar       ETETLPNSFYEAHISLIPKANKDTTKKENYRPISLMNTDAKILNKILANRIQQLIKKIIHHDQVGFIPGMQGWFNIRKSINIIHHINKRKDKNHIIISID
L1-1_ET          EDGKLPNSFYEASITLIPKQGKDPTGIENYRPISLMNIDAKILNKILANRIQKYIKHIIHHDQVGFIPGMQGWFNIRKSINIIHHIEKKKDKNHMIISID
L1-1_AMe         TEGKLPNSFYEANITLIPKPGKDPLKKENYRPISLMNMDAKILNKILANRIQQYIKRIIHHDQVGFIPGMQAWFNTRKSINVIHHINKKRLKNHMILSID
Meug             EEGVLPNSFYDTNMVLIPKPGRDKTEKENYRPISLMNIDAKILNKILAKRIQHLITRLIHYDQVGFIPGLQGWFNIRKTISIIDHINNKANKNHMIISID
Fcat             REGKLPDSFYEASITLIPKPDRDPVKKENYRPISLMNMDAKILNKILANRIQRHIKRIIHHDQVGFIPGMQGWFNIRKSINVIHHINKKKEKNHMILSID
Mmul             KEGILPNSFYEANIILIPKPGRDTTKKENFRPISLMNIDAKILNKILANRIQQHIKKLIHHDQVGFIPGMQGWFNIRKSINIIQHINRTKDKNHMIISID
Pham             KEGILPNSFYEANIILIPKPGRDTTKKENFRPISLMNIDAKILNKILANRIQQHIKKLIHHDQVGFIPGMQGWFNIRKSINIIQHINRTKDKNHMIISID
```

```
                 ---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
                          610       620       630       640       650       660       670       680       690       700
                 ---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
Conserved sites  AEKAFDXXQXXFXXXXLXXXGXXGXXXXXXXXXYXXPXAXIXXNGXXXXXXXXXXGXXQGCPLSPXLFXXVLEXLAXXXRXXXXIXGXXXXXXEXKXSLX
lineage 1(L1-2_PVa) AEKAFDKIQHPFMIKTLNKIGIEGKYLNIIKAIYDKPSANIILNGEKLKTFPLRSGTRQGCPLSPLLFNITLEVLASAIRQEKDIKGIQIGNEEVKLSLF
lineage 2(L1-1_PVa) AEKAFDKIQHPFMIKTLSKVGIEGTYLNIIKAIYDKPTANIILNGERLKAFSLRSGTRQGCPLSPLLFNIVLEVLATAIRQEREIKGIRIGKEEVKLSLF
L1-BT            AEKAFDKIQHPFMIKTLQKAGIEGTYLNIIKAIYDKPTANIILNGEKLKAFPLKSGTRQGCPLSPLLFNIVLEVLATAIRAEKEIKGIQIGKEEVKLSLF
L1-1_LA          AEKAFDKVQHPFMIKTLTKIGIEGKFLNIIKGIYAKPTANITLNGENLKAFPLRTGTRQGCPLSPLLFNIVLEVLARAIRLDKEIKGIRIGKEEVKLSLF
L1-1_Vpa         AEKAFDKIQHPFMIKTLAKVGIEGTYLNIIKAIYDKPTASIVLNGEKLKSFPLKSGTRQGCPLSPLLFNIVLEVLATAVRQEKEIKGIQIGKEEVKVSLY
L1-1_DN          AEKAFDKIQHPFLIKTLQKIGIQGNFLNMIKSIYEKPKANIVYNGEILKSFPLNSGTRQGCPLSPLLFNIVLEVLARALRQEPDIKGIQIGKEEVKISLF
L1-1_Cpo         AEKAFDKVQHLFMIRTLQKIGIDGLYLNIIKAIYDKPTASIILNGQKLKAFTLKSGTRQGCPLSPLLFNIVLEVLARAIRQEREIKGVKIGKEEVKLSLF
L1-1_SSc         AEKAFDKIQHPFLIKTLQKVGITGTYLNMIKAIYDKPTANIILNGEKLKEFPLRSGTRQGCPLSPLLFNIVLEVLATAIREVKEIKGIQIGKEEVKLSLF
L1-1B_Cho        AEKAFDKIQHPFLIKTLQKVGIEGNFLNMIKSIYEKPTASIVLNGERLKAFPLRSGMRQGCPLSPLLFNIVLEVLARAIRQDKEIKGIQIGKEEVKLSLF
L1-2_EC          AEKAFDKIQQPFMIKTLNKMGIEGNYLNIIKAIYDKPIANIILNGQKLNPIPLKTGTRQGCPLSPLLFNIVLEVLARAIRQEKGIKGIQIGREEVKLSLF
L1MAB2_ML        AEKAFDKIQHPFLIKTLRKVGIEGSYLNIIKAIYDRPTANIILNGQKLTPFPLRTGTRQGCPLSPLLFNIVLEVLAIAIRQEEEIKGIQIGKEEVKLSLF
L1-1_Str         AEKAFDKVQHPFMFKTLEKLGITGTYLNIVKAIYAKPQASIILNGEKLKAFPLKSGTRQGCPLSPLLFNLVLEVLARAIRQTKEIKGIKIGKEEIKLSLF
L1-1_MD          AEKAFDKIQHPFLLKTLESIGIEGSFLKIINSIYLKPTANIICNGDKLDAFPIRSGVKQGCPLSPLLFDIVLETLAVAIREDKEIEGIRIGKEETKLSLF
L1-1_TS          AEKAFDKIQHPFLIKTLEHLGIRGTYLKIVKAIYEKPTASILLNGQKLEPIPLKTGTRQGCPLSPLLFNIVLEVLARAIREEEAIRGIQIGKEEVKLSLY
L1-1_OP          AEKAFDKIQHHFLLKTLTKVGIDGKIHNIIKAIYEKPNASIILNGEKLEPFPLRSGTRQGCPLSPLLFNIVLEVLAEAIRQEKEIRGIQMGNEEVKLSLY
L1_RN            AEKAFDKIQHPFMIKVLERIGIQGPYLNIVKAIYSKPVANIKLNGEKLEAIPLKSGTRQGCPLSPYLFNIVLEVLARAIRQQKEIKGIQIGKEEVKISLF
L1A_Mim          AEKAFDKIQHPFMIRTLKKIGIEGTYLKMIQAIYDRPIANIILNGERLKSFPLRTGTRQGCPLSPLLFNIVLEVLATAIRQENGIKGIQIGAEEIKLSLF
L1-1_Cja         AEKAFDKIQQPFMLKTLNKLGIDGTYLKIIKAIYDKPTANIILNGQKLEAFPLKSGTRQGCPLSPLLFNIVLEVLARAIRQEKEIKGIQIGKEEAKLSLF
L1-1_EE          AEKAFDKIQHPFMIKTLQKMGIDGKFLKIVESIYSKPTANIILNGEKLEAFPLRSGTRQGCPLSPLLFNIVLEVLAIAIRQEQGIKGIQIGREEVKLSLF
L1-1_Tbel        AEKAFDKIQHPFMIKTLSKIGIEGSFLNTIKAIYEKPTANIILNGEKLKAFPLKTGTRQGCPLSPLLFNIVLEVLAIAIREENSIKGIRIGKEELKISLF
L1-1_Pca         AEKAFDKVQHPFMLKTLSKMGIDGKFLNTIRAIYTKPIAHIILNGERLKDFPLRSGTRQGCPLSPLLFNIVLEVLARAIRQDQEIKGICIGKEEVKISLF
L1A_OC           AEKAFDKIQHPFMMKTLSKLGLEGTFLNTIKAIYEKPTANILLNGEKLEAFPLRSGTRQGCPLSPLLFNIVLEVLARAIRQEKEIKGIQIGKEELKLSLF
L1HS             AEKAFDKIQQPFMLKTLNKLGIDGTYFKIIRAIYDKPTANIILNGQKLEAFPLKTGTRQGCPLSPLLFNIVLEVLARAIRQEKEIKGIQLGKEEVKLSLF
L1-2_Dor         AEKAFDKIQHPFMMKVLEKLGIHGTFLNTIKAVYDKPTASIILNGEKLKPFPLNSGTRQGCPLSPLLFNLVLECLATAIRQDADIKGIQIGKEEIKLSLF
L1-Y_CF          AEKAFDKIQHPFLIKTLQSVGIEGTFLDILKAIYEKPTANIILNGEALGAFPLRSGTRQGCPLSPLLFNIVLEVLASAIRQQKDIKGIQIGKEEVKLSLF
L1-1A2_Sar       AEKAFDKIQHPFMMKTLAKMGIEGTFLNIVKAIYHKPTASIVLNGEKLRAFPLRTGTRQGCPLSPLLFNIVLEVLAIAIRQEKDIKGIQVGKEEIKLSLF
L1-1_ET          AEKAFDNIQHPFLIKTLMKIGLEGKFLKLIQAIYERPTANIIVNGEKTKTIPLKKGTRQGCPLSPLLFNIVLEVLANNIRQRKDIKGIQMGEEEVKLSLF
L1-1_AMe         AEKAFDKIQHPFLIKTLQSVGIEGTFLNLIKAIYEKPTASIILNGEKLEAFPLRSGTRQGCPLSPLLFNIVLEVLATAIRQQKGIKGIQIGKEEVKLSLF
Meug             AEKAFDKIQHPFLLKTLENVGIKGTFHKIISSIYLKPSASIICNGDKLDAFPIRSGVKQGCPLSPLLFNMVLEMLAVAIRQDKDIQGIRIAKEETKLSLF
Fcat             AEKAFDKIQHPFLIKTLEKVGIEGTYLKIIKAIYEKPTANIILNGEKLRAFSLRSGTRQGCPLSPLLFNIVLEVLASAIRQQKEIKGIKIGKDEVKLSLF
Mmul             AEKAFDKIQQPFMLKTLNKFGIDGTYLKIIRAIYDKPTANIILNGQKLEKFPLKTGTRQGCPLSPLLFNIVLEVLARAIRQEKEIKGIQLGKEEVKLSLF
Pham             AEKAFDKIQQPFMLKTLNKFGIDGTYLKIIRAIYDKPTANIILNGQKLEKFPLKTGTRQGCPLSPLLFNIVLEVLARAIRQEKEIRGIQLGKEEVKLSLF
```


710 720 730 740 750 760 770 780 790 800
Conserved sites ADDMXXYXXXPXXSXXXXXXXXXXXXXXVXGYXINXXKXXXXXXXXXXXXXXXXXXXXXXFXXXXXXXXYLGXXLXXXXXXLXXXNXXXXXXXXXXXXXXWX
lineage 1(L1-2_PVa) ADDMILYIENPKDSTRTLLETISKYSKVSGYKINVQKSIAFLYSNNEVSEKEVEKIIPFAIATKRIKYLGINLTKHVKDLYNENYKTLLKEIEEDTKKWK
lineage 2(L1-1_PVa) ADDMILYIENPKDSTKKLLELINEFSKVAGYKINIQKSVAFLYTNDELSEREIKKTIPFTIASKTIKYLGIKLTKKVEDLFSENYKSLKKEIEEDTNKWK
L1-BT ADDMILYIENPKDSTRKLLEIINDYSKVAGYKINTQKSLAFLYTNNEKTEREIKETIPFTIATERIKYLGIYLPKETKDLYLENYKTLVKEIKEDTNRWR
L1-1_LA ADDMIXYTENPKESSRKLLKLIEEFGRVSGYKINIQKSLGFLYINKKNTEEEITKSIPFTVAPKKIRYLGINLTKDVKDLYKENYKALLQEIQKDILKWK
L1-1_Vpa ADDMLLYIENPKRSTQKLLELIEEFSKVAGYKINVQKSVAFLYTNDKSTEKESKETIPFKIAPKVIKYLGINLTKEAKELYTENYKPLMKEIKEDFKKWK
L1-1_DN ADDMILYIENPERSTTKLLELINEFSKVAGYKINAQKSVAFLYTNNEQDQEEIKKQIPFTIVNKKIKYLGINLTKEVKNLYTENYTRLFKEIKEDLNKWK
L1-1_Cpo ADDMILYIEDPLNSIERLLDTINKFSNVAGYKINTQKSIAFLYTNNKITEREIRETALFTLASKKMKYLGITLTKEVKDLYSENYNTLKKEIEDDLRKWK
L1-1_SSc ADDMILYLENPKDSTRKLLELIHEFGKVAGYKINTQKSIAFLYTNNEKAEKEIREAIPFTIASKRIKYLGVNLPKETKDLYSENYKPLMKEIKDDTNRWK
L1-1B_Cho ADDMILYLENPEKSMIQLLELINKFSKVAGYKINAHKSVMFLYARNERSEETLKKKIPFSIATKKIKYLGINLTKDVKDLYKENYITLLKEIERDLKRWK
L1-2_EC ADDMILYIENPKESIGKLLEVINNYSKVAGYKINLHKSVAFLYSSNEPTEKELKNTIPFTIATKRIKYLGVNLTKEVKDLYNENYKAFLRELDDDIRRWK
L1MAB2_ML ADDMILYIQNPRDSIKKLLDLIHEFGNVAGYKINPKKSEAFLYTNSELSEREIRKTIPFTIAPKKLRYLGINLTKEVKDLYSENYRTLKKEIEEDINRWK
L1-1_Str ADDMIIYLTDPKGSTKKLLELINEFSKVAGYKINTHKSKAFLYISNKTSEMEIRKTTPFTISSKKIRYLGINLTKEVKDLYNENYRTLKREIEEDLRRWK
L1-1_MD ADDMMVYLKNPRDSTKKLIEIINNFSKVAGYKINPHKSSAFLYISNTAQQQELEREIPFKITLDKIKYLGIYLPRQTQELYEHNYKTLATQLKLDLNNWK
L1-1_TS ADDMXVYLENPRESVKGLLTLIKAFGKVSGYKINVQKTIAFLYTNNKQTETQIKNTVPFTIATKKMKYLGIFLTRDVKDLYNENYKTLLKEIKADTNKWK
L1-1_OP ADDMILYVEEPRDSIQRLLELVREFGRVAGYKINEQKSTAIVYANSPKMEKDLTSKIPFKITEKSMKYLGINLTKNVGDLFEENYKLLKKEIEQDLKRWS
L1_RN ADDMIVYLSDPKSSTRELLKLINNFSKVAGYKINSNKSVAFLYTKEKQAEKEIRETTPFIIDPNNIKYLGVTLTKQVKDLYNKNFKTLRKEIEEDLRRWK
L1A_Mim ADDMILYLENPKDSTKKLLELINEFSKVSGYKINTQKSEAFIYANNNLIENQIKDSIPFTIATKKLKYLGIYLTKEVKDLYRENYETLRKEIAEDVNRWK
L1-1_Cja ADDMIVYLEDPIVSAQNLLKLISNFSKVSGYKINVQKSQAFLYTNNRLKESQIKNELPFTIATKRIKYLGIQLTRNVKDLFKENYKPLLNEIREDTNRWR
L1-1_EE ADDMIVYMEKPKESSKKLLEIIRQYSNVSGYKINIQKSVAFLYANTKLEEIEIQKSVPFSIATKTIKYLGVNLTKEVKDLYTENYESLLKEIEKDTKKWK
L1-1_Tbel ADDMMVYLEDPKVSMKKLLEVISEYSKVAGYKINIQKSIAFLFMNNKLNEAELKKGIPFTVATRCIRYLGINLTKKVKDLYKENYENLKKDIESDIKKWK
L1-1_Pca ADDMIVYAENPKESISKLLKLIEEFSKVSGYKINISKSVGFLYTNNETVKEEIIKSIPFTIAPKKIKYLGINLTRETTDLYKENYKPLLQETKRDLDKWK
L1A_OC ADDMILYLGDPKNSTKRLLELIEEFGKVAGYKINAQKSTAFVYTDNAMAEEELLRSIPFTIATKTIKYLGINLTKDVKDLYDENYKTLKKEIEEDTKKWK
L1HS ADDMIVYLENPIVSAQNLLKLISNFSKVSGYKINVQKSQAFLYTNNRQTESQIMGELPFTIASKRIKYLGIQLTRDVKDLFKENYKPLLKEIKEDTNKWK
L1-2_Dor ADDMILYLKNPIDSTPKLLKLIQNFGKVAGYKINPQKSMAFLYANNEKSVAEIRKATPFVIAPQKIKYLGINLTKEVKDLYDENFKILKKEINTELKKWQ
L1-Y_CF ADDMILYIENPKVSTPRLLELIQQFGSVAGYKINAQKSVAFLYTNNETEEREIKESIPFTIAPKSIRYLGINLTKDVKDLYPQNYRTLLKEIEEDTKRWK
L1-1A2_Sar ADDMILYLENPKTSTKKLLETIDSYSKVAGYRINTQKSMAFLYANNEREESDMXKAIPFTIAPQKIKYLGISLTKEVKDLYNENYKTLLQEIKEDTRKWK
L1-1_ET ADDMILYIENPKSSTAGVLTAIEEYGRVAGYKINKQKSVGFLYTSDRTMEEGIKKEVPFTVAKNKLKYLGIYLTKNTKDLYKENYKTLLQETKSDLHKWK
L1-1_AMe ADDMILYMENPKESTPKLLEVIEQFSKVAGYKINAQKSVAFLYTNNETEEREIRESIPFTITPKTMRYLGINLTRDVKDLYARNYRSLLKDIEEDIKRWK
Meug ADDMMIYLENPRDSSKKLLELINNFGKVAGYKINPHKSSAFLYISNKVQQQEIEREIPFKVRVDSIKYLGVYLPKQTQGLYEHNYKTLFAQIKSDLSKWK
Fcat ADDMILYMENPIDSTKSLLELIQEFSKVAGYKINVQKSVAFLYTNNEATERQIKKLIPFTIAPRSIKYLGINLTKDVKDLYAENYRKLMKEIEEDLKKWK
Mmul ADDMIVYLENPIVSAQNLLKLISNFSKVSGYKINVQKSQAFLYTSNRQTESQIRNELPFTIASKRIKYLGIQLTRDVKDLFKENYKPLLSEIKEDTNKWK
Pham ADDMIVYLENPIVSAQNLLKLISNFSKVSGYKINVQKSQAFLYTSNRQTESQIMNELPFTIASKRIKYLGIQLTRDVKDLFKENYKPLLSEIKEDTNKWK

```
                 ---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
                          810       820       830       840       850       860       870       880       890       900
                 ---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
Conserved sites  XXXCXWXGXXNXXKMXXLPXXXYXXXAXPIXXXXXXXXXXXXXXXXXFXWXXXXXRIXXXXXXXXXXXXGGXXXPXXXXYYXAXXXKXXWYWXXXRXXDXWX
lineage 1(L1-2_PVa) DIPCSWIGRINIVKMTILPKALYRFNAIPIKIPSAFFKEIEQKIIRFVWKHKRPRIAKAILRKKNEAGGITLPDFKLYYKATVIKTAWYWQQNRHTDQWN
lineage 2(L1-1_PVa) HIPCSWVGRINIVKMAILPKAIYRFNAIPIKIPMAFFRELEQIILKFIWSHKKPRIATAILRKKNKVGGITLPDIKLYYKATVIKTAWYWHKNRYIDQWN
L1-BT            NIPCSWIGRINIVKMSILPKAIYRFNAIPIKLPTVFFTELEQIISQFIWKYKKPRIAKAILRKKNGAGGINLPDFRLYYRATVIKTVWYWHKDRNMDQWN
L1-1_LA          NIPCSWIGRLNIVKMSILPKAIYTXNALPIQIPMSYFKGIEKQITNFIWKGKKPRISKALLKKKKKVGGLTLPDFRTYYTATVVKTAWYWYNNRHIDQWN
L1-1_Vpa         DIPCSWIGRINIVKMVTLPKAIYGFNAIPIQLPRTYFTELEQIIIKFIWNHQRPRIAKALLKRKKEAGGITLPDFRQYYRATVIKTAWYWYQNRHIDQWN
L1-1_DN          NIPCSWIGRLNIIKMSILPKLIYTFNAIPIKINAAFFKELEKLTMKFIWKGKRPRIAKDILKKKNEIGGITLPDFKTYYKATVVKTAWYWHKERHTDQWN
L1-1_Cpo         DIPCSWIGRTNIVKMAILPKLLYRFNAIPIKIPSTYLIDLEKSLLNFIWNQKRPRIAKAILSSKDKAGGITIPDLKLYYKATVVKSTWYWNQNRAEDQWN
L1-1_SSc         DIPCSWIGRVNIIKMTILPKAIYRFNAIPIKLPRTFFTELEQNILKFVWKHKRPRIAKDILKKKNGAGGIRLPDFRLYYKATVVKTAWYWHKDRHIDQWN
L1-1B_Cho        NIPCSWIGRLNVIKMSILPKLIYRFNAIPIKIPTTYFADLEKLVIKFIWKGKMPRIAKNTLKKKNEVGGLTLPDFEAYYKATVVKTAWYWHKDRYIDQWN
L1-2_EC          DIPCTWIGRINIVKMSILPKAIYRFNAIPIRIPMTFFTELEQRILKFIWGNKRPRIAKAILRKKNKTGGITIPDFKTYYKATVIKTAWYWYKNRCTDQWN
L1MAB2_ML        NIPCSWIGRINIIKMSILPKAIYKFNALPIKIPMAFFKDLERTLQKFIWNKKRPRIAAAILRK-NKVGGISMPDIKLYYKATVLKTAWYWHKNRHIDQWN
L1-1_Str         NVPCSWIGRTNIIKMAILPKVLYRFNAMPIRIPXAFLVEIDKAIMKFIWKNKRPRIAKAILSRKCESGGIAIPELKLYYKAIVTKTAWYWYQNRRVDQWY
L1-1_MD          NINCSWIGRANIIKMTILPKLIYLFSAIPIELPKYFFTDLEKTITKFIWKNKRSRISREIMKKNTYDGGLAVPDLKLYYKAAVIKTIWYWLRNRKEDQWN
L1-1_TS          NIPCSWIGRINIVKMSILPKAIYKFNAIPIKLPTTFFSDLEKTTQEFIWKHKRPRIARTILSKKNKAGGITIPDFKLYYKATIIKTAWYWYRNRHIDQWN
L1-1_OP          NIPCSWIGKINIIKMSILPKAIYTFNAIPIKLPKTFFTELETMIQRFIWKHKKPRIARTILKNRKLAGGITVPDLWTYYRAVVIKTAWYWHKDREEDQWS
L1_RN            DLPCSWIGRINIVKMAILPKAIYRFNAIPIKIPIQFFKELDRTICKFIWNNKKPRIAKAILNNKRTSGGITIPELKQYYRAIVIKTAWYWYRDRQIDQWN
L1A_Mim          SIPCSWIGRLNIIKMSILPKLIYRFNAIPIKIPSAFFTDIEKIILRFVWNQRRPRISRAILGNKNKMGGINMPDIKLYYKAVVIKTIWYWHKNRNIDQWN
L1-1_Cja         NIPCSWLGRINIVKMAILPKVIYRFNAIPIKLPMTFFTELEKTTLNFIWNQKRARIAKSILSKKNTAGGITLPDFKLYYKATVIKTAWYWYQNRDIDQWN
L1-1_EE          DIPCSWVGRINIIKMNILPRAIYKFNAIPIKIPSTFFRRIEKMLQMFIWNQKRPRIAKTILRKKNRTGGITLPDLKLYYRAIVIKTAWYWNMNRHTDQWN
L1-1_Tbel        DIPCSWIGRINIIKMAILPKLIYMFNAIPIKIPEIFFRDLEKTIIEFVWNHKRPRIAKAILRKKNGVGGISLPDPKIYYKAVVIKTAWYWHKNRHVDQWK
L1-1_Pca         NLPCSWIGRLNIVKMSILPKAIYLYNAIPIQVPASFFNEMEKQITNFVWKGKKPRISKVLLKKKNTLGGLSLPDLRTYYTATVVKTAWYWYKNRQIDQWN
L1A_OC           NLPCSWIGRINIIKMSILPKAIYRFNAIPIKIPKTFFSDLEKMVLKFIWRHKRPRIAKAILYNKNKAGGITIPDFRTYYRAVVIKTAWYWYRNRWIDQWN
L1HS             NIPCSWVGRINIVKMAILPKVIYRFNAIPIKLPMTFFTELEKTTLKFIWNQKRARIAKSILSQKNKAGGITLPDFKLYYKATVTKTAWYWYQNRDIDQWN
L1-2_Dor         SLPCSWIGKINIVKMAILPKAIYKFNAIPIKIPTSFFKEIEEAIQKFIWNNKRPRIAKTLLNGKNTAGGISIPNFKLYYKAIVIKTAWYWHKNRPQDQWN
L1-Y_CF          NIPCSWIGRINIVKMSMLPRAIYTFNAIPIKIPWTFFRELEQIILRFVWNQKRPRIARGILKKKTISGGITMPDFRLYYKAVVIKTVWYWHKNRHIDQWN
L1-1A2_Sar       DIPCSWIGRINIVKMAILPKALYKFNAIPIGIPLXFFKEMEQALLKFIWNNKPPRIAKAILGKK-KMGGINLPNFQLYYKAVVIKTAWYWNKGRAADQWN
L1-1_ET          NIPCSWIGRLNIVKMTILPKALYKFNAIPIQIPSTFFKELEKLTTNFIWRGKKPRISRELLKKKDTVGGLALPDFNAYYTATVVKTAWYWHNDRHSDQWK
L1-1_AMe         NIPCSWIGRINIVKMSILPRAIYTFNAIPIKIPRTFFKELEQIVLKFVWNQKRPRISKELLKRKNKAGGITMPDFELYYKAVITKTAWYWHKNRHIDQWN
Meug             NISCSWVGRANIIKMTILPKLIYLFSAIPIKLSDNYFLELDKIISKFIWKNKRSRISKGLMKRNAWEGGLALPDLKLYYKAAIIKTTWYWLRNREVDKWN
Fcat             DIPCSWIGKINIVKMSILPKAIYTFNAIPIKIAPAFFSKLEQAILKFIWNHKRPRIAKGILKKKTKAGGITIPDFSLYYKAVIIKTAWYWHKNRHIDQWN
Mmul             NIPCSWIGRINIVKMAILPKVIYRFNAIPIKLPMSFFTELEKTALKFIWNQKRARISKTILSQKNKAGGITLPDFKLYYKATVTKTAWYWYQNRDIDQWN
Pham             NIPCSWIGRINIVKMAILPKVIYRFNAIPIKLPMSFFTELEKTALKFIWNQKRARISKTILSQKNKAGGITLPDFKLYYKATVTKTAWYWYQNRDIDQWN
```

```
                 ---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
                          910       920       930       940       950       960       970       980       990       1000
                 ---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
Conserved sites  XXXXXXXXXXXXXXXXXXDXXXXXXXXXXXXXFXKXXXXXWXXXXXXXXXXXXXXPXTXXXSXWXXXLXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
lineage 1(L1-2_PVa) RIESPEVNPYLYGQIIFDKKAKNIQWRKESLFNKWCWENWKATCKRMKLDCCLSPHTKINSKWIKDLNIRPETIKCIEENIGTKLKDLGLKEDFMN-LTS
lineage 2(L1-1_PVa) RIESPEINPRLYAHLIYDKGGKNIQWGKDSLFNKWCWENWTDTCKKKKLDHLLTPYTRIDSKWIKDLNVRPETIKLLEENIGSKLSDITNRNFFSD-ISP
L1-BT            KIESPEINPRTYGHLIFDKGGKDIQWIKDNLFNKWCWEIWSTTCKRMKLDHFLTPYTKINSKWIKDLNVRPETIKLLEENIGKTLSDIYHSRILYD-PPP
L1-1_LA          RIENPDINPSTYEQLIFDKGPVSVNWGKDSLFNKWCWHNWISICKKMKQDPYLTPCTKTNSKWIKDLNIKTKTIKIMEEKIGTTLGALIQGINRIQNITK
L1-1_Vpa         RIESPEMNPQTFGQLIFDKGGKNIQWNKDSLFSKWCWENWTAACKTMKLEHTLTPYTKINSKWIKDLNIRQDTINLLEENIGKTLSDIHFKNFLLE----
L1-1_DN          RIESSDIEPHIYSHIIFDKATKPSQLGESGLFNKWCLENWIAICRRMKEDYHLTPYTKINSRWIKDLNIRAKTIKTLESSVGKHLQDLVIG-NGFMNITP
L1-1_Cpo         RLEDTTTTTNTLNHLIFDKGAKQVHWKNDSLFNKWCWKNWLSICRKLKLDPCLSPCTKLKSKWVKDLNIKTETLNLLEDKLGRNLEDIGVGREFMN-RTQ
L1-1_SSc         RIESPELNPRTYSQLIYDKGGKNIQWRKDSLFNKWCWENWTATWKRMKLEHSLTPYTKINSKWIKDLDIRPDTIKLLEENIGQTLSDINDSNIFSD-PPI
L1-1B_Cho        RIENSEIDPQIYGRLIFDKAPKATELGHNSLFNKWGWESWISISKRMKEDPYLTPYTKINSKWIKDLNIKDSTIKLLEDNVGRHLQDLVLG-GHFLDLTP
L1-2_EC          RIESPEIKPHIYGQLIFDKGAEGIQWRKESLFNKWCWENWKATCKRMKIDHSFSPFTKINSKWIKDLKVRPETIRLLEENVGSTLFDISIKRIFSDTMPS
L1MAB2_ML        RIESPEISPNQYAQLIFDKGGMNIQWSQDSLFNKWCWENWTDICKKMKLDHQLTPYTRINSKWIKDLNVRQETIKILEEYKGNKISDICQNNFFTD-TAP
L1-1_Str         RIEDTETNPQSYNYLIFDKGAKNMQWRKDSIFNKWCWENWKSICNKMKLNPFLSPATKVNSKWIKELDIKXETLRLIEGKVGYDLHIVGSGSKFLN-RTP
L1-1_MD          RLGEND-----LSKTVYDKPKDPSFWDKNPLFDKNCWENWKTVWERLGIDQHLTPYTKINSKWVSDLNIKKETISKLGKHRIVYMSDLWEGKGFKT-KQD
L1-1_TS          RIEIPEAKPQFLNQLIFDKAPTTYHWGEENLFSKWCWENWLTTCRRLKQDPYLSPCTKVNSKWIRDLNVKPQTIRTLEK-EGNTLMEIGTGIQFLY-KTR
L1-1_OP          RIETPEGNPHRYSQIIFDKKTNDNPGKWEGLFNKCCWDNWLIACRNKKIDPHLSPYTKIRSKWITDLNLHPETFKLLEENVGNTLQHLGVGPHFLK-KTP
L1_RN            RIEDPEMNPHTYGHLIFDKGAKTIQWKKDSIFSKWCWFNWRATCRRMQIDPCLSPCTKLKSKWIKDLHIKPDTLKLIEEKLGKHLEHMGTGKNFLN-KTP
L1A_Mim          RCENPDIKPSSYSHLIFDKADKNIRWGKESLFNKWCWENWIATCRRLKQDPHLSPLTKTNSRWITDLNLRYETIRTLEEKVGNTLLDIGLGKEFMK-KSP
L1-1_Cja         RTEASEATQHIYNHTIFDKPDKNKQWGKDSLFNKWCWENWLAMCRKQKLDPFLTPYTKINSRWIKDLNIRPNTIKTLEENLGKTIQDIGVGKDFMT-KTP
L1-1_EE          RIESPEMRPHTYGHLIFDKGAQTITWGKQSLFNKWCWKQWVETCRRMKLNHCISPNTKVNSKWIKDLDVRPETIRYLEENIGRTFFRINFKDIFNE-TNP
L1-1_Tbel        RIETPEITPKAYSQLIFDKGYQSIHWEKENLFSKWCWKNWVSTCRKMKLNPHLSPLTTVNSKWIKDLNLRPQTIKLLVENAGETLGDIGVGKDFLN-KTP
L1-1_Pca         RIETPEMTPSAYEQLIFDKGPKTLNWGKNSIFNKWCWQNWLSICNKLNQDPYLTPYTKTNSNWIKDLNIKPKTIKLIQEKVGTTMEVLIHNEDRLQNLSK
L1A_OC           RIETPEINPNIYSQLIFDQGSKTNSWSKDSLFNKWCWENWISTCRIMKQDPYLTPYTKIHSTWIKDLNLRPDTIKLLEN-IGETLQDIGTGKXFLE-KTR
L1HS             RTEPSEIMPHIYNYLIFDKPEKNKQWGKDSLFNKWCWENWLAICRKLKLDPFLTPYTKINSRWIKDLNVRPKTIKTLEENLGITIQDIGVGKDFMS-KTP
L1-2_Dor         RIENPEMNLQTYSHLIFDKGAKTIEWKKDSLFSKWCWQNWLTTCRKLKLDPYISPCTRINSKWIKDLEVKADTLKTLQDGIGETLGLLGSGRNFLN-RDP
L1-Y_CF          RIENPEVDPELYGQLIFDKGGKTIHWKKDSLFNKWCWENWTSTCRRMKLDHSLSPYTKINSKWMKDLNVRQDSIKILEKNTGNTLFELGHSNFLQD-TST
L1-1A2_Sar       RVEYSDTHPQIYDHLIFDKGARNVKWSKESMFNKLCWQNWTATCKKMGLDLHLSPCTKVRSKWIKDLNIRPESLRYIEDKVGKTLHDIEAKGIFKA-DTP
L1-1_ET          RIESPRIKPSAYRQLIFDKGPKTIKWEADALFNKWCWKQWISTCRKMKQDVYLTPCTRISSRWITDLEVKPQTIRTIKEGIGTNLRALAQGTHRLTAIGK
L1-1_AMe         RIENPEMDPRLFGQLIFDKAGKNIRWKKDSLFNKWCWENWTATCKRMKLDHSLTPYTKINSKWMKDLNVRQESIKILEENIGNNFYDIGQSNLFHD-TSP
Meug             RLGTQDAVGKEYSNLLFDKPKDPSFWDKNSLFDKNCWENWITVWRKLGIDPYLTPYTRIKSKWVHDLGIKIDTMNKLEKQGIVYLSDLWRREEFFT-KGE
Fcat             RIETPELDPQTYGQLIFDKAGKNIQWKKDSLFNKWCWENWTATCRRLKLDHFLTPFTKINSKWIKDLNVRQETIKTLEEKAGKDLSDLSRSNLLLD-TSP
Mmul             RTESSEIIPHIYSHLIFDKPERNKKWGKDSLFNKWCWENWLAISRKLKLDPFLTPYTKINSRWIRDLNVRPNTIKILEENLGSTIQDIGMGKDFMS-KTP
Pham             RTESSEIIPHIYSHLIFDKPERNKKWGKDSLFNKWCWENWLAISRKLKLDPFLTPYTKINSRWIRDLNVRPNTIKILEENLGSTIQDIGMGKDFMS-KTP
```


1010 1020 1030 1040 1050 1060 1070 1080 1090 1100
Conserved sites XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXWEXXXXXXXXDXXLXXXXXXXXXXXX XXXXXXXXXXXXXXXWXXXXXXXXXXXXXXXXXXXXX
lineage 1(L1-2_PVa) KAREVKAKINEWDYIKLKSFCSAKETVNKVKRQPSEWENIFASNASDKGLISKIYKELIRLNNNKKTND-----PIKKWAEDLNRHFSLEDIQMANKYMK
lineage 2(L1-1_PVa) RARETKEKINKWDYIKLKSFCTAKETINKTKRQPTVWEKIFANDTSDKGLISKIYKELIQLN-NKKTNN-----PIKKWAEDLNRYFSKEEIQMANRHMK
L1-BT RIMEIKAKINKWDLINLKSFCTSKETISKVKRQPSEWEKIIANEATDKQLISKIYKRLLQLN-SRKIND-----PIKKWAKELNRHFSKKDIQMANKHMK
L1-1_LA ---NDEEKPDNWELLKIKHLCSSKDFTKRVKRPPTDWERIFSYDISDQRLISKIYMILSKLN-HKKTNN-----PIKKWAKDMNTHFTKEDIQAANRYMR
L1-1_Vpa ---EIKARINKWDLMKLTSFCRAKETRNKTRRKPTEWEKIFASET-DKGLISKIYKQLIRLN-KKKINN-----PIQKWAEDLNKQFSKEDIQMIKKHMK
L1-1_DN KARAAKELIDKWDFLKIKAFCTSKEFVKKVKREPTQWEKIFGNHISDKKLITCIYKELLYLE-NKKINN-----PFKKWEKDLNRHFSKEEIQMAKKHMK
L1-1_Cpo TAQEILPRINNWDHFLLKSFCMSKEISSIVKRKPTNWEKILVNSLSDKGLLSKTYKELKKLR-PPKFKD-----PIQKWASEMNTHFSDEEMQMANKYMK
L1-1_SSc RVLTIKRKINKWDLIKLQSFCTAKETLNNTKRQPTEWEKIFASESTDKGLISKIYKXLLQLH-TKKTNN-----PIKKWAEDLNRQFSKEDIQMAEKHMK
L1-1B_Cho KAQATKEKIDKWELLKLRSFCTSKEFVKKVKRQPTQWEKIFGNHVSDKRLISCIYKEILQLN-DNSTDS-----PIIKWAKDMKRQFSEGEIQMAKKHMK
L1-2_EC QRRETIERINKWDFIRLKSFFKANENRIETKKQPTNWEKIFASHISDKGLISLIYKELSQLN-HKTSNN-----PIKKWAGDMNRHFSKEDILMANRHMK
L1MAB2_ML RAVETKEKMNKWDYIKIKSFCTAKETINKTTRKPTSWENIFANVISDKGLISKIYRELIQLN-KKKINN-----PIKKWAKDLNRHLSKEDIQKAKRHMK
L1-1_Str VAQELITRINKWDLLKLKSFFSARETVKEVNREPTSWEQIFTPHTSDRALISRMYKELKKLN-NKITNN-----PINKWAKDLNRHFSEEDIQSINKYMR
L1-1_MD IERITKCKINNFDYIKLKSFCTNKTNITKIRRETTNWEKIFIETS-DKGLITHIYNELNQLY-KKSSHS-----PIDKWAREMDRQFSDKEIKTINKHMK
L1-1_TS NPQDLREKIDKWDLIKLTSFCKAKETIKRAGRQPTDWEKVFANSRSDKGLTSWIYKELKRAE-KKKTNN-----PIIKWAKDMNRHFTKEDIRAANKHMK
L1-1_OP NAVEIKTKINNWDLIKLRSFCTARETINKVKRQPTEWEKIFAHDIGDRGLISRIYKELQNNQ-NVKTNK-----PLKKWAREMGKHFTKEQTQMANKHMK
L1_RN MAYALRSRIDKWDLIKLQSFCKAKDTVVRTKRQPTDWEKIFTNPTTDRGLISKIYKELKKLD-RRETNN-----PIKKWGSELNKEFTAEECRMAEKHLK
L1A_Mim KAITAATKINKWDMIKLQSFCTAKEIVMKVNRQPTEWEKIFASYASDKGLITRIYLELTKIR-KKKSNN-----PIKKWAKDLNRNFSKEDRRMANKHMK
L1-1_Cja KALATKAKIDKWDLIKLHSFCTAKETVIRVNRQPTEWEKIFAVYPSDKGLISRIYKELKQIY-KKKTNK-----PIQKWAKDMNRHFTKEDIHEANKHMK
L1-1_EE ITRKTKASINLWDYIKLKSFFTAKETTTQIKRPLTEWEKIFTCHTSDKSLITNIYKELARLN-NKTTNN-----PIQKWGEDLDRIFTTEEIQKAEKHMK
L1-1_Tbel KAQAIIPKIDKWDYIRLRSFCTAKDTVTSVNRQPSEWENIFVKYASDKGLITRIHRELKCLM-RKKMIN-----PIKRWESELNKSLSKEDIRTAKKHMR
L1-1_Pca KPQSTEEKIDKWELLRIKHFCSSKDFINRIKRTPTEWEKIFGNSMSDRRLISIIYKILQNIN-KKKINH-----PIKRWAKEMDRHFTKEEIQAAKKHMR
L1A_OC EAQAVKAKINYWDCIKLRSFCTAKETVRRVKRQPTEWEKIFANYATDKGLITRIYKEIKKLH-KNKTNN-----PLKRWAKDLNRHFSKEEIQMANRHMK
L1HS KAMATKAKIDKWDLIKLKSFCTAKETTIRVNRQPTTWEKIFATYSSDKGLISRIYNELKQIY-KKKTNN-----PIKKWAKDMNRHFSKEDIYAAKKHMK
L1-2_Dor ETQQIKERLDKWDCIKLQSFCRANDIASKINRKPTDWEKIFTSHTTDKGLISKIYLELKKLN-PPTTNPQRNNCPINKWAKDLKRNFSEEERRMANRHMK
L1-Y_CF KAKETKAKMNYWDFIKIRSFCTAKDTVNKTQRQPTEWEKIFANDISDKGLVSKIYKELIKLN-TKETNN-----PIMKWAKDMNRNLTEEDIDMANMHMR
L1-1A2_Sar LAKQVKTEINKWDYLKLRSFCTSRETVTKIQRQSTEWERIFTQYPSDKGLITRIYNALVELH-KKKTAN-----PIKKWGDEMNRNFPKEEIRMAERHMR
L1-1_ET GTHAGELEIDKWDLIRIKHLCTSKDFTKRVTRQPTDWERIFSNDTSDKGLITKIYNTLMACK-KRKTVN-----PLRKWAKELKRAFTREETXMANKHMR
L1-1_AMe KAREIKDKMNLWDFIRIKSFCTAKETVKKTKRQPTEWENIFAKDTTDKGLVSKIYKELLKLN-TRETNK-----QIIKWAEDMNRHFSNEDIQMANRHMK
Meug IECIMKCKMDNFDYIKLRSFCTTKPNATKIRRDVVNWERIFTAKLGDKGLISRIYRELTQMY-NHTSHS-----PIDKWSKDMNRQFSEEEIKAIYNHMK
Fcat KARELKAKVNYWDLMKIKSFCTAKETTNKTKRQPTEWEKIFANDISDKGLVSKIYKELTKLH-TRKTNN-----PVKKWAENMNRHFSKEDIRMANRHMK
Mmul KATAAKAKIDKWDLIKLKSFCTAKETTIRVNRQPTEWEKIFAIYSSDKGLISRTYKELKQIY-KKKTNN-----PIKKWAKDMNRHFSKEDIHTANRHMK
Pham KATAAKAKIDKWDLIKLKSFCTAKETTIRVNRQPTEWEKIFAIYSSDKGLISRTYKELKQIY-KKKTNN-----PIKKWAKDMNRHFSKEDIHTANRHMK

```
                 ---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
                          1110      1120      1130      1140      1150      1160      1170      1180      1190      1200
                 ---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+
Conserved sites  XCSXSLXXXEXQIKTXXRXHXTXXRXXXXXXXXXXXXCWXGCXXXGXLXHCWWXCXXXXPXWXXXWXXXXXXXXXXXPXXPXIXLLGXXXXXXXXXXXXXX
lineage 1(L1-2_PVa) RCSTSLAIREMQIKTTMRYHLTPVRIAIINRTSNNKCWRGCGEKGTLIHCWWDCKLVQPLWKTVWRFLK-KLRIDLPYDPAIPLLGIYPKDLKTHIRKDI
lineage 2(L1-1_PVa) RCSRSLIIREMQIKTTMRYHLTPVRMAIINKSKN-KCWRGCGEKGTLIHCWWDCKLVQPLWKTVWRFLK-KLKMDIPFDPAIALLGIYPKKTKTIIRKDL
L1-BT            RCSTSLIIREMQIKTTMRYHFTPVRMAAIQKSTNNKCWRGCGEKGTLLHCWWECKLVQPLWRTVWRFLK-KLEIEPPYDPAIPLLGIHTEK--TRRERDT
L1-1_LA          KCSRSLAIREMQIKTTMRFHLTPTRLALIQKTQNNKCWRGCGEIGTLIHCWWECKMVQPLWKSIWRYLK-QLEIELPYNPEIPLLGIYPRETRAFTQTDI
L1-1_Vpa         KCSISLIIREMQIKTTMRYHLTPVRMAVIQKSXNDKCWRGCGERGTLLHCWWECSLVQPLWKTVWRFLK-RLGIDLPYDPGIPLLGLYPEG--NLLQDDT
L1-1_DN          KCSKSLAIREMQIKTTMRYHLTPIRLAAMKKTEEYKCWRGCEERGTLIHCWWECRRIQPFWRTVWRFLK-KLAIDLPYDPAIPLLGIYPAELKTRTQTDI
L1-1_Cpo         KCSSSLVIREMQIKTTLRYHLTPERMARIKKTNNNKCWRGCGEKGTLLHCWWECRLVQPLWRSVWRFLK-KLGLEVPFIPAIPLLGIFPEELKASYHSDI
L1-1_SSc         RCSASLIIREMQIKTTMRYHLTPARMAIIQKSTNNKCWRGCGEKGTLVHCWWDCKLVQPLWKAVWRFLR-KLNIELPFDPAIPLLGIYPEK--TTTRKDT
L1-1B_Cho        KCSASLAIREMQIKTTMRYHLTPIRMAAIKQTGNYKCWRGCGEIGTLIHCWWDCIMVQPLWKTVWQFLR-KLDIELPFDPAIALLGIYPEDLKAVTRTDI
L1-2_EC          RCSSSLIIREMQIKTTLRYHLTPVRMTKISKTNSNKCWRGCGEKGTLIHCWWECKLVQPLWKTVWRFLK-KLKIELPYDPAIPLLGIYPKSLKSAIPKVL
L1MAB2_ML        TCSKSLIIREMQIKTTMRYHLTPVRMAIINKSTNDKCWRGCGEKGTLVHCWWECRLVQPLWKAVWRFLK-KLEMELPFDPVIPLLGIYPKKPETPIRKDI
L1-1_Str         KCSPSLAIREMQIKTTLRYHLTPVRLAAIMKSNNNKCWRGCGEKGTLVHCWWDCKMVRPIWKAVWRFLG-KLGMDPPFDPAIALLGLFPEDLKRAHYRDT
L1-1_MD          KCSTSLIIREMQIKTTLRYHLTPSRLANITAKESNECWRGCGKVGTLIHCWWSCELIQPFWRAIWNYAQRATKEYLPFDPAIALLGLYPKEIMDT---KT
L1-1_TS          KCSTSLIIREMQIKTTLRYHLTPVRMAIINNSKNNSCWRGCGEKGTLLHCWWECKLVQPLWKAVWRFLK-ALKIDLPYDPAIPLLGIYPEEHKSLYKKDT
L1-1_OP          KCSSSLAIREIQIKTSMRYHLTPVRLAHMNKSTNNTCWRGCGEKGSLLHCWWGCRLVQPLWKSVWRIFK-QLKFNIPYDPAIALLGIYPEH--LFYEKPT
L1_RN            KCSTSLVIREMQIKTTLRFHLTPVRMAKIKNSGDSRCWRGCGERGTLLHCWWDCRLVKPFWKSVWRFLR-KLDIELPEDPAIPLLGIYP-KDASTYKRDT
L1A_Mim          KCSTSLIIREMQIKTTMRYHLTPVRMAFIKKSPNNKCWRGCGERGTLLHCWWDCKLVQPLWKAIWRYLK-AIQVNLPFDPAIPLLGIYPNDPVTLYKKDT
L1-1_Cja         KCSSSLVIREMQIKTTLRYHLTPVRMAIIKKSGDNRCWRGCGEIGTLLHCWWECKLVQPLWKTVWRFLK-DLEIEIPFDPAIPLLGIYPKDYKSFYYKDT
L1-1_EE          KCSKSLIVREMQIKTTMRYHFTPVRMSHIRKGNSSKCWRGCGVKGTLLHCWWECQLVQPLWRTVWRTLR-RLEMDLPYDPAIPLLGIYPKEPNTSIQKDL
L1-1_Tbel        KCSTSLIIREMQIKTTMRYHLTPVRMAKIKNIKSDKCWRGCGEKGTLLHCWWECKLVQPLWRTIWRFLK-DLQIELPFDPAIPLLGIYPKDRRTLSQKSA
L1-1_Pca         TCSRSLVIKEIQIKTTMRYHLTPTRLAMIQKTQNNKCWRGCGESGTLLHCWWGCKMVQPLWKAIWRFLK-KMDIDIPFDPVIPLLGIYPTESRPFTPTDT
L1A_OC           KCSRSLAIREMQIKTTMRFHLTPVRMAHXQKSTNNRCWXGCGEKGTLTHCWWECKLVKPLWKSVWRFLR-NLNITLPFDPAIPLLGIYPKEFKLANKKAV
L1HS             KCSSSLAIREMQIKTTMRYHLTPVRMAIIKKSGNNRCWRGCGEIGTLLHCWWDCKLVQPLWKSVWRFLR-DLELEIPFDPAIPLLGIYPKDYKSCCYKDT
L1-2_Dor         KCSTSLAIKEMQIKTTLRYHLTPVRTAKIKNSNNNRCWRGCGQKGTLLHCWWECKLVQPLWKAVWRFLK-RLHIELPYDPAIPLLGIYPNVYKQDHTKAT
L1-Y_CF          KCSASLAIREIQIKTTMRYHLTPVRMGKINKAGNNKCWRGCGEKGTLLHCWWECELVQPLWKTVWRFLK-QLKIYLPYDPAIALLGIYPKDTNAMKRRDT
L1-1A2_Sar       KCSTSLIIREMQIKTTMRYHLTPQRLAHIPKNKSNRCWRGCGEKGTLLHCWWECRLVQPFWKTIWTILK-KLEIELPFDPAIPLLGIYPGEAKRYSRDGI
L1-1_ET          KCSRSLAIREMQIKTTMRYHLTPLRLAQISKSESNKCWRGCGEVGTLLHCWWSCRFVQPLWKAIWRYLK-KMEIDLPYDPAIPLLGIYPVEATNKTRPVI
L1-1_AMe         KCSKSLAIREIQIKTTLRYHLTPVRMAKIDKARNNNCWRGCGERGSLLHCWWECKLVQPLWKTVWRSLK-KLKIELPYDPAIALLGVYPKDTDVVKRRAI
Meug             KCSKSLLVREMQIKTTLRYHITLXRLANMTEQENDKCWRGCGKVGTLIHCWWSCERIQPFWRAIWNYAQRATKMCIPFDPAISLLGLYPQEIIKMGKGPT
Fcat             RCSASLLIREIQIKTTLRYHLTPVRVAKMNKSGDYRCWRGCGETGTLLHCWWECKLVQPLWKAVWRFLR-KLKIDLPYDPAIALLGIYPRDTGVLMHRAT
Mmul             KCSSSLAIREMQIKTTMRYHLTPVRMAIIKKSGNNRCWRGCGEIGTLLHCWWDCKLVQPLWKTVWRFLK-DLELDVPYDPAIPLLGIYPKDYKLCCYKDT
Pham             KCSSSLAIREMQIKTTMRYHLTPVRMAIIKKSGNNRCWRGCGEIGTLLHCWWDCKLVQPLWKTVWRFLK-DLELDVPYDPAIPLLGIYPKDYKLCCYKDT
```

```
                  ---------+---------+---------+---------+---------+---------+---------+---------+-----
                           1210      1220      1230      1240      1250      1260      1270      1280
                  ---------+---------+---------+---------+---------+---------+---------+---------+-----
Conserved sites   XXXXXXAAXXXXAXXWXXXXXPXXXXWXXKXWXXXXXEXYXXXXXXXXXXXXXXXXWXXXXXXXXXXXXXXXXXXXXXXXXXX
lineage 1(L1-2_PVa) CTPMFIAALFTVARTWKQPKCPTIDDWLKKLWYIYTMEYYSAIRRD-EILPFATTWMDLEIIVLSEISQTEKVENHMISLICGI.
lineage 2(L1-1_PVa) CTPIFMAAQFTIAKIWKQPKCPSTDDWIKKLWFIYTMEYYSAIKKN-EILPFATTWMDLEVIMLSEISQTEKDKYHMISLICGI.
L1-BT             CTPMFIAALFIIAKTWKQPRCPSADEWIRKLWYIYTMEYYSAIKKN-TFESVLMRWMKLEPIIQSEVSQKEKHKYSILTHIYGI.
L1-1_LA           CTPMFIAALFTIAKSWKQPRCPSTDEWVNKLWYIHTMEYYASIKNSDESVKHFITWRNLEGIMLSEISQRQKDKYCIRPLL.
L1-1_Vpa          CTPMFIAALFTIAKTWKQPKCPSTGDWIKKMWYIYTMEYYSAIKTD-NITPFAATWMLLENVILSEVSQKEKEKYHMRSLICGI.
L1-1_DN           CTPMFIAALFTIAKSWNQPKCPSTDEWINKMWYIHTMEYYSAVRTNT-LQTHVITWMNLENLMLSEATQALKDKYYMTSMI.
L1-1_Cpo          CAPMFIAAQFVIARSWKQPKCPSTEEWIKKLWYFYTMEYYSAIKKD-HIEIFINKWAQLETILISEINQSRMCEYRIVSLM.
L1-1_SSc          CTPMFIAALFTIAKTWKQPKCPSTEEWIQKMWYIYTMEYYSAIKKN-EIPAFLATWMDLETIMLSEVSHTMRHQHQMLSLTCGI.
L1-1B_Cho         CTPMFIAALFTIAKRWKQPKCPSTDEWINKMWYIHTMEYYAAVRRND-VVKHMTTWMNLEDIMLSEISQAQKEKYYMLPLM.
L1-2_EC           CTPMFIAALFTIAKTWKQPKCPSTDEWIKKMWYIYTMEYYSAAKQN-KIIPFAITWMDLERIMLSEISQREKDNLCMTPLI.
L1MAB2_ML         CTQMFIAAQFTIAKIWKQPKCPSVDEWIRKLWYIYTMEYYAAVKKK-ELLPFATAWMELESIMLSETSQSMKEKYHMISLFHG.
L1-1_Str          ATSMIIAAQFTIARLWNQPRCPSIDEWIKKMWHLYTMEYYAALRND-KIIEFAGKWMALEQIMLSEASQALKNKCQMSSLI.
L1-1_MD           CTKIFIAALFVVAQNWKTRGCPSIGEWLNKLWYMLVMEYYCAQRNN-KVEKFHGDWNNLQEVMQSERSRTRRTLYTETNTLWYNR
L1-1_TS           CTRMFIAALFTIARTWKQPCCPSKEDWIKKMWYIYTMEYYAAIKKN-KIMNFAATWMELESIILSDLSQKQRSEYHMFSLI.
L1-1_OP           CTPMLIAAQSVIAKTWKQPKCPSTEDWIRKLWFIYSMEYYSAIKKN-KMQFFVAKWAKLETIMLREMSQSQKVKYHMFALI.
L1_RN             CSTMFIAALFIIARSWKEPRCPSTEEWIQKMWYIYTMEYYSAIKNN-EFMKFVGKWLELENIILSELTQSQKDIHGMHSLISGY.
L1A_Mim           CTRMFIAAQFIIARLWKQPKCPSIQEWINKMWYMYTMEYYSALRNNGDIAHLIFSWLELEPILLSEVSQEWKNKHQIYSPANWY.
L1-1_Cja          CTRMFIAALFTIAKTWNQPKCPSMIDWTGKMWHIYTMEYYAAIKND-EFVSFVGTWMNLENIILSKLTQEQKMKYRMFSLIGGC.
L1-1_EE           CTHMFLAAQFVIAKTWKQPRCPTTDEWLSKLWYIYTMEYYSAVKNG-DFTVFSRSWMDLEKIMLSEISQKQKDEYGMISLSGRS.
L1-1_Tbel         CTSMFIAALFTIAKTWNQPKCPQTDEWIGKVWYIYSMEYYSAIKNN-KTLEFDRKWSDXEDLLLSEVSQAMKERYCMYPLNIYRL
L1-1_Pca          CTRMFIAAQSTIAKRWNQPRCPTTEEWIGKLWYIYSMEHYAMIKNNDDSRKHLLTWFQLEDIMLSESSQRQKDKYCMRPLR.
L1A_OC            CTLMFIAAQFTIAKTWNQPKCPSTVDWIKKLWDMYSLEYYTAVRNN-EIQSFATKWRNLEHIMLSEXSQSQRDKYHMFSLIGDN.
L1HS              CTRMFIAALFTIAKTWNQPKCPTMIDWIKKMWHIYTMEYYAAIKND-EFISFVGTWMKLETIILSKLSQEQKTKHRIFSLIGGN.
L1-2_Dor          STTMFIAALFTIAKIWNQPRCPSVDEWIKKMWYIYTMEFYASIRKN-DIAPFVRKWKNLEKIILSEVSQAQRNIGPMVSLICNI.
L1-Y_CF           CTPMFLAAMATIAKLWKEPRCPTKDEWIKKMWFMYTMEYYSAIRND-KYPPFASTWMELEGIMLSEVSQSEKDKHYMFSFIWGI.
L1-1A2_Sar        CISMFIAALFTIATIWKKPECPKTDDWLKKLWYIYTMEYYVAVRKH-EVMKFAYKWINMESIMLSEMSQKERDRHRKIALICGI.
L1-1_ET           CAPMFIAAQFTIAKTWKQPKFPSIDEWISKLWHIHTMEYYAALKSADDRMKHVASWEELEGIMLSEVSQXQKDRYNMSPLR.
L1-1_AMe          CTPMFIAALSTIAKSWKEPRCPSTDDWIKKLWSIYTMEYYSAIRKN-EFSTFAATWTALEEIMLSEISQAEKDNYHMISLIYGT.
Meug              CTKIFIAALYVVAKNWKSRGCPSIGEWLNKLWYMNVMEYYCAIRND-EQEDFREAWKDLYDLMLSERSRTRRTLCTATTTVCESF
Fcat              CTPMFIAALSTIAKLWKEPKCPSTDEWIKKLWFIYTMEYYVAMRKN-EIWPFVATWMELESVMLSEISHTEKDRYHMVSLLCGS.
Mmul              CTRMFIAALFTIAKTWNQPKCPSVTDWIKKMWHIYTMEYYAAIKKD-EFVSFVGTWMQLETIILSKLSQEQKTKHRMFSLIGGN.
Pham              CTRMFIAALFTIAKTWNQPKCPSVTDWIKKMWHIYTMEYYAAIKKD-EFASFVGTWMQLETIILSKLSQEQKTKHRMFSLIGGN.
```