ENDOGENOUS RETROVIRUS EVOLUTION IN MAMMALIAN GENOMES

by

Xiaoyu Zhuo

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Human Genetics

The University of Utah

May 2017

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Xiaoyu Zhuo**

has been approved by the following supervisory committee members:

Cédric Feschotte	, Chair	**May 3rd, 2016** Date Approved
Ellen Jean Pritham	, Member	**May 3rd, 2016** Date Approved
Nels Christian Elde	, Member	**May 3rd, 2016** Date Approved
Mark Yandell	, Member	**May 3rd, 2016** Date Approved
Bradley Cairns	, Member	**May 3rd, 2016** Date Approved

and by **Lynn B. Jorde**, Chair/Dean of

the Department/College/School of **Human Genetics**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Endogenous retroviruses (ERVs), derived from exogenous retroviruses (XRVs), comprise about 5 to 10 % of most mammalian genomes. We can study retroviral infection which originated millions years ago and understand long term evolution of infectious viruses by working on ERVs.

At the same time, it has been suggested that multiple, new emerging viruses that infect human populations have been come from different bat species, and bats have become recognized as the reservoir of zoonotic viruses. However, we know little about retroviruses in bats. Here, we mined ERVs in the little brown bat genome, and found that the overall ERV amount in the little brown bat is comparable to other mammals. However, we still find hundreds of lineage-specific ERVs in the little brown bat genome.

With identified bat ERVs, we subsequently investigated if there is any related retroviral cross-species transmission and independent endogenization. Using sequence homologous method to search bat ERV sequences against 107 available mammalian genomes, we found highly similar sequences in cat, tiger, and pangolin genomes in addition to related bat genomes. We found the ERV sequence is patchy distributed among mammalian lineages, and their high sequence similarity is incongruent with their host divergence. We also narrowed down the ERV insertion time to 10 to 20 million years ago. To understand how they evolved in different lineages, we investigated their evolution after integration in both bat and cat genomes. In the cat genome, the ERV lost its envelope domain and transformed to intracellular retrotransposon. While in the bat genome, multiple related infectious viruses became endogenized, and, at least in one lineage, the infectious capability has been maintained.

Finally, I developed a computational pipeline and statistical framework which allows our method to be applied to the ERV population of virtually any species. When applied to 53 available vertebrate genomes, the approach identified ERVs previously known to have spread by reinfection in humans, mouse, and pig as well as additional ERV families carrying signature of recent infections in these and other species, including nonhuman primates, revealing their potential for zoonotic transmission.

This dissertation is dedicated to my parents.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First of all, I want to thank all my friends in China who encouraged me to pursue a doctoral degree in the U.S.

I want to thank everyone in the Feschotte and Pritham Lab for creating a friendly environment, particularly Cédric Feschotte for his patience and guidance. Thomas Carter, Ed Chuong, Rachel Cosby, Aurélie Kapusta, John McCormick, Alesia McKeown, Jainy Thomas and former members Clement Gilbert, Ray Malfavon-borja, Claudia Marquez, Sarah Schaack Mahima Varma and Qi Wang, I want to thank all of you for your support and constructive discussion, particularly Thomas Carter for helping me with my dissertation, Rachel Cosby, and Jainy Thomas for correcting my defense presentation.

I want to thank all members of my committee. They have been supportative and very cooperative.

I want to thank Mina Rho for running MGEScan-LTR for me, Saher Sue Hammound and Jingtao Guo in Brad Cairns' lab for providing WGBS methylation data. Every member of Elde lab, I want to thank you for spending hours in participating in joint meetings. I want to thank the Yandell lab, particularly Daniel Ence for weekly meetings, Zev Kronenberg for cooperating on the lncRNA paper, Steve Flygare for statistics consultation. I want to thank Dr. Fred Adler. I learned how to use R and most of my advanced statistics from your class, and you have been helping me with my research since then.

Finally, I want to thank everyone in the Human Genetics Department, it is such a great program.

CHAPTER 1

INTRODUCTION

It has been proposed that retroviruses derived from LTR retrotransposons by acquiring an envelope protein.[1] Acquiring an envelope gives retroviruses the capability to spread between cells and organisms. Similar to other LTR retrotransposons, they must insert their genome into the host's and utilize host transcription machinery to express retroviral genes and replicate. When these extracellular viral particles invade susceptible cells, the envelope proteins interact with specific cellular receptors to initiate the invasion of a new cell. If retroviruses infect a germline cell, the integrated provirus becomes transmittable from generation to generation, and may eventually become fixed in the population as an endogenous retrovirus (ERV).[2] During vertebrate evolution multiple exogenous retroviruses (XRVs) completed this transition and became ERVs in the host genome. Today, sequences derived from ERVs make up 8 % of our genome.[3]

Most of ERV insertions are tolerated by the host, but sometimes they can be deleterious. Moreover, through millions of years of coevolution with their host, some of these insertions become co-opted by their hosts, providing essential cellular functions. The most well-known example of a co-opted retroviral gene is syncytin, an essential gene for placenta development and derived from ERV envelope gene.[4] The placenta is a eutherian mammalian specific tissue that supports embryo development by mediating material exchange between fetus and mother. Surprisingly, different mammalian lineages have their own syncytin genes independently derived from different ERVs.[5] The most popular hypothesis is that an original syncytin gene existed in the common ancestor of all placental mammals but has been repeatedly replaced by new envelope-derived genes in different lineages. Multiple ERVs have been found to be transcriptional active in the placenta, and it has been suggested that the ERV activity facilitated the turnover of syncytin genes during mammal evolution.[6,7]

ERV envelopes have contributed to other cellular and biological processes, including multiple other envelope proteins that may also be captured by hosts to fight against viral infection.[8] Other retroviral genes besides envelopes have also been co-opted. In mouse

genome, Fv1 was likely derived from gag gene of an ancient Class III ERV MERV-L, and experimental evidence demonstrated that Fv1 can inhibit murine leukemia virus (MLV) infection possibly by interacting with retroviral capsids.[9]

Besides these adapted coding genes, there is mounting evidence suggesting noncoding region of ERVs can also be adapted by hosts. Britten and Davidson first proposed that repetitive Deoxyribonucleic acid (DNA) sequence could be used to regulate multiple gene expression under the same regulation signal.[10] Recent technology advances enable us to investigate cellular function of repetitive ERVs. Wang et al. found many p53 binding sites located in ERV sequences.[11] Chuong et al. identified MER41 has been co-opted in the interferon- response pathway.[12] Furthermore, multiple other ERVs have been shown to act as enhancers downstream of many transcription factors.[13]

Noncoding region of ERVs can be co-opted for other functions as well. Long noncoding ribonucleic acid (lncRNA) are loosely defined as mature ribonucleic acid (RNA) length 200 nt without coding capacity. There are 10,000 annotated lncRNA genes in the human genome and their cellular function is being increasingly appreciated. It has been demonstrated that ERVs make up many lncRNAs. For instance, HERV-H and associated LTR7 can act as cis regulatory binding sites for OCT4 and NANOG, producing hundreds of inducible lncRNAs in human embryonic stem cells.[14]

ERVs are not only important fuel for host genome evolution, they also provide an invaluable historical record for virology study. Retroviruses have a much higher mutation rate compared to their mammalian hosts, and their high mutation rate has generally constrained the analysis of retroviral evolution to periods spanning the past few thousands of years. However, once an ancient XRV become endogenzied in the host genome, its evolution slowed down from the retroviral mutation rate to the mammalian host's mutation rate.[15] Since most of ERV insertions are neutral, it is possible to reconstruct computationally and even synthetically infectious viruses millions of years ago knowning host's neutral mutation rate.[16] Therefore, ERVs shed lights on retroviral evolution as well as vertebrate genome evolution. For instance, by studying endogenized lentiviruses in lemur genomes, the origin of lentiviruses has been pushed back to ˜14 millions years ago.[17,18] And we can even reconstruct an infectious retrovirus that disappeared millions years ago based on current ERV sequences.[16]

ERVs are not just relics of past infections, some of them can still produce infectious viral particles. *Emv* and *Xmv* ERVs are endogenied murine leukemia virus (MLV). Most of them have intact open reading frames and are capable of producing infectious viruses.[19] Another

spreading ERV is koala retrovirus, which can transmit vertically as part of the genome and horizontally from individual to individual as infectious virus.[20] Therefore, studying ERVs may also help us preventing potential zoonosis.

Bats are increasingly recognized as a reservoir of zoonotic viruses.[21] To better understand which retroviruses infected bats and which proviruses in the bat genome may still be active, I began my thesis research by mining ERVs in the little brown bat genome. Then I described a cross-species retroviral transmission among bats, cats, and pangolin 10-20 million years ago. At last, I developed a new method to distinguish reinfecting ERVs from others undergo retrotransposition within host genome. We validated our method using the wealth of knowledge available for human ERVs, and identified many potentially reinfecting ERVs in other mammal genomes.

1.1 Bibliography

[1] Malik, H. S.; Eickbush, T. H. *J. Virol.* **1999**, *73*, 5186–5190.

[2] Bannert, N.; Kurth, R. *Annu. Rev. Genomics Hum. Genet.* **2006**, *7*, 149–173.

[3] Lander, E. S. et al. *Nature* **2001**, *409*, 860–921.

[4] Lavialle, C.; Cornelis, G.; Dupressoir, A.; Esnault, C.; Heidmann, O.; Vernochet, C.; Heidmann, T. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **2013**, *368*, 20120507.

[5] Dewannieux, M.; Heidmann, T. *Curr. Opin. Virol.* **2013**, *3*, 646–656.

[6] Chuong, E. B.; Rumi, M. A. K.; Soares, M. J.; Baker, J. C. *Nat. Genet.* **2013**,

[7] Chuong, E. B. *Bioessays* **2013**, *35*, 853–861.

[8] Malfavon-Borja, R.; Feschotte, C. *J. Virol.* **2015**, *89*, 4047–4050.

[9] Yap, M. W.; Colbeck, E.; Ellis, S. A.; Stoye, J. P. *PLoS Pathog.* **2014**, *10*, e1003968.

[10] Britten, R. J.; Davidson, E. H. *The Q. Rev. Biol.* **1971**, *46*, 111–138.

[11] Wang, T.; Zeng, J.; Lowe, C. B.; Sellers, R. G.; Salama, S. R.; Yang, M.; Burgess, S. M.; Brachmann, R. K.; Haussler, D. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 18613–18618.

[12] Chuong, E. B.; Elde, N. C.; Feschotte, C. *Science* **2016**, *351*, 1083–1087.

[13] Sundaram, V.; Cheng, Y.; Ma, Z.; Li, D.; Xing, X.; Edge, P.; Snyder, M. P.; Wang, T. *Genome Res.* **2014**, *24*, 1963–1976.

[14] Kapusta, A.; Feschotte, C. *Trends Genet.* **2014**, *30*, 439–452.

[15] Feschotte, C.; Gilbert, C. *Nat. Rev. Genet.* **2012**, *13*, 283–296.

[16] Dewannieux, M.; Harper, F.; Richaud, A.; Letzelter, C.; Ribet, D.; Pierron, G.; Heidmann, T. *Genome Res.* **2006**, *16*, 1548–1556.

[17] Gifford, R. J.; Katzourakis, A.; Tristem, M.; Pybus, O. G.; Winters, M.; Shafer, R. W. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 20362–20367.

[18] Gilbert, C.; Maxfield, D. G.; Goodman, S. M.; Feschotte, C. *PLoS Genet.* **2009**, *5*, e1000425.

[19] Kozak, C. A. *Viruses* **2014**, *7*, 1–26.

[20] Tarlinton, R. E.; Meers, J.; Young, P. R. *Nature* **2006**, *442*, 79–81.

[21] Brook, C. E.; Dobson, A. P. *Trends Microbiol.* **2015**, *23*, 172–180.

CHAPTER 2

GENOME-WIDE CHARACTERIZATION OF ENDOGENOUS RETROVIRUSES IN THE BAT MYOTIS LUCIFUGUS REVEALS RECENT AND DIVERSE INFECTIONS

Journal of Virology, 2013 vol. 87 (15) pp. 8493-8501. Genome-wide Characterization of Endogenous Retroviruses in the Bat Myotis lucifugus Reveals Recent and Diverse Infections. Xiaoyu Zhuo, Mina Rho, Cédric Feschotte.

JVI Journals.ASM.org

Genome-Wide Characterization of Endogenous Retroviruses in the Bat *Myotis lucifugus* Reveals Recent and Diverse Infections

Xiaoyu Zhuo,[a] Mina Rho,[b] Cédric Feschotte[a]

Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, Utah, USA[a]; Department of Biostatistics and Bioinformatics, Roswell Park Cancer Institute, Buffalo, New York, USA[b]

Bats are increasingly recognized as reservoir species for a variety of zoonotic viruses that pose severe threats to human health. While many RNA viruses have been identified in bats, little is known about bat retroviruses. Endogenous retroviruses (ERVs) represent genomic fossils of past retroviral infections and, thus, can inform us on the diversity and history of retroviruses that have infected a species lineage. Here, we took advantage of the availability of a high-quality genome assembly for the little brown bat, *Myotis lucifugus*, to systematically identify and analyze ERVs in this species. We mined an initial set of 362 potentially complete proviruses from the three main classes of ERVs, which were further resolved into 13 major families and 86 subfamilies by phylogenetic analysis. Consensus or representative sequences for each of the 86 subfamilies were then merged to the Repbase collection of known ERV/long terminal repeat (LTR) elements to annotate the retroviral complement of the bat genome. The results show that nearly 5% of the genome assembly is occupied by ERV-derived sequences, a quantity comparable to findings for other eutherian mammals. About one-fourth of these sequences belong to subfamilies newly identified in this study. Using two independent methods, intraelement LTR divergence and analysis of orthologous loci in two other bat species, we found that the vast majority of the potentially complete proviruses identified in *M. lucifugus* were integrated in the last ~25 million years. All three major ERV classes include recently integrated proviruses, suggesting that a wide diversity of retroviruses is still circulating in *Myotis* bats.

With 1,116 known extant species in 202 genera, bats (order *Chiroptera*) constitute more than 20% of living mammal species (1). The family *Vespertilionidae*, which contains about one-third of all bat species and more than 100 species in the genus *Myotis*, ranks among the most species rich of all mammal families. Bats display many exceptional developmental and physiological characteristics, including the extreme elongation of digits to form webbed wings enabling powered flight, the capacity of several species to undergo extended hibernation, and extraordinary life spans for their size and metabolic rate (up to 34 years in the wild for *Myotis*), making them emerging models for research in limb development (2, 3) and aging (4). Bats have also gained attention in biomedical research because a number of bat species have been identified as zoonotic reservoirs for some of the most sinister viruses infecting humans, such as rabies, Ebola, Marburg, Hendra, Nipah, and SARS-like viruses (5–10). A recent study suggests that bats host almost twice as many zoonotic viruses per species as rodents, another important reservoir of zoonotic viruses (11).

The growing notoriety of bats as reservoirs for zoonotic viruses has generated considerable interest in the scientific community and prompted a broad effort to characterize the viruses naturally infecting bats, including recent metagenomic surveys of the "virome" of several bat species (12–15). Together, these studies have led to the detection of a large number of viruses affiliated with diverse mammalian families of (mostly) RNA viruses, as well as insect and plant viruses (12–15).

Retroviruses are unique among vertebrate viruses in that they possess an obligatory chromosomal integration stage in their replication cycle. Integration may occasionally occur in the germ line, which can result in vertical inheritance and fixation in the host population (16–18). Such endogenous retroviruses (ERVs) have been identified in nearly all vertebrate genomes examined (16–18), and they often occupy a substantial fraction of mammalian genomes, accounting for about 8% of human (19) and 10% of mouse nuclear genome sequences (20). The infiltration and amplification of ERVs in vertebrate genomes are pervasive and represent a source of genetic variation thought to have had a strong impact on the biology and evolution of host species (18, 21, 22). Furthermore, because ERV integration events can often be dated, they provide a precious fossil record of past retroviral infections that have afflicted the host species or its ancestors (22–25).

Despite the prevalence of ERVs in mammalian genomes and their biological relevance, relatively few bat retroviral sequences have been reported in the literature (26–29). Initially, these were only short ERV fragments isolated by PCR with degenerate primers designed to amplify conserved *pol* domains of retroviruses (26, 27). More recently, traces of foamy viruses (spumaviruses) were identified in bat viromes (15), and Cui et al. reported an apparently complete sequence for an exogenous gammaretrovirus (*Rhinolophus ferrumequinum* retrovirus [RfRV]) in the greater horseshoe bat, as well as defective gammaretroviral sequences in other bat species (30). Lastly, the same group identified ~50 copies of endogenous gammaretroviruses in the draft genome sequences of *M. lucifugus* and of the megabat *Pteropus vampyrus* and were able to recover a total of 16 proviruses with both of the long terminal repeats (LTRs) but apparently defective coding capacity (28).

Received 5 April 2013 **Accepted** 17 May 2013
Published ahead of print 29 May 2013
Address correspondence to Cédric Feschotte, cedric@genetics.utah.edu.
Supplemental material for this article may be found at http://dx.doi.org/10.1128/JVI.00892-13.

doi:10.1128/JVI.00892-13

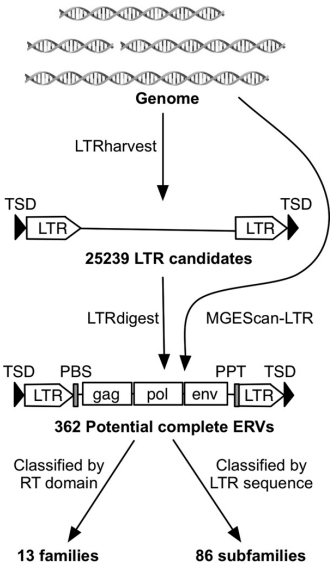

FIG 1 Potential complete ERV identification pipeline. We identified 25,239 LTR candidates with pairs of putative LTRs and TSD using LTRharvest, and all of them were annotated using LTRdigest. LTR candidates with canonical retroviral features were extracted as potential complete ERVs. We also used the independent pipeline MGEScan-LTR to identify potential complete ERVs. By combining the two independent pipelines, we identified 362 potential complete ERVs. They were further classified into 13 families based on RT domain phylogeny and into 86 subfamilies based on LTR sequence similarity.

These results suggested that bats are host to a large diversity of gammaretroviruses, including endogenous elements (28, 30). Early this year, endogenous betaretroviruses were also identified and reported in megabats and microbats (29). However, the overall diversity and evolution of ERVs in bat genomes remain largely unexplored.

In this study, we take advantage of the recent public release of a high-quality, 7× genome assembly (http://www.genome.gov/25521745) of the little brown bat *Myotis lucifugus*, one of the most common species in North America, to perform a comprehensive mining and analysis of ERVs in a bat species. We found that the amount and diversity of ERVs in *M. lucifugus* rival those observed in other mammalian genomes and include both ancient and recent integration events. Our study suggests that the vespertilionid bats have been subject to considerable levels of retroviral infections over the last ~25 million years (My) and that diverse retroviruses are likely still circulating among natural populations of *M. lucifugus*.

MATERIALS AND METHODS

ERV mining. The *M. lucifugus* 7× genome assembly Myoluc 2.0 was downloaded from NCBI (NCBI accession number AAPE02000000) and used as the input for two ERV identification pipelines (Fig. 1). First, we identified pairs of putative LTRs separated by 1 to 15 kb and flanked by target site duplications (TSD) by using LTRharvest (31). The LTR nucleotide similarity threshold used in LTRharvest was >80%, with other parameters set to their defaults. Internal retroviral sequence features of ERV candidates, including protein domains, primer-binding sites (PBS), and polypurine tracts (PPT), were predicted using LTRdigest (32). The *M. lucifugus* tRNA library used for PBS annotation was generated for the Myoluc 2.0 genome assembly using tRNAscan-SE (33), and 32 retroviral-related protein domain profiles (see File S1 in the supplemental material) used for putative domain annotation were downloaded from the Pfam database (34). To remove false positives and arrive at a list of high-confidence full-length ERVs, we applied two additional filters. First, we performed a tblastn search against all repeat libraries in Repbase (version 17.11) (45) to remove candidates whose reverse transcriptase (RT) domains were most closely related to those of non-LTR retrotransposons. Second, we required each candidate to contain at least 3 of the 5 canonical retroviral protein domains (Gag, PR, RT, IN, and RH) identified by LTRharvest. We also observed that some of the predicted LTR boundaries were truncated, so we manually refined the LTR termini for each of the filtered full-length ERVs using genomic alignments with blastn. The second ERV identification pipeline employed MGEScan-LTR (36) with the default parameters. The outputs from LTRdigest and MGEScan-LTR were also submitted to CENSOR (37) to systematically identify any other known repetitive elements inserted within the candidate ERVs. This approach was also used to eliminate several false positives where a pair of short interspersed elements (SINEs) flanking putative retroviral domains was misidentified as LTRs.

ERV classification and phylogenetic analysis. We used MUSCLE (38), complemented by manual refinements, to build an amino acid multiple alignment of the RT domain from 177 full-length bat ERVs and 20 known exogenous and endogenous retroviruses (see File S2 in the supplemental material). A neighbor-joining phylogeny was built from the RT domain alignment using MEGA5 (39) with 1,000 bootstrap replicates, applying the pairwise deletion option and using JTT as the amino acid substitution model (40). A Bayesian phylogenetic reconstruction was built using MrBayes 3.1.2 (41) with two runs of 5 million generations, employing a mixed-rate model. The tree was sampled every 100 generations. Posterior probabilities supporting family clustering are summarized in Table 1. For subfamily clustering of LTR sequences, we used Vmatch with parameters set according to the LTRdigest protocol (32).

Dating ERV insertions by using LTR divergence. LTR pairs from 362 full-length ERVs were aligned using the Smith-Waterman algorithm (42). CpG sites in all LTR sequences were removed, and the pairwise evolutionary distance K of LTR pairs was corrected using the Jukes-Cantor model (43). A previously estimated substitution rate (r) of 2.692×10^{9} for the *M. lucifugus* lineage (44) was used for dating each insertion. The date of ERV integration was calculated as $K/2r$.

RepeatMasker analysis. To generate a systematic annotation of ERVs in the *M. lucifugus* genome assembly, we first collected the LTR sequences and internal regions of potentially complete proviruses identified *ab initio*. We first separate the LTR sequences and internal regions from complete elements and extracted representative or consensus sequences from each of the 86 subfamilies. These 172 sequences formed our *M. lucifugus* ERV (MLERV) library. To remove any repetitive elements nested within the MLERV library, we screened this library using RepeatMasker (version 3.3.0) (3.0.1996-2010 [http://www.repeatmasker.org]) with a library of non-ERV repetitive elements from Repbase (version 17.11) (45). We then combined our 172 MLERV entries with a Repbase ERV library (version 17.11) (45) to build a custom ERV library. This library was used to subsequently run RepeatMasker on the *M. lucifugus* genome assembly, using the sensitive Crossmatch alignment program with the default parameters.

Estimation of full-length ERVs and solitary LTRs. A Perl script was used to parse the RepeatMasker output to systematically identify LTR pairs flanking internal ERV regions masked on the same DNA strand. A potential full-length ERV was considered when a pair of similar LTR fragments were separated by less than 20 kb and the alignment of the pair of LTR fragments spanned at least 100 bp. We also required that at least 500 bp of internal region were masked as internal ERV sequences.

To estimate solitary LTR numbers, we parsed the RepeatMasker output to map solo LTRs. In many cases, we found LTRs to be fragmented. To

TABLE 1 Summary of 13 MLERV families

Family	Class	Neighbor-joining bootstrap	Bayesian posterior probability	Age (My[a])	LTR length (bp)	Internal length (bp)	Copy no.
MLERV1	I	99.5	1	4.2	442	6,554	33
MLERV2	I	100	1	6.9	418	7,531	5
MLERV3	I	86.8	0.99	6.8	433	5,961	71
MLERV4	I	100	0.99	15.8	339	7,098	48
MLERV5	I	0.75	1	15.0	358	7,380	16
MLERV6	I	100	1	13.0	425	9,007	3
MLERV7	III	100	1	3.0	868	10,596	2
MLERV8	II	100	0.99	4.8	334	5,042	23
MLERV9	II	98.3	0.99	13.1	393	5,216	19
MLERV10	II	76	0.99	7.6	444	4,344	25
MLERV11	II	100	0.99	10.5	546	5,515	48
MLERV12	II	0.54	0.93	8.7	432	6,529	20
MLERV13	II	100	0.87	9.6	421	7,170	41

[a] My, million years.

better estimate solitary LTR copy numbers from fragmented pieces in the genome, we aligned fragmented LTR sequence to their consensus sequence and calculated the occurrence of each base in the alignment. Theoretically, the abundance of each base should be the same. In reality, it fluctuates because of genome rearrangements. Therefore, we used the median occurrence of an LTR as a proxy for its genomic copy number. Paired LTRs from full-length elements are included in the genomic copy number as well, so we subtracted twice the full-length ERV copy number from the genomic copy number and used it as the solitary LTR number.

Identification of orthologous MLERV insertion sites in other bat genomes. A Perl script was designed to find orthologous loci of MLERVs in other bat genome assemblies. The first and last 100 bp of each MLERV plus 300 bp of their flanking sequences were extracted from the *M. lucifugus* genome assembly and used as queries to search other genome assemblies using blastn. Genome sequences matching only the flanking region in the queries were labeled "empty sites," while sequences matching both the flanking and repeat regions were labeled "occupied." A given MLERV was considered present or absent at an orthologous locus when at least one end could be unambiguously labeled an occupied or empty site. All of the orthologous loci were validated by manual inspection.

RESULTS

***De novo* detection of ERVs in the *M. lucifugus* genome.** We used two different ERV mining pipelines (Fig. 1). The first strategy relies on the combination of LTRharvest (31) and LTRdigest (32). We used LTRharvest to define ERV candidates by scanning the genome sequence for putative LTR pairs (100 to 1,000 bp) separated by 1,000 to 15,000 bp and flanked by target site duplications (TSD). LTRdigest then screens and annotates each internal sequence of the ERV candidates for putative protein-coding domains (e.g., reverse transcriptase, integrase, etc.), primer-binding sites (PBS), and polypurine tracts (PPT), characteristic of complete proviruses. A filter is then applied to retain complete or nearly complete ERVs based on the presence of a subset of these features (see Materials and Methods). To complement this approach, we applied a second computational tool, MGEScan-LTR, designed to identify full-length LTR retrotransposons, including ERVs (36). LTRdigest and MGEScan-LTR both use HMMER (46) to identify protein domains; however, LTRdigest outputs all retroviral protein domains with an E value of $<1^{-6}$ for further identification, while MGEScan-LTR retains candidates with a set of protein domains with a combined E value of $<1^{-10}$ or a longest open reading frame (ORF) length of >700 bp.

With LTRharvest (31), we identified 25,239 ERV candidates with a pair of predicted LTRs in the *M. lucifugus* genome. This large output was filtered down to 217 ERV candidates by LTRdigest (32). Applying MGEScan-LTR (36) revealed 245 putative full-length ERVs (Fig. 1). While the total numbers of ERVs identified by the two pipelines were similar, only a small subset of elements were identified by both programs, as determined based on their location in the genome assembly. After removing redundant elements and false positives (see Materials and Methods), we arrived at a total of 362 distinct and potentially complete proviruses identified in the *M. lucifugus* genome assembly (hereinafter referred to as potentially complete ERVs) (Fig. 1). The LTR lengths of these 362 ERVs vary from 154 to 840 bp, and their total internal lengths range from 2,291 to 12,503 bp after removing secondary transposon insertions (see Table S1 in the supplemental material). Of the 362 ERVs, 252 (~70%) have perfect TSD, and 27 have identifiable TSD with 1 or 2 mutations.

Phylogenetic analysis and classification of ERVs from *M. lucifugus*. Of the 362 ERVs identified as described above, 177 had a reverse transcriptase (RT) domain conserved enough to be aligned confidently for phylogenetic analysis. We used this conserved RT domain (47) to build a multiple alignment and compute phylogenetic trees using the neighbor-joining method implemented in MEGA (39) and the Bayesian method implemented in MrBayes (48). Both methods produced trees with nearly identical topologies, allowing us to classify bat ERVs into 13 major families, denoted MLERV1 to -13 (Fig. 2 and Table 1). We defined all families as monophyletic groups of closely related branches with bootstrap support of at least 75% in neighbor joining and posterior probability of at least 0.75 in Bayesian trees (except for the MLERV12 family, which was supported by 54% bootstrap but a posterior probability of 0.93). Representatives of the known retroviral classes were included in our phylogenetic analysis in order to assign the MLERV families to one of the three major ERV classes. We were able to identify 6 MLERV families (MLERV1 to -6) comprised of 145 elements as class I ERVs (gammaretroviruses), 6 MLERV families (MLERV7 to -12) accounting for 157 elements as class II ERVs (betaretroviruses), and one family (MLERV13) represented by two elements as class III ERVs (spumaretroviruses).

To further classify MLERVs into subfamilies, we compared

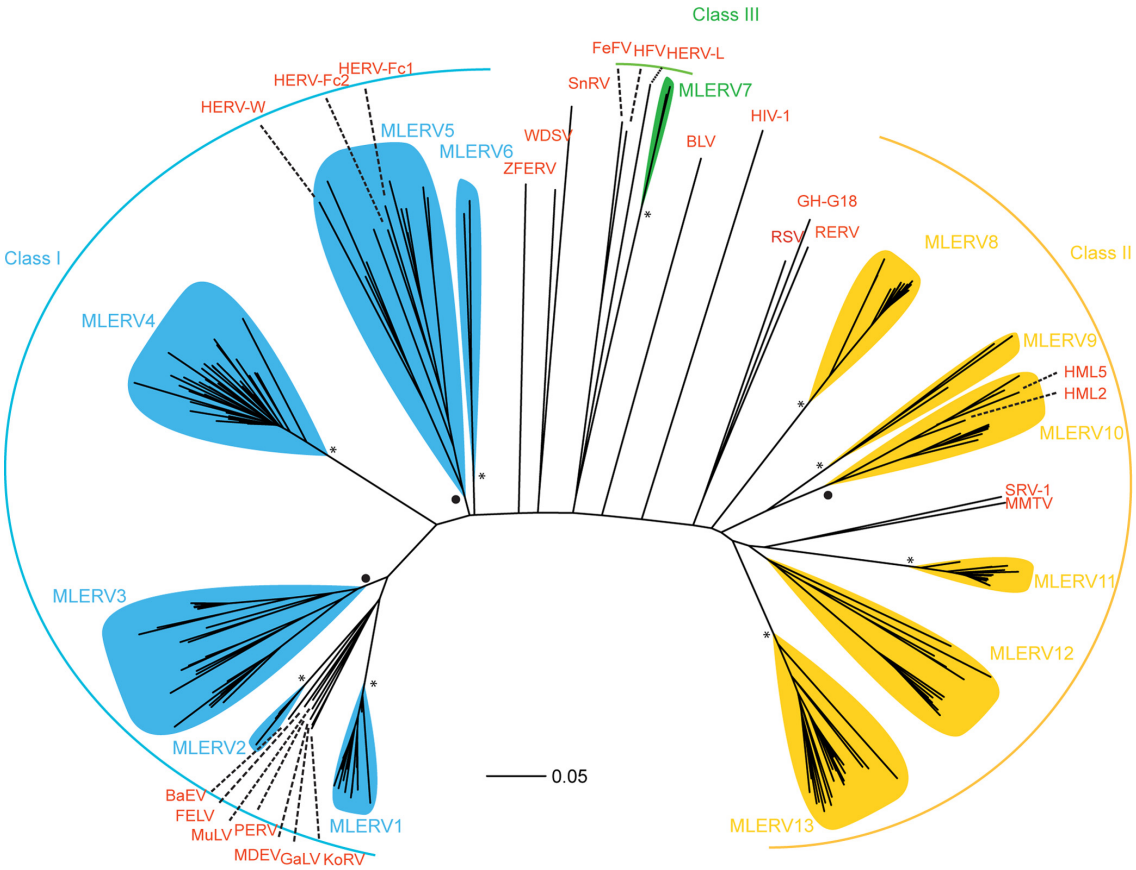

FIG 2 Phylogeny of 13 MLERV families and reference retroviral sequences. Class I, class II, and class III ERVs are illustrated with blue, yellow, and green, respectively, and reference retroviral sequences are shown in red. MLERV families with neighbor joining bootstrap values higher than 95 are labeled with an asterisk at the root, and those with a bootstrap value of between 75 and 95 are labeled with a dot at the root. KoRV, koala retrovirus; GaLV, gibbon ape leukemia virus; MDEV, *Mus dunni* endogenous virus; PERV, porcine endogenous retrovirus; MuLV, murine leukemia virus; FELV, feline leukemia virus; BaEV, baboon endogenous virus; HERV, human endogenous retrovirus; ZFERV, zebrafish endogenous retrovirus; WDSV, walleye dermal sarcoma virus; SnRV, snakehead fish retrovirus; FeFV, feline foamy virus; HFV, human foamy virus; BLV, bovine leukemia virus; RSV, Rous sarcoma virus; GH-G18, Golden hamster intracisternal A-particle H18; RERV, rabbit endogenous retrovirus; HML, human MMTV-like; SRV, simian type D retrovirus; MMTV, mouse mammary tumor virus.

their LTR sequences, which are among the most rapidly evolving sequences in retroviruses (49, 50). Based on a 75% interelement LTR nucleotide similarity cutoff, the program Vmatch (www.vmatch.de) clustered the 362 potential complete ERVs into 86 subfamilies (including 40 singletons) (see Table S2 in the supplemental material). Although families and subfamilies were defined independently, we found that the two classification levels were congruent in that ERVs falling within a given subfamily also belonged to the same family. One advantage of the classification based on LTR sequences is that we could generally assign elements with highly diverged, partial or missing RT domains to one of the families defined upon RT phylogeny. By combining these different classification methods, we were able to assign 354 ERVs to one of the 13 families defined in Table 1, leaving only 8 ERVs presently unclassified.

Census of the ERV population in the *M. lucifugus* genome assembly. To comprehensively assess the abundance of ERV-derived sequences in *M. lucifugus*, we ran RepeatMasker to annotate the 7× genome assembly using a custom library combining consensus or representative sequences for each of the 86 MLERV subfamilies defined above and all nonredundant ERV sequences deposited in Repbase (45) (see Materials and Methods). The total length of ERV-related sequences annotated by RepeatMasker amounted to 89 Mb, which represents 4.9% of the 1.8-Gb genome assembly after removing gaps.

To further delineate the ERV composition of the bat genome, we implemented custom scripts (available upon request) to parse the RepeatMasker output and estimate the numbers of full-length ERVs (as defined by the presence of a pair of LTRs flanking a sequence masked as an internal ERV region) and solitary LTRs for each major class of ERVs (see Materials and Methods). Solitary LTRs typically arise as a result of intraelement recombination between the 5′ and 3′ LTRs of a full-length provirus.

For class I ERVs, the approach identified 464 full-length proviruses and 35,404 solitary LTRs in the genome assembly. The

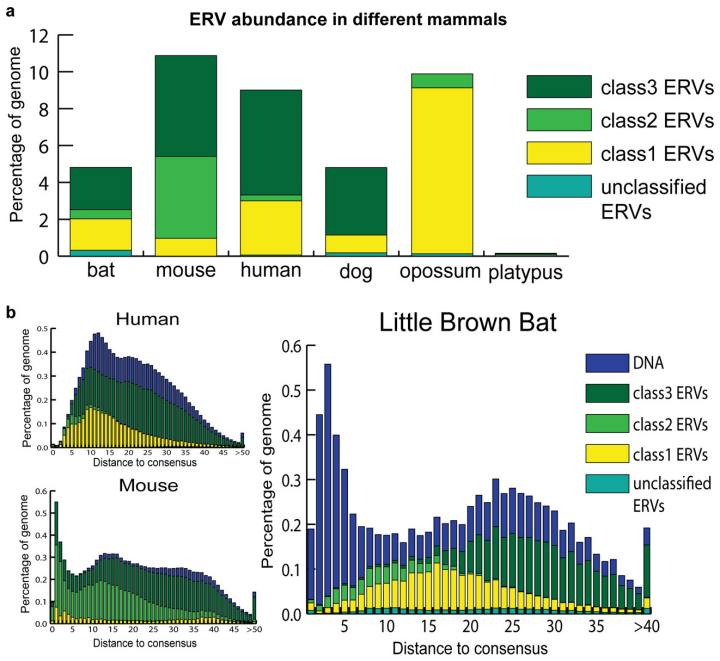

FIG 3 Comparison of ERV abundance and dynamics in different genomes. Different ERV classes and DNA transposons are labeled with different colors. (a) Comparison of percentages of genomes derived from different classes of ERVs in little brown bat and other mammals. (b) ERV and DNA transposon dynamics in little brown bat, human, and mouse genomes. Distance to consensus was corrected using the Jukes-Cantor model. Older elements are more distant from the consensus. The abundance is illustrated also, using the percentage of the genome.

sizes of full-length class I ERVs in *M. lucifugus* typically range from 6 to 9 kb (Table 1). MLERV3 is the most diverse family in this class, including 15 distinct subfamilies (Fig. 2 and Table 1). Together, the total genomic length occupied by class I elements is estimated at 31.5 Mb (1.66% of the genome assembly).

Class II ERVs were represented by 638 full-length proviruses and 10,858 solitary LTRs. The lengths of full-length class II ERVs range from 4.5 to 9.5 kb. The most abundant class II family is MLERV11 (Fig. 2 and Table 1). Notably, subfamily MLERV11_2 includes 123 potentially full-length copies, more than any other MLERV subfamily. In total, class II ERVs occupy 9.1 Mb (0.48%) of the genome assembly.

Covering 49.2 Mb (or 2.6%) of DNA, class III ERVs account for the largest amount of ERV-derived sequences in the genome assembly. This result was somewhat surprising in light of our initial *ab initio* mining of ERVs, which had retrieved a single class III family (MLERV7) represented by only 2 complete canonical copies (Fig. 2 and Table 1). Nonetheless, our parsing of the RepeatMasker output identified 571 full-length and 81,967 solitary LTRs affiliated with class III ERVs. Manual inspection of a subset of these sequences revealed that they represent relatively ancient and often nonautonomous class III elements previously identified in other mammalian genomes, such as mammalian apparent LTR retrotransposons (MaLRs) (51). Thus, the discrepancy between the results of the *ab initio* search and the RepeatMasker annotation can be explained by the fact that most class III ERVs are represented by highly decayed copies and nonautonomous elements, as well as abundant solitary LTRs derived from ancient families (see below). By design, such incomplete or highly diverged copies cannot be identified by the two pipelines used for our *ab initio* mining (31, 32, 36). The difficulty in identifying class III ERVs using *ab initio* approaches has been reported for other mammals (52–58).

Overall, the ERV coverage of the bat genome (89 Mb, 4.9%) is less than that in the human (261 Mb, 9.0%) and mouse (285 Mb, 10.9%) genomes but similar to the ERV coverage of the dog genome (115 Mb, 4.8%) (RepeatMasker) (Fig. 3a). However, the bat genome assembly is less complete and of poorer quality than the mouse and human genome assemblies. Because ERVs and other repeats tend to be overrepresented in nonassembled regions of sequenced genomes (gaps), our estimate of ERV abundance in *M. lucifugus* should be viewed as a conservative estimate.

Comparative demography of ERVs in bat, human, and mouse. The RepeatMasker output provides a measure of sequence divergence for each DNA segment annotated to its closest consensus sequence in our ERV library, enabling us to examine the tempo and evolutionary dynamics of ERV invasions in *M. lucifugus* in comparison to those in human and mouse (Fig. 3b). Overall, the demographic profile of *M. lucifugus* ERVs is more similar to that of human ERVs: class III ERVs are the most abundant and the most diverged (ancient), class II ERVs are the least abundant but the most recent, while class I ERVs occupy an intermediate position both in abundance and divergence. The similar histories of ERV accumulation in the bat and human (and to some extent mouse) lineages are to be contrasted with the dramatic differences in DNA transposon activity, which is strikingly elevated in the *M.*

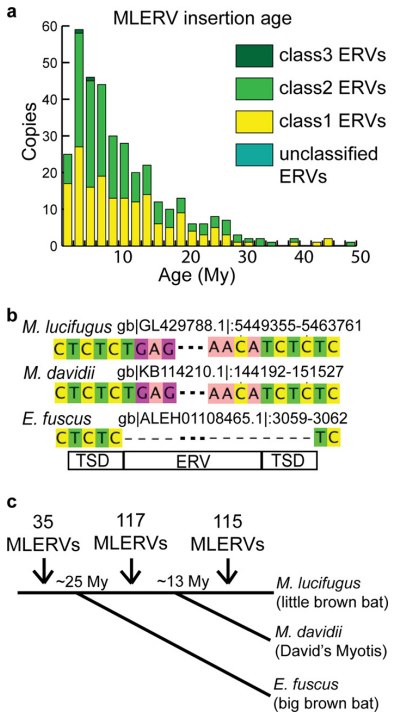

FIG 4 Recent ERV invasion in *M. lucifugus* genome. (a) Most of the 362 complete ERVs invaded the *M. lucifugus* genome recently. Copy numbers of different ERV classes are shown in different colors; age was estimated by LTR pair divergence. (b) An example of an MLERV integration event (TG. . .CA) specific to the *Myotis* lineage. The MLERV is present in the *M. davidii* genome with target site duplication (TCTC), but a precise empty site is found in *E. fuscus*. (c) Orthologous loci of MLERVs in *E. fuscus* and *M. davidii* indicate that most of the ERVs invaded after divergence from *E. fuscus*, and ERVs were active before and after speciation of *M. lucifugus* from *M. davidii*. The speciation times between *M. lucifugus* and *M. davidii* and between *M. lucifugus* and *E. fuscus* were around 13 My and around 25 My, respectively. Numbers of ERV insertions are labeled between time points.

lucifugus lineage (Fig. 3b), consistent with previous reports (59, 60).

Recent ERV infiltrations in the *M. lucifugus* lineage. The 5′ and 3′ LTR sequences from a given provirus are typically identical upon chromosomal integration and are expected to diverge subsequently by accumulating substitutions at the neutral rate of the host species. Thus, if the host neutral substitution rate is known, the age of an individual ERV integration event can be estimated by measuring the pairwise distance between LTR sequences (61). We applied this method and a rate of neutral substitution previously estimated for the *M. lucifugus* lineage (44) to date the integration of the 362 potential complete ERVs predicted *ab initio*, as we expected these to represent some of the youngest elements in the genome (Fig. 4a).

The results show that, indeed, the vast majority of the ERVs surveyed integrated relatively recently, with 232 of 362 (64%) proviral integrations estimated to be less than 10 My old according to this analysis. Twenty-three of these elements have strictly identical LTR pairs, and another 58 elements have LTRs that are >99% identical, indicating that all these ERVs have inserted very recently, probably within the past 2 My (Fig. 4a). The most recently active subfamily according to this analysis is MLERV3_15 of the ERV I class. We estimated that each of the 12 copies of MLERV3_15 have inserted within the last 2.5 My, including 4 copies with identical LTRs (see Table S2 in the supplemental material). These data suggest that the *M. lucifugus* lineage has been subject to many recent ERV infiltrations.

We measured the age of MLERV integration events alternatively by assessing their presence or absence at orthologous genomic positions in closely related bat species. Recently, draft genome assemblies of two additional vespertilionid bats, *Eptesicus fuscus* and *Myotis davidii*, were released (NCBI accession number ALEH01000000 and ALWT01000000, respectively) (62). *E. fuscus* has been estimated to share an ancestor with *Myotis* bats at ~25 My ago (63, 64), and the time of divergence of *M. lucifugus* and *M. davidii* is predicted to be around 10 to 15 My (64–66). We used BLAST with queries representing the termini of each of the individual full-length MLERV copies plus 300 bp of flanking genomic sequences to identify orthologous regions in the *E. fuscus* and *M. davidii* genomes (Fig. 4b shows an example). After combining information from these two other bat genomes and manually inspecting each locus, about 35% of orthologous MLERV loci could be unambiguously identified in *E. fuscus* and about 70% in *M. davidii* (see Tables S1 and S3 in the supplemental material). Among these, we found evidence for 137 MLERVs present at orthologous positions in *M. davidii*, while 115 MLERVs were missing at the orthologous site in this species (Fig. 4c) (52 of these loci are precisely missing the MLERV and have only one copy of the TSD). In the *E. fuscus* draft genome assembly, we identified 35 MLERVs present at orthologous loci, while 94 MLERVs were missing at orthologous positions (see Tables S1 and S3 in the supplemental material). Together, these data indicate that the vast majority of potential complete ERVs detected in *M. lucifugus* integrated after speciation of *E. fuscus* and *Myotis*, and many ERVs continued to accumulate during *Myotis* evolution and integrated after the divergence of *M. lucifugus* and *M. davidii* (Fig. 4c and Table 2).

Our age estimates based on these cross-species genomic comparisons were largely concordant with the age of ERV integrations calculated by LTR divergence. Indeed, MLERVs with orthologous empty sites in *E. fuscus* were on average much younger (7.2 My) than those with occupied sites (23.1 My). The oldest MLERV insertion with an empty site in *E. fuscus* was predicted to be 27 My old according to LTR divergence, which is roughly consistent with

TABLE 2 Ortholog status of identifiable complete ERVs in *M. davidii* and *E. fuscus*

Ortholog status in other genomes	Copy no.	Avg age (My)[a]	Oldest ERV infiltration (My)	Latest ERV infiltration (My)
M. davidii				
Empty site	115	4.2	24.3	0.0
Occupied site	137	15.9	54.8	1.4
E. fuscus				
Empty site	94	7.1	27.5	0.0
Occupied site	35	23.1	52.3	5.4

[a] Age is estimated using LTR pair comparison.

the divergence time of ~25 My estimated between these two bat species (63). However, we note that the age of the youngest MLERV insertions with occupied orthologous sites in *E. fuscus* was significantly underestimated by LTR divergence (5.4 My). Similar trends were found in *M. davidii* (summarized in Table 2). This discrepancy between the results of the two dating methods could be caused by gene conversion homogenizing LTR sequences, leading to underestimation of the timing of integration, as previously reported in other genomes (67). These data emphasize the need to apply multiple methods to confidently date ERV integration events.

DISCUSSION

Census of ERVs in the *M. lucifugus* genome. By combining two different *ab initio* mining strategies, we identified 362 potentially complete proviruses in the *M. lucifugus* genome. Nearly all of these elements fall within 86 subfamilies that enabled us to identify a multitude of related sequence fragments using RepeatMasker, including nearly 1,700 full-length ERVs and 130,000 solitary LTRs in the *M. lucifugus* genome assembly. When used in conjunction with mammalian ERV sequences catalogued in Repbase, our collection allowed us to estimate that ERVs occupy 4.9% of the bat genome, a substantial fraction comparable to that observed in other eutherian genomes (Fig. 3a) (19, 20, 69).

Our data complement previous findings by Cui et al. (28), who identified 3 major groups (A, B, and C) of gammaretroviruses in the *M. lucifugus* genome by BLAST searches. Our approach identified these three groups as the MLERV2, MLERV1, and MLERV3 families, respectively. We discovered three additional gammaretrovirus families (MLERV4 to -6) (Fig. 2 and Table 1). The total length of sequences derived from the MLERV4 family alone is 9.2 Mb, or ~0.5% of the genome assembly. At the time of this study, there were 5 entries of internal (coding) ERV regions and 132 entries of LTR sequences for *M. lucifugus* in Repbase, a comprehensive database for transposable element sequences, including ERVs (45). We identified both LTR and internal sequences for 13 families and 86 subfamilies of ERVs, most of which were not reported in Repbase (see Table S2 in the supplemental material). Furthermore, through manual examination, we found that several of the *M. lucifugus* LTR sequences deposited in Repbase were actually truncated at their 5′ end (data not shown). Thus, our manually curated collection of 86 reference ERV sequences will be useful to replace or complement existing Repbase entries. Overall, the coverage of MLERV families newly identified in this study amounts to 23 Mb of the genome assembly, thereby substantially improving the census of ERVs in this bat species.

Comparison of ERV diversity in *M. lucifugus* with that of other mammals. With regard to ERV diversity within *M. lucifugus*, we found that class I (gammaretroviruses) and class II (betaretroviruses) ERVs are similarly diverse (each composed of 6 major families), but the total amount of genomic DNA derived from class II ERVs (9.1 Mb) is considerably smaller than that derived from class I ERVs (31.5 Mb). Class III (spumaviruses) ERVs are the most abundant (49.2 Mb) in the genome, but they are generally older and more degraded than class I and II elements, which hampered the identification of full-length class III ERVs using *ab initio* methods, as reported for other mammalian genomes (52–58, 70). Using RepeatMasker, we identified 571 apparently full-length class III ERVs, but we observed that a large fraction of these elements are nonautonomous MaLR-like elements that are comparable to those abundantly populating the human and mouse genomes (19, 20). Nonetheless, we note that the only class III family we detected *ab initio* in *M. lucifugus* (MLERV7) is a relatively young family, with an age estimated at ~4 My (Table 1). Thus, all three major ERV classes are represented by relatively recent insertions in the *M. lucifugus* genome.

Overall, the demographic profile of the three ERV classes in *M. lucifugus* was more similar to that seen in the human genome (Fig. 3b). While the bulk of class III ERVs likely predate the radiation of eutherian mammals and, thus, have essentially been inherited through vertical decent, the amplification of class I and II ERVs is much more recent and largely lineage specific (Fig. 3). We conclude that there was a parallel invasion and expansion of these two classes of ERV in the human and bat lineages.

An important motivation for our analysis of ERVs in *M. lucifugus* relates to recent findings of massive lineage-specific DNA transposon activity in *M. lucifugus* (Fig. 3b) (59, 60). There is strong evidence that several of these DNA transposons have been acquired horizontally (71, 72), possibly reflecting a peculiar sensitivity of the germ line of this group of bats to lateral infiltration of mobile elements. Because retroviral endogenization also represents a form of horizontal transfer to the germ line, it was of interest to see whether these bats also display a greater vulnerability to ERV invasions. While we found clear evidence of recent ERV colonization in the genome of *M. lucifugus*, neither the diversity nor the sheer amount of ERV sequences depart dramatically from the diversity or amount observed in other mammalian genomes (Fig. 3). Thus, while the mobile element landscape of *M. lucifugus* is exceptional in terms of recent DNA transposon invasions, *M. lucifugus* does not appear to be an outlier among eutherian mammals in terms of its ERV population. We conclude that the apparent vulnerability of vespertilionid bats to horizontal transfer of DNA transposons is not generalizable to all types of mobile elements.

Superspreader hypothesis. Recently, Magiorkinis et al. (73) proposed the "superspreader" hypothesis, which postulates that ERVs lacking coding capacity for an envelope (*env*-less ERVs) amplify more efficiently within the genome than those encoding an intact envelope. The hypothesis was supported by a detailed phylogenetic analysis of intracisternal A-type particles (IAPs) from several mammalian genomes (73) and for several primate ERV families (74). In *M. lucifugus*, we classified MLERVs to 13 families and 86 subfamilies. At the family level, we found no clear relationship between the presence of an envelope domain and family copy number; however, at the subfamily level, we observed that the most successful subfamilies are predominantly composed of *env*-less elements (see Table S2 in the supplemental material). For example, the two largest subfamilies in our data set (MLERV4_6 and MLERV11_2) are entirely composed of copies lacking an identifiable envelope domain. Thus, the pattern of MLERV subfamily expansion brings further support to the superspreader hypothesis.

Bats as possible zoonotic reservoirs of retroviruses. We found several clear examples of very recent ERV families in *M. lucifugus*. A good illustration is MLERV3_15, a subfamily of class I elements. Four of the 12 copies identified in the genome have identical LTR pairs, while the other eight have LTR pairs that are >99% identical, indicative of nearly contemporary integration events (see Table S2 in the supplemental material). All 12 copies are also absent at orthologous positions in *M. davidii* (see Table

S3). Nonetheless, none of the MLERV3_15 copies identified appear to retain intact coding capacity, suggesting that they are currently incapable of replicating autonomously.

However, in a recently active class I ERV subfamily, MLERV2_2 (0 to 3 My old), we identified one copy (entry 74) with apparently intact *gag*, *pro*, *pol*, and *env* coding regions, suggesting that this copy might be replication competent. In addition, another apparently intact and functional class II ERV was recently identified in *M. lucifugus* (29). Together, these results suggest that both class I and II ERVs in *M. lucifugus* are potentially capable of autonomous replication and of producing infectious viral particles.

Among the most recently integrated (<10 My ago) potentially complete proviruses supported by both LTR-LTR divergence and cross-species analysis, we were able to detect members of all three main retroviral classes (see Table S2 in the supplemental material). Our finding of recently integrated spumaretroviruses and gammaretroviruses is consistent with the identification of exogenous members of these retroviral taxa in several bat species, including microbats (15, 30). We also identified proviral copies of betaretroviruses (e.g., MLERV12_4) that have retained identical LTRs flanked by perfect TSD and are absent in *M. davidii* (see Tables S2 and S3), which suggests that *M. lucifugus* was also infected by exogenous betaretroviruses in the recent past. Together, these data indicate that a wide diversity of retroviruses have recently infected these bats and are likely still circulating in natural populations of *M. lucifugus*. Given the apparent propensity of bats to act as reservoir species for zoonotic viruses that are highly pathogenic to humans, these observations raise concerns that these animals may also be capable of transmitting zoonotic retroviruses to humans.

ACKNOWLEDGMENTS

We thank Aurelie Kapusta, Ellen Pritham, and Claudia Marquez for helpful discussions and Ray Malfavon-Borja for critical reading of the manuscript.

X.Z. and C.F. were supported by grant R01-GM077582 from the National Institutes of Health.

REFERENCES

1. **Simmons NB.** 2005. Order Chiroptera, p 312–529. *In* Wilson DE, Reeder DM (ed), Mammal species of the world: a taxonomic and geographic reference, 3rd ed, vol 1. Johns Hopkins University Press, Baltimore, MD.
2. **Hockman D, Mason MK, Jacobs DS, Illing N.** 2009. The role of early development in mammalian limb diversification: a descriptive comparison of early limb development between the Natal long-fingered bat (Miniopterus natalensis) and the mouse (Mus musculus). Dev. Dyn. **238:**965–979.
3. **Behringer RR, Rasweiler JJ, Chen CH, Cretekos CJ.** 2009. Genetic regulation of mammalian diversity. Cold Spring Harbor Symp. Quant. Biol. **74:**297–302.
4. **Munshi-South J, Wilkinson GS.** 2010. Bats and birds: exceptional longevity despite high metabolic rates. Ageing Res. Rev. **9:**12–19.
5. **Calisher CH, Childs JE, Field HE, Holmes KV, Schountz T.** 2006. Bats: important reservoir hosts of emerging viruses. Clin. Microbiol. Rev. **19:**531–545.
6. **Dobson AP.** 2005. What links bats to emerging infectious diseases? Science **310:**628–629.
7. **Lau SKP, Woo PCY, Li KSM, Huang Y, Tsoi H-W, Wong BHL, Wong SSY, Leung S-Y, Chan K-H, Yuen K-Y.** 2005. Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. Proc. Natl. Acad. Sci. U. S. A. **102:**14040–14045.
8. **Li W, Shi Z, Yu M, Ren W, Smith C, Epstein JH, Wang H, Crameri G, Hu Z, Zhang H, Zhang J, McEachern J, Field H, Daszak P, Eaton BT, Zhang S, Wang L-F.** 2005. Bats are natural reservoirs of SARS-like coronaviruses. Science **310:**676–679.
9. **Wong S, Lau S, Woo P, Yuen K-Y.** 2007. Bats as a continuing source of emerging infections in humans. Rev. Med. Virol. **17:**67–91.
10. **Drexler JF, Corman VM, Wegner T, Tateno AF, Zerbinati RM, Gloza-Rausch F, Seebens A, Müller MA, Drosten C.** 2011. Amplification of emerging viruses in a bat colony. Emerg. Infect. Dis. **17:**449–456.
11. **Luis AD, Hayman DTS, O'Shea TJ, Cryan PM, Gilbert AT, Pulliam JRC, Mills JN, Timonin ME, Willis CKR, Cunningham AA, Fooks AR, Rupprecht CE, Wood JLN, Webb CT.** 2013. A comparison of bats and rodents as reservoirs of zoonotic viruses: are bats special? Proc. Biol. Sci. **280:**20122753. doi:10.1098/rspb.2012.2753.
12. **Li L, Victoria JG, Wang C, Jones M, Fellers GM, Kunz TH, Delwart E.** 2010. Bat guano virome: predominance of dietary viruses from insects and plants plus novel mammalian viruses. J. Virol. **84:**6955–6965.
13. **Donaldson EF, Haskew AN, Gates JE, Huynh J, Moore CJ, Frieman MB.** 2010. Metagenomic analysis of the viromes of three North American bat species: viral diversity among different bat species that share a common habitat. J. Virol. **84:**13004–13018.
14. **Ge X, Li Y, Yang X, Zhang H, Zhou P, Zhang Y, Shi Z.** 2012. Metagenomic analysis of viruses from bat fecal samples reveals many novel viruses in insectivorous bats in China. J. Virol. **86:**4620–4630.
15. **Wu Z, Ren X, Yang L, Hu Y, Yang J, He G, Zhang J, Dong J, Sun L, Du J, Liu L, Xue Y, Wang J, Yang F, Zhang S, Jin Q.** 2012. Virome analysis for identification of novel mammalian viruses in bat species from Chinese provinces. J. Virol. **86:**10999–11012.
16. **Gifford R, Tristem M.** 2003. The evolution, distribution and diversity of endogenous retroviruses. Virus Genes **26:**291–315.
17. **Blikstad V, Benachenhou F, Sperber GO, Blomberg J.** 2008. Evolution of human endogenous retroviral sequences: a conceptual account. Cell. Mol. Life Sci. **65:**3348–3365.
18. **Jern P, Coffin JM.** 2008. Effects of retroviruses on host genome function. Annu. Rev. Genet. **42:**709–732.
19. **Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, et al.** 2001. Initial sequencing and analysis of the human genome. International Human Genome Sequencing Consortium. Nature **409:**860–921.
20. **Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, et al.** 2002. Initial sequencing and comparative analysis of the mouse genome. Mouse Genome Sequencing Consortium. Nature **420:**520–562.
21. **Kurth R, Bannert N.** 2010. Beneficial and detrimental effects of human endogenous retroviruses. Int. J. Cancer **126:**306–314.
22. **Feschotte C, Gilbert C.** 2012. Endogenous viruses: insights into viral evolution and impact on host biology. Nat. Rev. Genet. **13:**283–296.
23. **Katzourakis A, Gifford RJ, Tristem M, Gilbert MTP, Pybus OG.** 2009. Macroevolution of complex retroviruses. Science **325:**1512.
24. **Emerman M, Malik HS.** 2010. Paleovirology—modern consequences of ancient viruses. Plos Biol. **8:**e1000301. doi:10.1371/journal.pbio.1000301.
25. **Holmes EC.** 2011. The evolution of endogenous viral elements. Cell Host Microbe **10:**368–377.
26. **Tristem M, Kabat P, Lieberman L, Linde S, Karpas A, Hill F.** 1996. Characterization of a novel murine leukemia virus-related subgroup within mammals. J. Virol. **70:**8241–8246.
27. **Baillie GJ, van de Lagemaat LN, Baust C, Mager DL.** 2004. Multiple groups of endogenous betaretroviruses in mice, rats, and other mammals. J. Virol. **78:**5784–5798.
28. **Cui J, Tachedjian G, Tachedjian M, Holmes EC, Zhang S, Wang L-F.** 2012. Identification of diverse groups of endogenous gammaretroviruses in mega- and microbats. J. Gen. Virol. **93:**2037–2045.
29. **Hayward JA, Tachedjian M, Cui J, Field H, Holmes EC, Wang L-F, Tachedjian G.** 2013. Identification of diverse full-length endogenous be-

taretroviruses in megabats and microbats. Retrovirology **10:**35. doi:10.1186/1742-4690-10-35.
30. **Cui J, Tachedjian M, Wang L, Tachedjian G, Wang L-F, Zhang S.** 2012. Discovery of retroviral homologs in bats: implications for the origin of mammalian gammaretroviruses. J. Virol. **86:**4288–4293.
31. **Ellinghaus D, Kurtz S, Willhoeft U.** 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinform. **9:**18. doi:10.1186/1471-2105-9-18.
32. **Steinbiss S, Willhoeft U, Gremme G, Kurtz S.** 2009. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. Nucleic Acids Res. **37:**7002–7013.
33. **Lowe TM, Eddy SR.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25:**955–964.
34. **Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A, Finn RD.** 2012. The Pfam protein families database. Nucleic Acids Res. **40:**D290–D301.
35. Reference deleted.
36. **Rho M, Choi J-H, Kim S, Lynch M, Tang H.** 2007. De novo identification of LTR retrotransposons in eukaryotic genomes. BMC Genomics **8:**90. doi:10.1186/1471-2164-8-90.
37. **Kohany O, Gentles AJ, Hankus L, Jurka J.** 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinform. **7:**474. doi:10.1186/1471-2105-7-474.
38. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32:**1792–1797.
39. **Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S.** 2011. MEGA5: Molecular Evolutionary Genetics Analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. **28:**2731–2739.
40. **Jones DT, Taylor WR, Thornton JM.** 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. **8:**275–282.
41. **Ronquist F, Huelsenbeck JP.** 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19:**1572–1574.
42. **Smith TF, Waterman MS.** 1981. Identification of common molecular subsequences. J. Mol. Biol. **147:**195–197.
43. **Jukes TH, Cantor CR.** 1969. Evolution of protein molecules, p 21–132. *In* Munro HN, Mammalian protein metabolism, vol III. Academic Press, San Diego, CA.
44. **Pace JK, Gilbert C, Clark MS, Feschotte C.** 2008. Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. Proc. Natl. Acad. Sci. U. S. A. **105:**17023–17028.
45. **Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J.** 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. **110:**462–467.
46. **Eddy SR.** 2011. Accelerated profile HMM searches. PLoS Comput. Biol. **7:**e1002195. doi:10.1371/journal.pcbi.1002195.
47. **Xiong Y, Eickbush TH.** 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. **9:**3353–3362.
48. **Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP.** 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. **61:**539–542.
49. **Slattery JP, Franchini G, Gessain A.** 1999. Genomic evolution, patterns of global dissemination, and interspecies transmission of human and simian T-cell leukemia/lymphotropic viruses. Genome Res. **9:**525–540.
50. **Fernández-Medina RD, Ribeiro JMC, Carareto CMA, Velasque L, Struchiner CJ.** 2012. Losing identity: structural diversity of transposable elements belonging to different classes in the genome of Anopheles gambiae. BMC Genomics **13:**272. doi:10.1186/1471-2164-13-272.
51. **Smit AF.** 1993. Identification of a new, abundant superfamily of mammalian LTR-transposons. Nucleic Acids Res. **21:**1863–1872.
52. **McCarthy EM, McDonald JF.** 2004. Long terminal repeat retrotransposons of Mus musculus. Genome Biol. **5:**R14. doi:10.1186/gb-2004-5-3-r14.
53. **Polavarapu N, Bowen NJ, McDonald JF.** 2006. Identification, characterization and comparative genomics of chimpanzee endogenous retroviruses. Genome Biol. **7:**R51.
54. **Polavarapu N, Bowen NJ, McDonald JF.** 2006. Newly identified families of human endogenous retroviruses. J. Virol. **80:**4640–4642.
55. **Garcia-Etxebarria K, Jugo BM.** 2010. Genome-wide detection and characterization of endogenous retroviruses in Bos taurus. J. Virol. **84:**10852–10862.
56. **Martínez Barrio Á, Ekerljung M, Jern P, Benachenhou F, Sperber GO, Bongcam-Rudloff E, Blomberg J, Andersson G.** 2011. The first sequenced carnivore genome shows complex host-endogenous retrovirus relationships. PLoS One **6:**e19832. doi:10.1371/journal.pone.0019832.
57. **Garcia-Etxebarria K, Jugo BM.** 2012. Detection and characterization of endogenous retroviruses in the horse genome by in silico analysis. Virology **434:**59–67.
58. **Brown K, Moreton J, Malla S, Aboobaker AA, Emes RD, Tarlinton RE.** 2012. Characterisation of retroviruses in the horse genome and their transcriptional activity via transcriptome sequencing. Virology **433:**55–63.
59. **Ray DA, Feschotte C, Pagan HJT, Smith JD, Pritham EJ, Arensburger P, Atkinson PW, Craig NL.** 2008. Multiple waves of recent DNA transposon activity in the bat, Myotis lucifugus. Genome Res. **18:**717–728.
60. **Pritham EJ, Feschotte C.** 2007. Massive amplification of rolling-circle transposons in the lineage of the bat Myotis lucifugus. Proc. Natl. Acad. Sci. U. S. A. **104:**1895–1900.
61. **Dangel AW, Baker BJ, Mendoza AR, Yu CY.** 1995. Complement component C4 gene intron 9 as a phylogenetic marker for primates: long terminal repeats of the endogenous retrovirus ERV-K(C4) are a molecular clock of evolution. Immunogenetics **42:**41–52.
62. **Zhang G, Cowled C, Shi Z, Huang Z, Bishop-Lilly KA, Fang X, Wynne JW, Xiong Z, Baker ML, Zhao W, Tachedjian M, Zhu Y, Zhou P, Jiang X, Ng J, Yang L, Wu L, Xiao J, Feng Y, Chen Y, Sun X, Zhang Y, Marsh GA, Crameri G, Broder CC, Frey KG, Wang L-F, Wang J.** 2013. Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. Science **339:**456–460.
63. **Miller-Butterworth CM, Murphy WJ, O'Brien SJ, Jacobs DS, Springer MS, Teeling EC.** 2007. A family matter: conclusive resolution of the taxonomic position of the long-fingered bats, miniopterus. Mol. Biol. Evol. **24:**1553–1561.
64. **Lack JB, Roehrs ZP, Stanley CE, Jr, Ruedi M, Van Den Bussche RA.** 2010. Molecular phylogenetics of Myotis indicate familial-level divergence for the genus Cistugo (Chiroptera). J. Mammal. **91:**976–992.
65. **Agnarsson I, Zambrana-Torrelio CM, Flores-Saldana NP, May-Collado LJ.** 2011. A time-calibrated species-level phylogeny of bats (Chiroptera, Mammalia). PLoS Curr. **3:**RRN1212. doi:10.1371/currents.RRN1212.
66. **Stadelmann B, Lin LK, Kunz TH, Ruedi M.** 2007. Molecular phylogeny of New World Myotis (Chiroptera, Vespertilionidae) inferred from mitochondrial and nuclear DNA genes. Mol. Phylogenet. Evol. **43:**32–48.
67. **Kijima TE, Innan H.** 2010. On the estimation of the insertion time of LTR retrotransposable elements. Mol. Biol. Evol. **27:**896–904.
68. Reference deleted.
69. **Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ, Zody MC, Mauceli E, Xie X, Breen M, Wayne RK, Ostrander EA, Ponting CP, Galibert F, Smith DR, deJong PJ, Kirkness E, Alvarez P, Biagi T, Brockman W, Butler J, Chin C-W, Cook A, Cuff J, Daly MJ, DeCaprio D, Gnerre S, Grabherr M, Kellis M, Kleber M, Bardeleben C, Goodstadt L, Heger A, Hitte C, Kim L, Koepfli K-P, Parker HG, Pollinger JP, Searle SMJ, Sutter NB, Thomas R, Webber C, Baldwin J, Abebe A, Abouelleil A, Aftuck L, Ait-zahra M, et al.** 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature **438:**803–819.
70. **Han K, Konkel MK, Xing J, Wang H, Lee J, Meyer TJ, Huang CT, Sandifer E, Hebert K, Barnes EW, Hubley R, Miller W, Smit AFA, Ullmer B, Batzer MA.** 2007. Mobile DNA in Old World monkeys: a glimpse through the rhesus macaque genome. Science **316:**238–240.
71. **Gilbert C, Schaack S, Pace JK, Brindley PJ, Feschotte C.** 2010. A role for host-parasite interactions in the horizontal transfer of transposons across phyla. Nature **464:**1347–1350.
72. **Thomas J, Schaack S, Pritham EJ.** 2010. Pervasive horizontal transfer of rolling-circle transposons among animals. Genome Biol. Evol. **2:**656–664.
73. **Magiorkinis G, Gifford RJ, Katzourakis A, De Ranter J, Belshaw R.** 2012. Env-less endogenous retroviruses are genomic superspreaders. Proc. Natl. Acad. Sci. U. S. A. **109:**7385–7390.
74. **Belshaw R, Katzourakis A, Paces J, Burt A, Tristem M.** 2005. High copy number in human endogenous retrovirus families is associated with copying mechanisms in addition to reinfection. Mol. Biol. Evol. **22:**814–817.

CHAPTER 3

CROSS-SPECIES TRANSMISSION AND DIFFERENTIAL FATE OF AN ENDOGENOUS RETROVIRUS IN THREE MAMMAL LINEAGES

PLoS Pathogens, 2015 vol. 11 (11) p. e1005279. Cross-species transmission and differential fate of an endogenous retrovirus in three mammal lineages. Xiaoyu Zhuo, Cédric Feschotte.

PLOS | PATHOGENS

RESEARCH ARTICLE

Cross-Species Transmission and Differential Fate of an Endogenous Retrovirus in Three Mammal Lineages

Xiaoyu Zhuo, Cédric Feschotte*

Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, Utah, United States of America

* cedric@genetics.utah.edu

OPEN ACCESS

Citation: Zhuo X, Feschotte C (2015) Cross-Species Transmission and Differential Fate of an Endogenous Retrovirus in Three Mammal Lineages. PLoS Pathog 11(11): e1005279. doi:10.1371/journal.ppat.1005279

Editor: Robert Belshaw, Plymouth University, UNITED KINGDOM

Received: July 5, 2015

Accepted: October 23, 2015

Published: November 12, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The work is supported by grant R01-GM077582 from National Institutes of Health (http://www.nih.gov/). XZ is also supported by University of Utah graduate research fellowship (https://gradschool.utah.edu/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Endogenous retroviruses (ERVs) arise from retroviruses chromosomally integrated in the host germline. ERVs are common in vertebrate genomes and provide a valuable fossil record of past retroviral infections to investigate the biology and evolution of retroviruses over a deep time scale, including cross-species transmission events. Here we took advantage of a catalog of ERVs we recently produced for the bat *Myotis lucifugus* to seek evidence for infiltration of these retroviruses in other mammalian species (>100) currently represented in the genome sequence database. We provide multiple lines of evidence for the cross-ordinal transmission of a gammaretrovirus endogenized independently in the lineages of vespertilionid bats, felid cats and pangolin ~13–25 million years ago. Following its initial introduction, the ERV amplified extensively in parallel in both bat and cat lineages, generating hundreds of species-specific insertions throughout evolution. However, despite being derived from the same viral species, phylogenetic and selection analyses suggest that the ERV experienced different amplification dynamics in the two mammalian lineages. In the cat lineage, the ERV appears to have expanded primarily by retrotransposition of a single proviral progenitor that lost infectious capacity shortly after endogenization. In the bat lineage, the ERV followed a more complex path of germline invasion characterized by both retrotransposition and multiple infection events. The results also suggest that some of the bat ERVs have maintained infectious capacity for extended period of time and may be still infectious today. This study provides one of the most rigorously documented cases of cross-ordinal transmission of a mammalian retrovirus. It also illustrates how the same retrovirus species has transitioned multiple times from an infectious pathogen to a genomic parasite (i.e. retrotransposon), yet experiencing different invasion dynamics in different mammalian hosts.

Author Summary

The cross-species transmission of viruses poses a continuous threat to public health. Bats are increasingly recognized as a major reservoir for zoonotic RNA viruses, including

rabies, Ebola, and possibly MERS, but little is known about their capacity to harbor and transmit retroviruses. Here we investigated past incidents of cross-species transmission involving bat retroviruses, by screening for the presence of endogenous retroviruses (ERVs) previously identified in the genome of the little brown bat in more than 100 diverse mammal species. This screen revealed an intriguing case of a gammaretrovirus that independently infiltrated the germ line of species belonging to three mammalian orders: vesper bat, felid cat and pangolin. We found that the ERV initiated its genomic invasion of the three lineages around the same timeframe ~13–25 million years ago, but experienced a different fate in each lineage. In the pangolin lineage, the ERV's genomic propagation stalled shortly after endogenization, while it amplified continuously throughout felid and vesper bat evolution to generate hundreds of species-specific insertions in each lineage. Furthermore, in the cat lineage genomic amplification appears to have occurred predominantly via retrotransposition; while in bats the ERV has expanded via a mixture of retrotransposition and reinfection activity that may still be ongoing.

Introduction

Viral cross-species transmission (CST) represents a major threat to both human and animal populations. Most viral diseases of humans are zoonotic: they stem from CST of viruses from domestic or wild animals [1]. The explosion and development of human society, including modern transportation, over the last 100 years has exposed us to an increasing number of pathogens [2]. AIDS, which has caused more than 25 million deaths over the past ~30 years (aids.gov), is one of the most notorious examples of a pandemic initiated by viral CST [3,4]. The pathogens causing AIDS (HIV-1 and HIV-2) are retroviruses, a family of RNA viruses that use reverse transcription to replicate their genome [5]. Other retroviral CST events have been documented within primates, felids and ruminants, suggesting that retroviral CST represents a continuous threat to human and animal health [6–10].

Retroviruses are unique amongst animal viruses in that chromosomal integration of so-called proviruses is an obligatory step in their replication cycle [5]. As a consequence, retroviral infection of germ cells or their progenitors result in proviruses that may be vertically inherited along with the host genome. Such inheritable proviruses are called endogenous retroviruses (ERVs). Under some circumstances, which are still poorly understood, ERVs can further propagate within the genome and spread in the population, resulting in the formation of large families of interspersed repeats in the host genome [11]. Despite the potentially deleterious consequences associated with the genomic propagation of ERVs, the process has been remarkably pervasive during mammalian evolution. Indeed every mammalian genome thus far examined harbor a great abundance and diversity of ERVs, which are mostly lineage-specific. For example, 8% of the human genome is composed of ERV sequences derived from a wide variety of retroviruses acquired at different time points during primate evolution [12–14]. Once integrated and endogenized, most ERVs appear to evolve at the host's neutral mutation rate, which is much slower than the mutation rate of exogenous retroviruses (XRVs) [15]. Therefore ERVs provide a valuable fossil record of past retroviral infections and a unique opportunity to investigate retroviral evolution at a deep time scale, including CST events [16–20].

Many ancient CST events have been inferred by comparing ERV sequences across species [21–28]. Most of the well-documented cases of retroviral CST events involve closely related host species (e.g. from the same order). Indeed, it is thought that viral CST is often constrained by the evolutionary distance between donor and recipient species [19,20,29]. The observation

that all retroviruses known to infect humans have been acquired from other primates is consistent with this notion [7]. However, retroviral CST events can also occur between distantly related species. For example, the cat RD114 gammaretrovirus is a recombinant containing an envelope domain mostly closely related to Baboon endogenous virus (BaEV), and is thought have been acquired by the domestic cat from an Old World monkey [30,31]. Also, the koala retrovirus (KoRV), which is currently spreading and undergoing endogenization in the wild, is very closely related to gibbon ape leukemia virus (GALV) and to ERVs found in Asian rodents, from which it was most likely acquired [32]. It has also been reported that reticuloendotheliosis virus (REV) was likely transmitted from mammals to birds [10]. Recent phylogenomics surveys of ERVs across a wide variety of vertebrate species suggested that CST between widely diverged species (i.e. from different orders or classes) may be more common than initially anticipated [19,20,33,34]. However, the evidence remains limited and more detailed case studies are needed to confirm this idea.

Bats (order Chiroptera) are increasingly regarded as exceptionally potent reservoirs of zoonotic viruses [35–40]. Indeed, a variety of bat species have been implicated in the spillover of diverse and highly pathogenic RNA viruses such as Rabies, Nipah, Hendra, SARS, Marburg, and Ebola viruses in the human population [41]. Very recently, one potential case of CST of an endogenous betaretrovirus involving phyllostomid bats, rodents and New World monkeys was reported [28]. We previously produced a comprehensive catalog of ERVs in the vespertilionid bat *Myotis lucifugus* [42] (referred to as MLERVs hereafter), documenting a rich and recent history of retroviral infections in this species lineage. Here, we have taken advantage of this resource to seek evidence of CST events implicating MLERVs. We identified an intriguing case of a gammaretrovirus that colonized independently the genomes of vespertilionid bats, felids and pangolin but followed a different fate and amplification dynamics in these lineages.

Results

ERVs closely related to MLERV1 are present in species from three mammal orders

To detect possible CST events involving *M. lucifugus* ERVs, we used the sequence of the reverse transcriptase domain (RVT_1) (642 nt) from members of each of the 86 MLERV subfamilies previously identified [42] as queries in megaBLAST searches of all mammal genomes deposited in the NCBI whole genome shotgun (WGS) database as of February 2015 (107 mammal species). Excluding hits to *M. lucifugus*, the most significant hits (>80% nucleotide identity over the entire domain; e-value $< 10^{-80}$) were obtained with a query representing the MLERV1 family [42] against the genome assemblies of the domestic cat (*Felis catus*) [43], Amur tiger (*Panthera tigris*) [44] and Chinese pangolin (*Manis pentadactyla*). In addition, and less surprisingly, many highly significant hits to MLERV1 were also obtained in the genomes of vespertilionid bat species closely related to *M. lucifugus* (Brandt's myotis, *Myotis brandtii* [45]; David's bat, *Myotis davidii* [46]; big brown bat, *Eptesicus fuscus*).

Further examination revealed that the hits in the feline genomes corresponded to an endogenous gammaretrovirus family initially described in the domestic cat. Two proviruses of this family were initially documented in cat as FERVmlu1 and FERVmlu2 [47]. In 2011, this ERV family was also reported in Repbase [48] as ERV1-1_Fca. In a more recent and more systematic inventory of ERVs in the cat genome [49], this family was designated as FcERV_γ6, a nomenclature we will adopt hereafter. Most recently, this family was identified as part of "lineage VII" by Mata et al. [34] who also reported the presence of closely related gammaretroviral elements in several wildcat species, including jaguar, puma, jaguarundi and tiger. To our knowledge, the related elements in the pangolin have not been previously characterized elsewhere. Hereafter we refer to this novel ERV family as MPERV1 for *Manis pentadactyla* ERV1 and deposited its

Table 1. Copy number of MLERV1 related proviruses in different species.

	Tiger	Cat	*E. fuscus*	*M. davidii*	*M. brandtii*	*M. lucifugus*	pangolin
Full-length ERV	55	88	3	48	51	204	2
Solo LTR	675	744	67	1042	948	1638	27
total	730	832	70	1090	999	1842	29

doi:10.1371/journal.ppat.1005279.t001

consensus sequence in Repbase. For simplicity, we refer to all the elements detected in vespertilionid bats as MLERV1 and all the elements in different felids as FcERV_γ6.

To determine the ERV copy number in each species, we used the LTR sequences to mask their corresponding genome assembly using the Repeatmasker program and parsed the positional output to infer the number of putative full-length proviruses (i.e. containing two LTRs) and solitary (solo) LTR (see Methods). The results of this analysis (Table 1) show that each species harbors a relatively small number of full-length proviruses (2–50) but often numerous solo LTRs (up to 1600 + in *M. lucifugus*). It should be noted that the vast majority of proviruses we inferred to be full-length (based on the occurrence of a pair of LTRs within 10 kb) contain sequencing/assembly gaps. Thus we cannot ascertain whether they contain all the coding domains of a complete provirus.

A retroviral CST event involving bat, cat and pangolin

To illustrate the exceptional level of sequence similarity among MLERV1, FcERV_γ6 and MPERV1, we generated nucleotide pairwise alignments of FcERV_γ6 and MLERV1 and of FcERV_γ6 and MPERV1 using the most closely related full-length proviruses from each family and performed a sliding window analysis of nucleotide identity across the two pairwise alignments (Fig 1A). As a comparison, we performed the same analysis for proviral sequences representative of HIV-1 (Group M subtype B) and its closest relative from the chimpanzee SIVcpz [50,51]. The results show that the two representatives of the MLERV1 and FcERV_γ6 families and two representatives of FcERV_γ6 and MPERV1 are highly similar throughout their entire length, with an average level of nucleotide identity (~85%) comparable to that between HIV-1 and SIVcpz (Fig 1A). The most divergent segment corresponds to the predicted surface (SU) domain of the envelope protein (~50% identity in the N-terminal region). Elevated divergence in the SU region is also apparent between the two lentiviruses, as previously documented [52], and is thought to reflect the rapid adaptation of retroviral envelope to diverged host cell receptors [53]. In summary, MLERV1, FcERV_γ6 and MPERV1 are just as closely related to each other as HIV-1 and SIVcpz, and thus these three elements and their relatives in the bat, cat and pangolin genomes can be considered as endogenous elements descended from the same retrovirus.

The overall level of nucleotide similarity between MLERV1, FcERV_γ6 and MPERV1 is strongly incongruent with a scenario of vertical inheritance of an ancestral ERV present in the common ancestor of chiropterans, felids and pangolins, which dates back to ~85 million year ago (MYA) [54,55]. Furthermore, we could not find any close relative of MLERV1 or FcERV_γ6 (no megaBLAST hit with sequence identity >80%) in the genome assemblies of species representative of other chiropteran (e.g. flying fox, Pteropodidae) or carnivore families (e.g. dog, Canidae; bear, Ursidae; ferret, Mustelidae; seal, Phocidae; walrus, Odobenidae). The Chinese pangolin genome is the only available representative of the order Pholidota, which is considered sister to Carnivora, and thus equally related to Perissodactyla (horse, rhino) and Cetartiodactyla (cow, pig, hippo, whales), all of which appear to lack related ERVs (Fig 1B). Thus, the taxonomic distribution of MLERV1/FcERV_γ6 elements is extremely patchy, being detected in four vespertilionid bats, two feline species (cat and tiger), and one pangolin, but not in any of the numerous phylogenetically intermediate species represented in the NCBI WGS

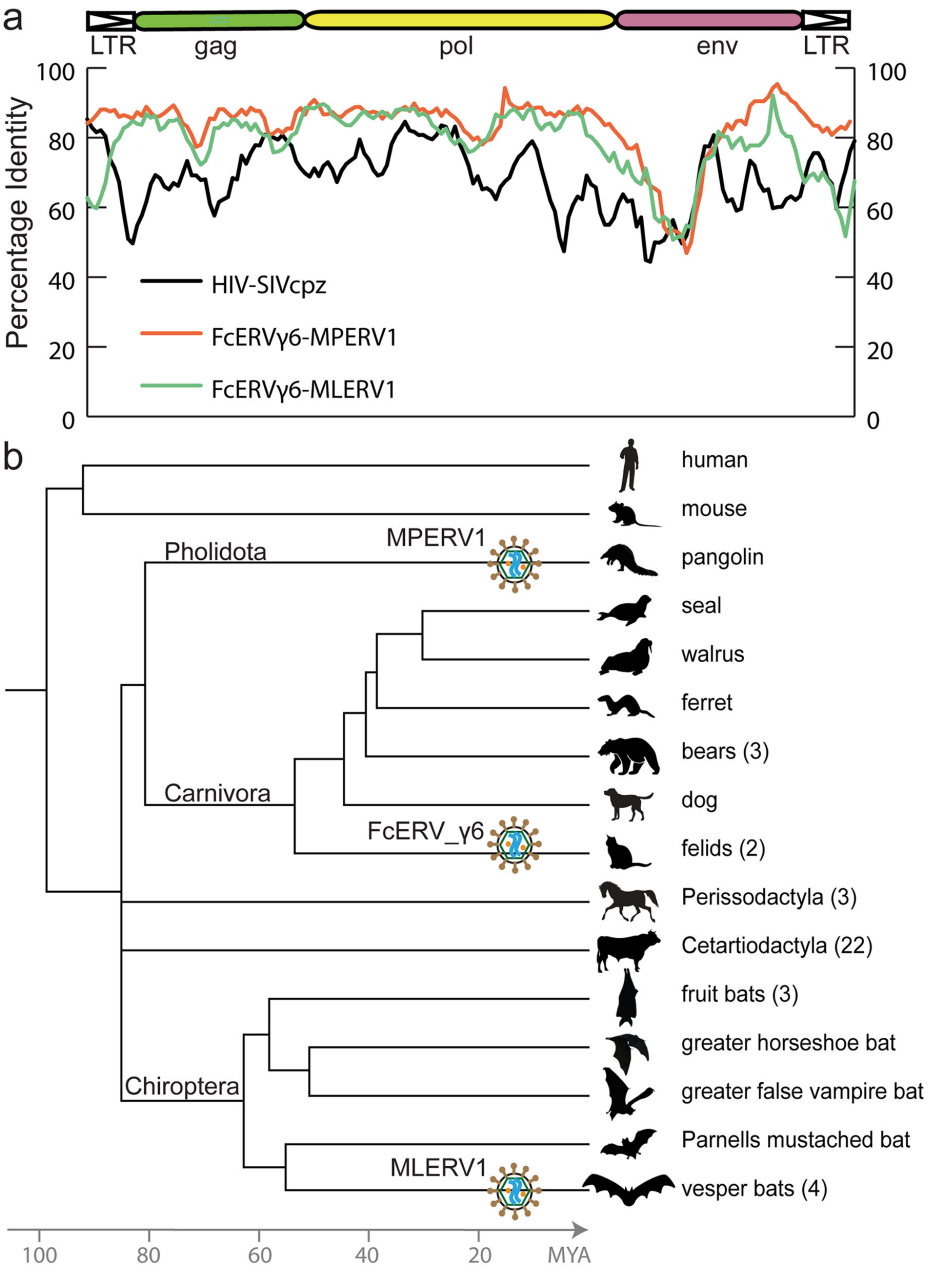

Fig 1. High sequence similarity and taxonomic distribution of MLERV1, FcERV_γ6 and MPERV1. (a) Sliding window analysis of percent sequence identity along pairwise alignments of entire proviruses. DNA sequence distance is corrected using kimura 2 parameter substitution model. (b) Taxonomic distribution of MLERV1, FcERV_γ6 and MPERV1. A schematic of the phylogenetic relationship of the 55 species from the clade "Scrotifera" currently

represented in the NCBI whole genome sequence database, with human and mouse shown as outgroups. The 55 species fall within 6 mammal orders: Pholidota, Carnivora, Cetacea, Artiodactyla, Perissodactyla, Chiroptera. Some of the species are collapsed by order/family with the number of species for each clade indicated into parentheses. The three independent retroviral invasions of MLERV1, FcERV_γ6 and MPERV1 are depicted above each of the mammal lineages affected. The placement of retroviral particles does not imply the timing of corresponding CST.

doi:10.1371/journal.ppat.1005279.g001

database (Fig 1B). This taxonomic distribution suggests that the retrovirus that gave rise to MLERV1, FcERV_γ6 and MPERV1 underwent at least two CST events and was endogenized at least 3 times independently in the vespertilionid, felid, and pangolin lineages.

Dating MLERV1/FcERV_γ6 insertions using comparative genomics

To gain further insights into the evolutionary history of these ERVs, we next sought to estimate when they first infiltrated their host genomes. Given that the likelihood of the same endogenous retrovirus to integrate at the same exact genomic location independently in different lineages is negligible, the presence of an element at orthologous position in different species can be interpreted as having inserted prior to their divergence time [56,57]. Conversely, since ERVs are not known to excise from the genome, the absence of an element in one species at a genomic location occupied by an ERV in another species strongly suggests that the ERV integrated after the split of the two species [58,59]. Such 'empty' sites can be corroborated by the presence of a single copy of the host target sequence duplicated upon proviral integration (typically 4-bp target site duplication for gammaretroviruses). This cross-species presence/absence approach has been widely applied to date a variety of mobile element insertions, including ERVs [14,42,59,60]. It is possible that the age of some integration events may be underestimated because of incomplete lineage sorting. Therefore, orthologous insertion analysis should be interpreted with caution when applied to rapidly radiating species such as the three *Myotis* considered here.

We first examined the sharing of FcERV_γ6 elements between the cat and tiger, which diverged ~10.8 MYA [61]. Out of a total of 1,419 putative full length proviruses and solo LTRs detected in the current whole genome assemblies of the two species, we were able to ascertain that 256 occupy orthologous positions, while 261 and 201 are specific to the cat and tiger lineages, respectively. None of these elements were detectable in other available carnivore genome assemblies (e.g. dog, panda, ferret, seal), while some of their flanking host sequences were readily detected (e.g., the flanking sequence of FcERV_γ6–68 is found in dog chromosome 14). These data indicate that FcERV_γ6 first invaded a felid ancestor sometime between ~10.8 million years (MY) and ~55 MYA and has continued to amplify to generate many insertions specific to the cat and tiger lineages (Fig 2).

Our previous phylogenetic analysis [42] has shown that the MLERV1 family of the little brown bat *M. lucifugus* can be divided into 3 subfamilies. Here we performed a systematic analysis of the presence/absence of MLERV1 elements (including solo LTRs) from the 3 subfamilies in the genome assemblies of three other vespertilionid bats currently available: Brandt's myotis (*Myotis brandtii*), David's bat (*Myotis davidii*) and the big brown bat (*Eptesicus fuscus*), which have been estimated to diverge from *M. lucifugus* ~10 MYA, ~13 MYA and ~25 MYA, respectively [62–65]. The vast majority of MLERV1 elements and their close relatives were found to be species-specific (Fig 2). Only 3 elements were present at orthologus loci across the 3 *Myotis* genomes (Fig 2) and we could not find a single insertion shared between *E. fuscus* and any of the *Myotis*. Another interesting observation is that members of the MLERV1_3 subfamily, which contributes the vast majority (>80%) of MLERV1 elements in the 3 *Myotis* genomes, could not be identified at all in the *E. fuscus* genome. Indeed, all 29 elements detected in *E. fuscus* cluster with either one of the other two subfamilies (S6 Fig). Together these data suggest

that the MLERV1 family expanded independently in the *Myotis* and *Eptesicus* lineages, but achieved a much higher copy number in the *Myotis* lineage due to the amplification of the MLERV1_3 subfamily, which has generated numerous species-specific insertions (Fig 2).

Dating of individual provirus insertions using LTR-LTR divergence

Another widely applied method to date retroviral and other LTR-bearing retroelement insertions relies on the divergence of the 5' and 3' LTR of individual elements. This is because their retrotransposition mechanism results in two identical LTRs at the time of chromosomal integration. Given that most ERV LTR sequences are assumed to evolve neutrally once integrated in the host chromosome, the age of a provirus can be estimated based on LTR divergence by applying the host neutral substitution rate [49,59,66,67]. To eliminate the inflated divergence caused by hypermutable methylated CpG sites [68], we excluded all the CpG sites from our calculation of LTR-LTR divergence. We applied this method to calculate the age of all complete (i.e. with two LTR) proviruses detected in cat, *M. lucifugus* and pangolin genome assemblies. We use previously estimated neutral substitution rates of 2.7×10^{-9} and 1.8×10^{-9} per year for vespertilionid bats and felids respectively [44,69], and an "average" mammal neutral substitution rate of 2.2×10^{-9} per year [70] for the pangolin.

The results of these calculations predict that the oldest MLERV1 and FcERV_γ6 proviruses would be ~10 MY and ~20 MY respectively (Fig 3A). The amplification of the bat MLERV1 family would have peaked sharply in the last 2 MY, while the cat FcERV_γ6 elements inserted more continuously over the past ~15 MY (Fig 3A). The two MPERV1 proviruses identified in the pangolin genome are estimated to be ~10 and ~18 MY based on this approach.

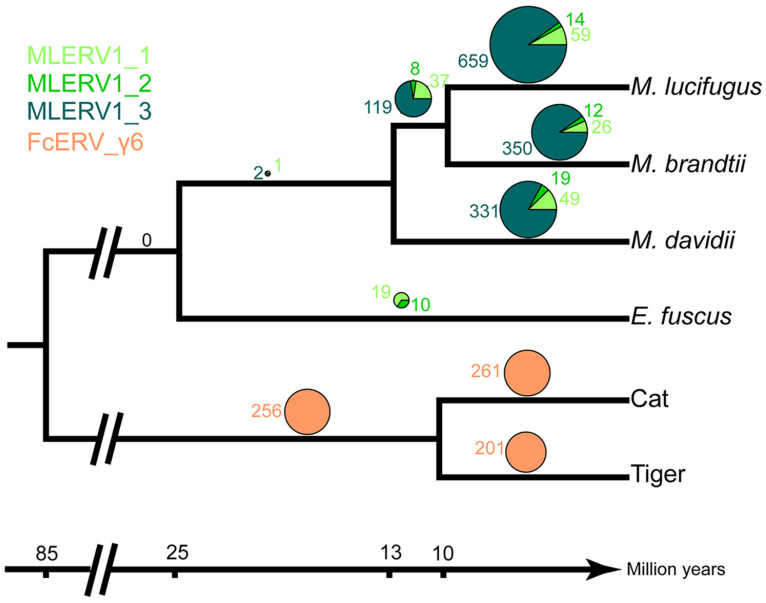

Fig 2. Distribution of MLERV1/FcERV_γ6 insertions in vesper bats and felids. The numbers of ERV insertions detected as orthologous or species-specific are shown as pies above each branch of the phylogeny of the vesper bats and felids examined. Different colors are used to illustrate FcERV_γ6 and the three different MLERV1 subfamilies.

doi:10.1371/journal.ppat.1005279.g002

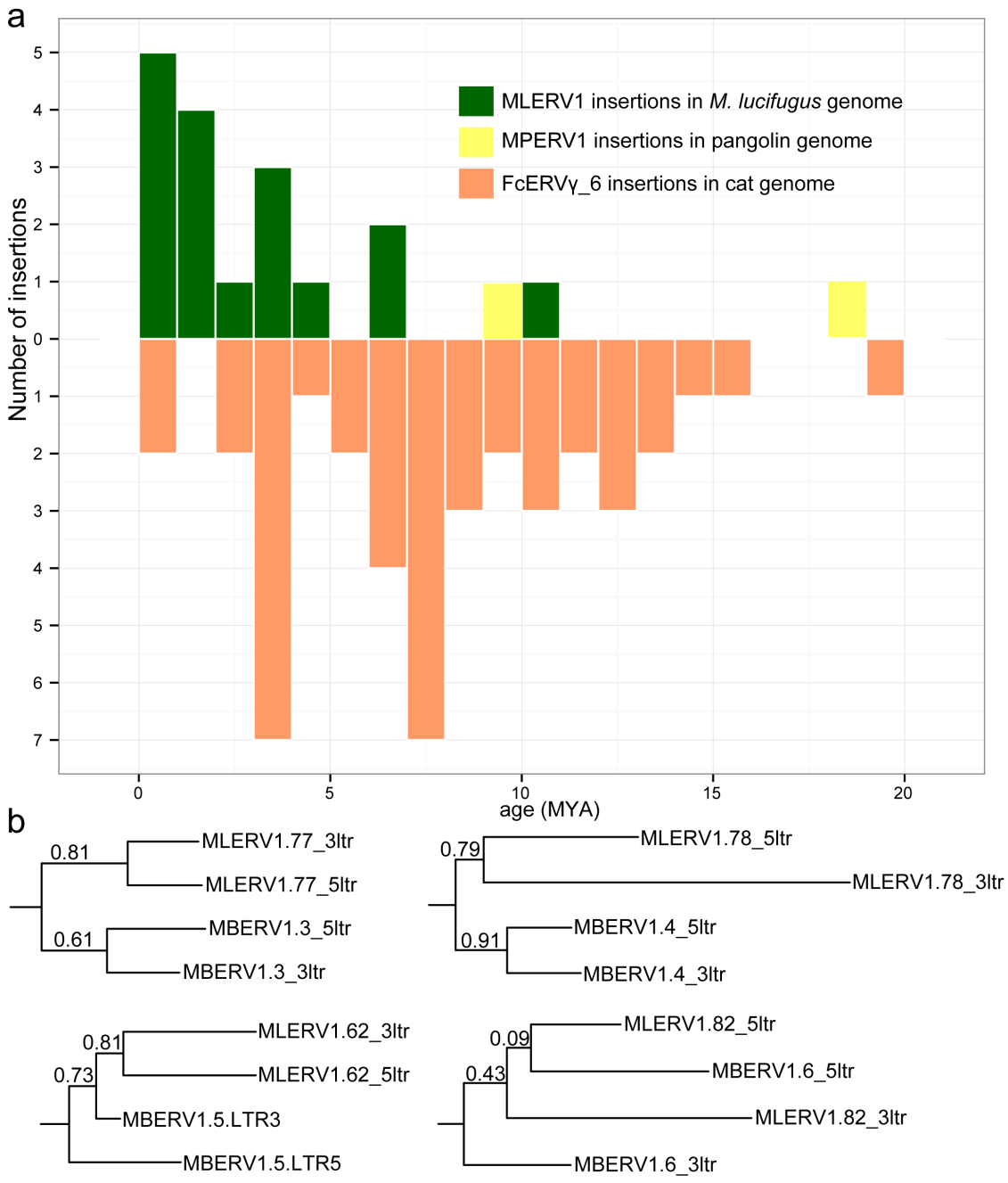
a
MLERV1 insertions in M. lucifugus genome
MPERV1 insertions in pangolin genome
FcERVγ_6 insertions in cat genome
Number of insertions
age (MYA)
b
0.81
MLERV1.77_3ltr
MLERV1.77_5ltr
0.61
MBERV1.3_5ltr
MBERV1.3_3ltr
0.79
MLERV1.78_5ltr
MLERV1.78_3ltr
0.91
MBERV1.4_5ltr
MBERV1.4_3ltr
0.81
MLERV1.62_3ltr
0.73
MLERV1.62_5ltr
MBERV1.5.LTR3
MBERV1.5.LTR5
0.09
MLERV1.82_5ltr
MBERV1.6_5ltr
0.43
MLERV1.82_3ltr
MBERV1.6_3ltr

Fig 3. Dating individual proviral insertions based on LTR-LTR divergence. (a) Age distribution of proviral insertions inferred from LTR-LTR divergence. The *y* axis shows the number of insertions for each age class binned in MY on the *x* axis. Each ERV family is shown as bars of different colors. (b) Evidence of 'gene' conversion between 5' and 3' LTR of the same provirus. Four LTR trees are shown for four pairs of orthologous proviruses shared by *M. lucifugus* (MLERV) and *M. brandtii* (MBERV). Each maximum likelihood tree was built from a multiple alignment of the 5' and 3' LTRs from each provirus rooted with a non-orthologous LTR from *M. lucifugus* (also illustrated in S2 Fig). The support for each node as determined with an approximate likelihood ratio test (aLRT) is shown. The fact that 5' and 3' LTR from the same provirus tend to group together rather than by species is indicative of gene conversion between the LTRs along the two species lineages following proviral insertion in their common ancestor.

doi:10.1371/journal.ppat.1005279.g003

While these estimates are consistent with independent ERV invasions of the vespertilionid, felid and pangolin lineages, we noticed that the age of individual insertions based on LTR divergence were generally lower than those estimated based on their presence/absence at orthologous position across species. For instance, we found that 27 FcERV_γ6 proviruses were orthologous in cat and tiger, which indicates that all must have inserted prior to speciation of these felids, which has been robustly estimated at ~10.8 (8.4–14.5) MY [61]. However only 13 of these 27 insertions were estimated to be older than 10 MY based on LTR divergence (S1 Table). Similarly, *M. lucifugus* and *M. brandtii* are thought to have diverged ~10 MYA [62,64], but the age of the four MLERV1 insertions orthologous between these two species was estimated to be 10.5, 6.8, 4.5 and 1.2 MY based on LTR divergence (S1 Table). One possible explanation for these discrepancies between the two dating methods is the phenomenon of gene conversion between two LTRs adjacent in the genome, which essentially erases some of the divergence accumulated over time through point mutations occurring in each of the LTRs, causing to underestimate the date of proviral insertion [71,72]. Indeed, a phylogenetic analysis of the LTR sequences from the four MLERV1 proviruses orthologous in *M. lucifugus* and *M. brandtii*, shows topologies consistent with LTR homogenization through gene conversion events for at least two of the proviruses examined (corresponding to the two upper trees in Fig 3B): their 5' and 3' LTR cluster together rather than by species (Fig 3B). This is the topology predicted if conversion events in one or both of the species lineages had removed nucleotide divergence accumulated between 5' and 3' LTR prior to speciation [72]. Thus, estimates of the age of proviruses based on LTR divergence should be interpreted with caution, as they are likely to be underestimates. Nonetheless, the results are in agreement with the other lines of evidence that the vespertilionid, felid, and pangolin lineages were independently infiltrated by the same ERV during an evolutionary timeframe ranging from ~25 to ~13 MYA.

Phylogenetic analysis of FcERV_γ6, MLERV1 and MPERV1 families

To further characterize the evolution history of MLERV1, FcERV_γ6 and MPERV1, we examined the phylogenetic relationship of elements within these families using a maximum-likelihood tree built from an alignment of their 3' LTR sequences. We used only 'complete' proviruses (30 in bats, 43 in cats and 2 in pangolin), since we observed that including tiger provirus and solo LTRs did not yield any new major clade in the phylogeny (S1 Fig). Also the general topology of the tree is identical if the 5' LTR sequences are included (S2 Fig). Trees generated using internal coding sequences also displayed the same general topology (S3 Fig), but offered less phylogenetic resolution due to the more constrained nature of retroviral coding sequences relative to LTRs [73,74].

The unrooted tree resulting from the phylogenetic analysis (Fig 4) clearly shows that FcERV_γ6 and MPERV1 elements are more closely related to each other than to the bat MLERV1 elements. Another striking observation is that elements within the FcERV_γ6 family fall within a single clade with uniformly short branches, whereas the MLERV1 elements, as we previously reported [42] can be divided into 3 distinct subfamilies separated by long branches, with MLERV1_2 and MLERV1_3 being closer to each other and more distant from FcERV_γ6

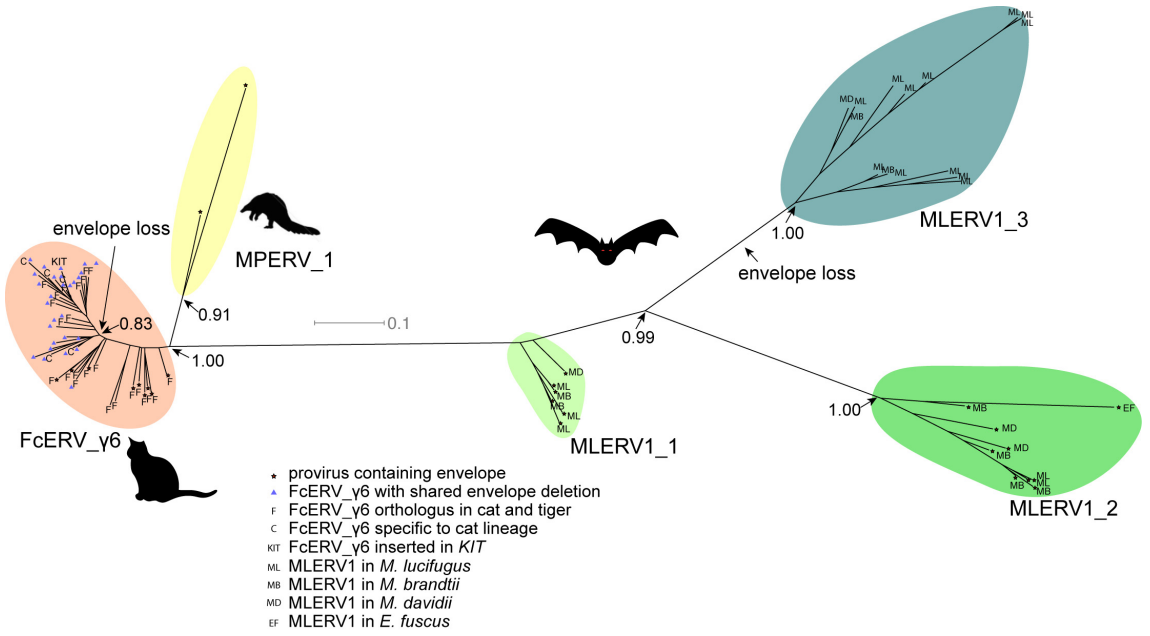

Fig 4. Phylogenetic analysis of MLERV1, FcERV_γ6, MPERV1 families. A maximum likelihood phylogenetic tree built from a multiple alignment of 3' LTR sequences of 75 proviruses. The support for each node as determined with an approximate likelihood ratio test is shown. Information on the species origin as well as the presence/absence of envelope sequence is labeled at each node. Two independent losses of envelope by deletion are highlighted.

doi:10.1371/journal.ppat.1005279.g004

than MLERV1_1 (Fig 4). These data are consistent with a scenario whereby the FcERV_γ6 family was amplified from a single infectious progenitor, while MLERV1 elements might have originated from at least three distinct infectious progenitors.

Selection analysis on coding sequences reveal different amplification dynamics

To further explore the history of the FcERV_γ6 and MLERV1 families, we next turn to an analysis of selection regimes that have acted on their coding sequences during their amplification. Such analysis can help discern whether ERVs have spread primarily through reinfection or retrotransposition events because the latter mechanism, which is strictly intracellular, is predicted to be associated with the loss of envelope function. Indeed, the envelope protein binds to host cell membrane receptor to promote virion entry in the host cell and therefore is required for most retroviral infection [5]. Thus, proviruses that originate from infection events should show evidence of functional constraint on envelope domains [75,76], Magiorkinis:2012gy}. To perform this analysis, we used all MLERV1 (n = 30) and cat FcERV_γ6 (n = 43) proviruses with complete (or nearly complete) coding capacity. Given that only 2 proviral MPERV1 copies could be identified, we did not perform selection analysis for this family.

To evaluate how natural selection may have constrained the different coding regions of MLERV1 and FcERV_γ6, we computed the dN/dS ratio (ω) applying the branch model implemented in PAML, where dN denotes the non-synonymous substitution rate and dS denotes the synonymous substitution rate, along the branches of the phylogeny of FcERV_γ6 and

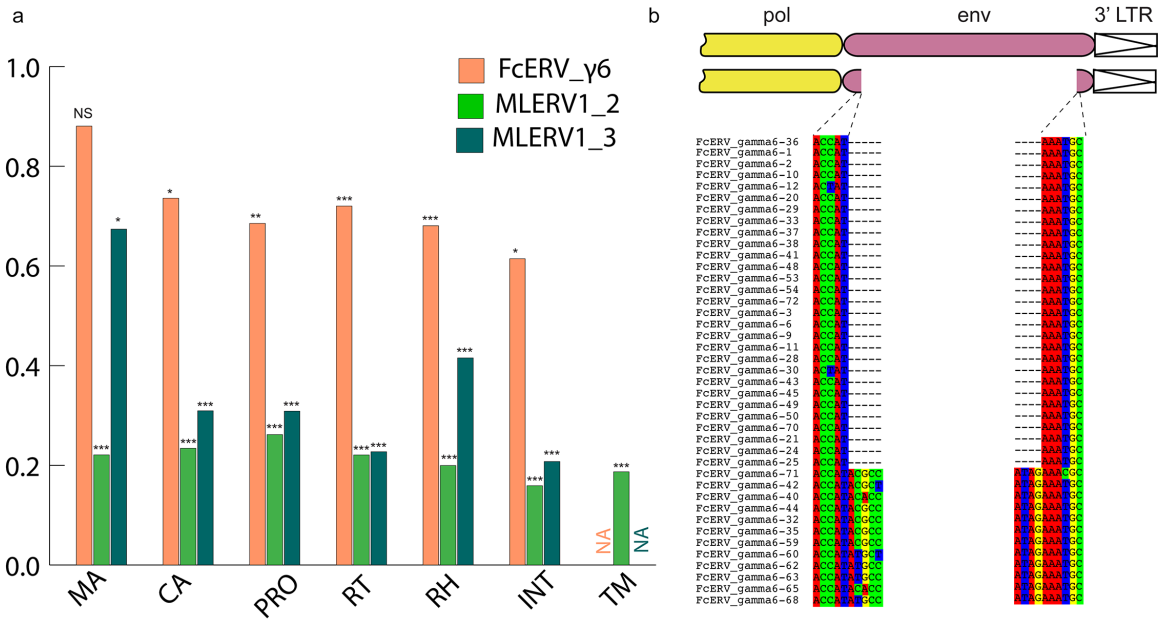

Fig 5. Selection analysis on coding domains. (a) dN/dS ratio (ω) of each coding domain in FcERV_γ6, MLERV1_2 and MLERV1_3. MA, CA, PRO, RT, RH, INT and TM denote matrix, capsid, aspartyl protease, reverse transcriptase, RnaseH, integrase and envelope transmembrane domain, respectively. Asterisks denote the level of significance of departure from ω = 1 (likelihood ratio test, see Methods) with * = p<0.05; ** = p<0.01; *** = p<0.001. NS = not significant (p>0.05); NA = not applicable (domain deleted). (b) Shared breakpoints at the site of envelope deletion in a subset of FcERV_γ6 elements. A schematic of the prototypical proviral coding regions showing the approximate position of the envelope deletion in 29 FcERV_γ6 elements marked with blue triangles in Fig 4 and (below) an alignment with a subset of envelope-containing FcERV_γ6 elements, showing that they share the same deletion breakpoints. These data indicate that these 29 elements likely arose from amplification of a progenitor copy that had suffered a large deletion in the envelope region.

doi:10.1371/journal.ppat.1005279.g005

MLERV1 elements for each of their predicted coding domains [77]. ω values significantly smaller than 1 are indicative of purifying selection acting to maintain a functional protein sequence, while ω values not significantly different from 1 are indicative of neutral evolution or relaxed functional constraint. To test for significant deviation of ω from 1, we apply a likelihood ratio test [78].

Within the FcERV_γ6 family, the analysis reveals that purifying selection has acted on all coding domains (ω value ranging from ~0.6 to ~0.9, $p < 0.05$), with the notable exception of Gag matrix and envelope domains (Fig 5A). The ω value is not significantly different from 1 (neutral evolution) for the Gag matrix domain (Gag_MA). Besides, all but nine of the 43 FcERV_γ6 proviruses lack an envelope domain (TLV_coat). The nine copies that have retained a recognizable envelope domain occupy basal branches in the phylogeny (Fig 4) and have orthologs in the tiger genome suggesting that they predate the envelope-less copies (S1 Table). Furthermore, 29 of the 43 FcERV_γ6 proviruses examined, including all cat-specific copies, share the same deletion breakpoint removing most of envelope gene (Fig 5B). These data suggest that FcERV_γ6 copies potentially coding an envelope were inserted prior to the speciation of cat and tiger (~10.8 MYA), while copies integrated more recently lacked the envelope domain. In addition, the envelope open reading frames of these nine ancient FcERV_γ6 elements accumulated multiple indels or missense substitutions. Thus, none of the FcERV_γ6

elements in the cat genome appear to have retained a functional envelope domain. These data suggest that FcERV_γ6 rapidly lost its infectious capacity in the cat lineage but has continued to amplify primarily via retrotransposition amplified primarily via retrotransposition.

By contrast, selection analysis suggests that the MLERV1 family has experienced a more complex amplification history. We focused our analysis on the MLERV1_2 and MLERV1_3 subfamilies because they are the two best-supported monophyletic subfamilies with sufficient number of proviruses to draw solid conclusions. First, we observe that generally the signature of purifying selection is more pronounced on the bat elements than on the cat elements, as indicated by much lower ω values (Fig 5A). The only exception is the Gag matrix domain of the MLERV1_3 subfamily, which exhibits relatively higher ω value ($\omega = 0.67$, $p = 0.02$) (Fig 5A). In addition, all MLERV1_3 elements appear to have lost their envelope domain through the same deletion event (Fig 4). This pattern contrasts with elements within the MLERV1_2 subfamily, for which all coding domains, including envelope, have evolved under strong purifying selection during the spread of these elements (ω from 0.15 to 0.27, $p<0.001$) (Fig 5A). These data suggest a scenario whereby MLERV1_3 has amplified primarily by retrotransposition, while the spread of MLERV1_2 has been driven by multiple infection events.

We also observe that in both FcERV_γ6 and MLERV1_3, the losses of envelope coincided with the elevation of the dN/dS ratio in their Gag matrix domain (Fig 5A). To evaluate whether this reflects a loss of function (neutral evolution) or a relaxation of purifying selection, we further examined the integrity of open reading frames (ORFs) in each ERV family by computing the frequency of stop codons and frameshift mutations occurring in each of the domains (see Methods). Overall the results indicate that the coding integrity of the Gag matrix domains of FcERV_γ6 and MLERV1_3 elements is not significantly different from that of the other ERV subfamilies or that of the other protein domains (S4 Fig). These results suggest that the Gag matrix domain is not dispensable for retrotransposition, as previously demonstrated functionally for IAP elements [79], but appears to evolve faster in retrotransposing ERVs.

Discussion

Cross-ordinal transmission of a mammalian retrovirus

Until recently, most retroviral CST events that have been documented rigorously have implicated closely related species [6,9], suggesting that the phylogenetic distance between species is an important determinant of the host range of a retrovirus [7,29]. Indeed, many previous studies have illustrated how the divergence of host cellular factors that either facilitate or restrict viral replication can modulate the host range of a virus [80–82]. The systematic analysis of retroviruses fossilized in the genome as ERVs is progressively revealing a more nuanced picture whereby some retroviruses appear to have been capable to infect widely diverged species (i.e. belonging to different orders) without seemingly much changes occurring in their own sequences. Recent large-scale pylogenomics analyses have suggested that cross-order transmission may actually be fairly common for some groups of retroviruses, including gammaretroviruses [19,20,34] and IAP betaretroviruses [33]. While these studies disclosed phylogenetic patterns suggestive of multiple CST events, they did not explicitly rule out alternative hypotheses, such as vertical persistence and stochastic loss of the ERV in some lineages, and thus they generally await confirmation through more detail analyses such as the one presented here.

Our study provides multiple lines of evidence supporting the notion that a gammaretrovirus infiltrated independently the germline of bat, cat and pangolin species representing three mammalian orders (Chiroptera, Carnivora, Pholidota, respectively). First, elements found in these species display a level of nucleotide sequence similarity (~85%) along their entire length that is comparable to that observed between closely related retroviruses that have undergone very

recent CST (such as SIVcpz and HIV-1). Such a level of sequence similarity between ERVs inhabiting species diverged by ~85 MYA [54] is incompatible with a scenario of vertical descent from an ERV inherited from their common ancestor. The CST hypothesis is also bolstered by the highly discontinuous taxonomic distribution of this particular ERV family. Out of 107 mammal species for which whole genome assemblies are publicly available, we could only detect members of this ERV family in vespertilionid bats, felids and pangolin, but not in several species representing related mammal families (6 additional Chiroptera species from 4 families and 7 additional Carnivora species from 5 families). Thus, a scenario evoking a single introduction of this ERV family in the common ancestor of bats, cats and pangolins followed by vertical inheritance would necessitate at least 5 independent losses (Fig 1B) to account for its current taxonomic distribution. A more parsimonious scenario is that this ERV family was acquired horizontally and independently in each of the three species lineages where it is currently detected. It is also possible that the ancestral retrovirus infected other species but failed to endogenize in their genomes, or it could also be that additional species lineages hosted this ERV family but have gone extinct. Finally, our estimation of the dates at which these elements first entered their host genomes, which relies on two independent approaches (cross-species comparison of orthologous ERV loci and LTR-LTR divergence), converges to a bracket of 13 to 25 MYA, which far postdates the divergence of their host species (~85 MYA). Together these data indicate that a progenitor gammaretrovirus infiltrated the germlines of ancestral vespertilionid bat, felid and pangolin species.

It is conceivable that this retrovirus could have transferred directly between these ancestral species because their geographic distribution likely overlapped in Eurasia during the estimated period of initial ERV infiltration (~13–25 MYA) [61,62,83]. Given that cats are known to prey on both bats [84–87] and pangolins [88], a direct transfer from bat or pangolin to cat is plausible. Indeed, predation has been put forward as the most likely explanation for the spillover of bat lyssaviruses (rabies) into domestic cats [89]. On the other end, both bats and pangolins are capable of surviving a cat attack, which makes the transfer from predator to prey conceivable as well. Nonetheless, multiple lines of evidence indicate that MLERV1 colonized these bat genomes more recently (Figs 2 and 3), which may suggest a CST from cat to bat. Furthermore, we cannot rule out that one or several intermediate hosts were involved in the introduction of the retrovirus in these species.

Repeated transition from retrovirus to retrotransposon

Our data suggest that, shortly after infiltration of the felid genome, FcERV_γ6 lost the capacity to infect cells and transformed into a retrotransposon. Envelope domain remnants are only found in the basal branches in their phylogeny, and all the FcERV_γ6 elements amplified in the domestic cat lineage clearly derive from a progenitor that lacked coding capacity for a functional envelope protein (Figs 4 and 5B). Together these data suggest that FcERV_γ6 lost its infectious capability soon after it became endogenous, but continued to propagate by retrotransposition, much like the IAP elements in the mouse genome [90,91]. Coincided with the envelope loss, we found gag matrix domain evolves at a relaxed rate in FcERV_γ6 family. A recent study showed that a FcERV_γ6 insertion in the *KIT* gene currently segregating in domestic cats is responsible for the "Dominant White" and white spotting pigmentation phenotypes [92], which supports our findings that some FcERV_γ6 insertion activity is very recent and likely ongoing. Interestingly, the FcERV_γ6 element inserted at the *KIT* locus lacks envelope domain and clusters with other recently active FcERV_γ6 copies in our phylogenetic analysis (Fig 4). Collectively these data suggest that FcERV_γ6 has morphed into a successful retrotransposon that may still be active in the domestic cat.

In contrast to FcERV_γ6, the sequence diversity and phylogenetic structure of MLERV1 elements in the vesper bat genomes are indicative of a more complex amplification history characterized by a mixture of retrotransposition and reinfection events. Our phylogenetic analysis delineates at least three highly diverged MLERV1 subfamilies. The separation between three subfamilies suggested that this family stemmed from at least three related infectious progenitors independently, which is conceivable considering the independent introduction of multiple HIV-1 strains in human population [93]. Furthermore, selection analyses suggest that different subfamilies have adopted different evolutionary trajectories. The MLERV1_2 subfamily is characterized by a signature of intense purifying selection acting on all coding regions throughout the whole clade (Fig 5A). These data strongly suggest that elements within that subfamily have retained their infectious capacities for extended period of time and most likely spread primarily through reinfection events. It is even possible that MLERV1_2 is still active and infectious: most insertions are very recent (Fig 2 and S1 Table) and at least one copy (MLERV1.80) contains apparently full-length and intact *gag*, *pol*, and *env* genes.

The MLERV1_3 subfamily appears to have followed a different evolutionary path whereby the divergence of the elements was accompanied by a strong signature of purifying selection in all coding regions with the notable exception of the Gag matrix domain which has been evolving faster than other domains (Fig 5A) and the envelope domain which was apparently deleted altogether. This selection pattern resembles that of FcERV_γ6 and is indicative of proliferation primarily via retrotransposition as opposed to reinfection. Consistent with this hypothesis and the so-called superspreader model [33], the MLERV1_3 family has been by far the most successful at spreading during *Myotis* evolution: it has the highest copy number, including many species-specific insertions (Table 1).

Interestingly, none of the 29 MLERV1 elements identified in the big brown bat *E. fuscus* belong to the MLERV1_3 subfamily. This is consistent with the idea that the MLERV1_3 subfamily originated after the split of Eptesicus-Myotis split ~25 MYA and amplified during the diversification of the *Myotis* lineage. At present it remains unclear whether the MLERV1 elements present in *E. fuscus* and *Myotis* descend from element(s) introduced in their common ancestor or if they result from independent acquisition of the same retrovirus. On the one hand, the observation that both *E. fuscus* and *Myotis* harbor elements from two diverged subfamilies may be interpreted as evidence that these subfamilies descend from a single progenitor ERV acquired in the common ancestor of these species. On the other hand, the fact that none of the MLERV1 insertions are shared (orthologous) between *E. fuscus* and any of the 3 *Myotis* genomes (Fig 2) and that none of the provirus insertions dated in any of these bat species appear older than 13 MY (considerably less than the estimated divergence between the two genera, 25 MYA) (Fig 3) supports a scenario of multiple, independent acquisition. This scenario, while requiring at least two CST events, is conceivable because *Eptesicus* and *Myotis* bats likely occupied a widely overlapping geographic distribution at the estimated time of MLERV1 invasions [62] and these congeners are currently known to frequently come into contact within the same roost [94,95].

Regardless of the origin of MLERV1, the data summarized above illustrate how the same retrovirus has infiltrated widely diverged mammals and transitioned multiple times (at least twice: FcERV_γ6 and MLERV1_3) from an infectious pathogen to a genomic parasite (i.e. a retrotransposon). The biological factors and sequence of events underlying such transition remain poorly understood. In a seminal study, Ribet et al. showed that the loss of envelope gene combined to the gain of an endoplasmic reticulum targeting signal were apparently sufficient for an infectious progenitor of the mouse IAP elements to turn into a highly active retrotransposon [79]. Magiorkinis et al. [33] have extended this paradigm and proposed that the passive loss of envelope lead ERVs to become "superspreaders" in the genome. Through a study of IAP-like elements across a wide range of species, these authors observed that

envelope-less elements generally achieve much higher copy numbers than those maintaining a functional envelope. Our results support this model. First, envelope-less FcERV_γ6 elements have proliferated to high copy numbers in the domestic cat (n = 832) and tiger (n = 730). In addition, in the bats the only subfamily of MLERV1 elements that has attained similarly high copy number is MLERV1_3, which conspicuously lack a functional envelope gene (Fig 5). MLERV1_3 elements have generated many species-specific insertions consistently outnumbering the MLERV1_2 subfamily, which appears to have spread primarily by reinfection (659 vs. 14 in *M. lucifugus* genome, 350 vs. 12 in *M. brandtii* and 331 vs. 19 in *M. davidii*) (Fig 2). Thus, our study is consistent with the notion that the loss of infectious capacity correlates with ERV expansion by retrotransposition, as proposed previously for the rodent IAP families [33,79].

One important difference is that the shift between infection and retrotransposition in the MERV1 family was apparently accompanied by little changes in the sequence of MLERV1 elements. Indeed, members of the MLERV1_2 and MLERV1_3 subfamilies diverge by ~15–20% in their RT domain nucleotide sequences with Kimura correction. By comparison, infecting and retrotransposing IAP subfamilies diverge more substantially (~65% in RT domain). Thus our findings suggest that the transition between the two modes of ERV amplification can occur relatively fast during ERV evolution.

Does host biology affect ERV proliferation?

An intriguing finding of this study is that the same or very similar retrovirus was endogenized in three different mammalian hosts, but followed quite different evolutionary trajectories in the three species lineages. In the pangolin lineage, the ERV family failed to amplify (only 2 detectable copies) and was essentially 'dead-on-arrival'. In the cat lineage, the ERV progenitor apparently lost its infectious capacity shortly after endogenization and subsequently amplified to high copy numbers by retrotransposition through an extended period of time ranging from at least 10 MYA (256 insertions orthologous in cat and tiger) to modern times (*KIT* insertion segregating in domestic cats). Meanwhile, in the bat lineage, the ERV followed a more complex evolutionary path characterized by multiple episodes of reinfection, and at least one burst of amplification by retrotransposition. These observations beg the question whether the loss of infectious capacity of an ERV and its conversion to a retrotransposon is a purely stochastic process, largely owing to the stochastic mutation of gag matrix and loss of envelope functions, or possibly the characteristics of different proviral ancestors, or if it can be influenced by some biological characteristics of the host species? For instance, it has been recently reported that the level of endogenous retroviral activity may be partly governed by host body size [96]. The pattern of sustained reinfection of MLERV1 in the bat lineage is particularly intriguing in light of the growing appreciation that bats seem to frequently act as reservoir for viruses otherwise lethal to other mammals [35,39]. The reasons for bats' propensity to support high and diverse loads of viral pathogens are poorly understood, but it is thought that some physiological (e.g. immunopathological tolerance) and/or ecological features (e.g. flight, roosting) allow these animals to tolerate higher level of viral replication and/or facilitate viral transmission [39,97,98]. By the same token, it is tempting to speculate that the same properties might predispose bats to support higher level of ERV reinfection compared to other mammals such as cats. However, we only investigated one ERV family in three mammal species here. It is possible that MLERV1 is just a unique ERV family in bat genome, and overall the ERV replication in bats is not significantly different from other mammals. Testing this hypothesis will necessitate a more systematic examination of the amplification dynamics of ERVs in a wide range of mammals to assess whether the tendency toward maintenance of infectious capacity is a general trademark of bats or possibly other groups of mammals.

Methods

Initial detection of CST events involving MLERVs

Nucleotide sequences of all RVT_1 domains of previously identified MLERVs [1,42] were used as queries to search the whole genome sequence database from the National Center for Biotechnology Information (NCBI) using default MegaBLAST parameters [99]. An 80% similarity over 80% region was used as filter to exclude non-specific hits.

Identification of complete proviruses, putative full-length ERVs and solo LTRs

Complete MLERV1 and FcERV_γ6 proviruses in the *M. lucifugus* and cat genomes were collected from previous publications [42,49]. To ensure we only considered elements from these families, we only retained elements with 80% nucleotide similarity to another family member, a procedure which resulted in the exclusion of the FcERV_γ6_46 copy from the FcERV_γ6 family (S3 Fig).

To identify complete proviruses in other vesper bat genomes, the RVT_1 domain sequence of MLERV1.71 in *M. lucifugus* was used as query in blastn search of the *M. brandtii*, *M. davidii* and *E. fuscus* genome assemblies available in NCBI. In parallel, we applied LTRharvest [100] and LTRdigest [101] as described previously [42] to identify all putative proviruses in each of the three bat genome assemblies. We then used BEDTools to intersect the coordinates of RVT_1 domain blastn hits with that of the candidate proviruses [102]. All the candidate proviruses intersecting with a MLERV1 RVT_1 hit were 'manually' inspected to refine their termini and confirm their identity as members of the MLERV1 family.

To comprehensively retrieve all proviruses and solo LTRs related to the FcERV_γ6/MLERV1/MPERV1 families in each of their respective genomes, we run RepeatMasker [103] with default setting and a custom repeat library with representitives from all MLERV1/MPERV1/FcERV_γ6 subfamilies against each genome assembly. The RepeatMasker output was then parsed using script parse_RMout_count_solo_and_full.pl to produce bed files of all complete solo LTRs and full length ERVs. We define a complete solo LTR as a sequence matching the LTR with missing less than 150 bp at their 5' termini and missing less than 10 bp at their 3' termini. We identified elements as putative proviruses those delimited by two LTRs in the same orientation separated by 3 kb to 10 kb of intervening sequence. Manual inspection of a subset of putative proviruses identified by this approach confirmed that most contained typical ERV coding sequences, though frequently interrupted by large sequence/assembly gaps. The LTR libraries and PERL scripts used for these analyses have been deposited on Github (https://github.com/xzhuo/orthologusLTR.git).

In the pangolin genome, our initial MEGAblast search yielded only 2 significant hits to the MLERV1 RVT_1 domain, but many more related ERVs could be retrieved using the RT domain from these two initial hits in reiterative blast searches against the pangolin genome assembly. To examine the relationship of these RT elements to each other and to the MLERV1 and FcERV_γ6 families, we conducted a phylogenetic analysis using the Maximum Likelihood package PhyML3.1 with the GTR+Γ model [104]. The resulting tree (S3 Fig) revealed that only the two initial hits clustered with FcERV_γ6/MLERV1 and were considered part of the MPERV1 family. The other elements form a distinct family we called MPERV2. MPERV1 and MPERV2 elements share less than 80% nucleotide sequence similarity in their coding regions, but still retain substantial level of sequence similarity in their LTRs. Thus, to correctly estimate the number of solo LTRs for the MPERV1 family, we had to examine their position on a phylogenetic tree of LTRs (S5 Fig). Using this approach, 27 solo LTRs could be assigned to the MPERV1 family in the pangolin genome assembly (as reported in Table 1). Because MPERV2

was much more distantly related to FcERV_γ6 and MLERV1 (<80% sequence similarity), we did not analyze further MPERV2 in this study. Reference sequences for MPERV1 and MPERV2 have been deposited in Repbase [48].

All identified ERVs are available as bed format (S1 File).

Sliding window pairwise similarity calculation

To generate the sliding window analysis shown in Fig 1A, we used MUSCLE [105] to align the nucleotide sequence for three pairs of proviruses: MLERV1_77 vs. FcERV_γ6_62; FcERV_γ6_62 v.s MPERV1_ltr106; SIVcpz(AF115393) vs. HIV-1(NC_001802) and used SEAVIEW [106] to manually adjust each alignment. Each pairwise alignment was then split into 300 bp windows with step size 50bp (i.e. = 170 segments for the MLERV1_77 vs. FcERV_γ6_62 alignment) and the percentage of sequence similarity was computed and corrected using kimura 2 parameter model for each window [107].

ERV orthologous loci identification

Orthologous ERV loci were detected similarly as we described previously [42]. Briefly, we used the Perl script extract_flanking_fasta.pl to extract 200 bp at both ends of each query element along with 200 bp of flanking sequences. The output file is then used as query in a batch blastn search against the target genome assembly with default parameters. The csv format blast output is then parsed using orthoblast_finder.pl to pair 5' end hits with 3'end hits. Finally, the paired hits output was parsed using the script final_annotation.pl to infer the presence/absence of each element in the target genome. All these perl scripts were deposited on Github (https://github.com/xzhuo/orthologusLTR.git).

Estimation of individual provirus insertions using LTR–LTR divergence

Sequence divergence between 5' and 3' LTRs from the same provirus was computed as previously described [42]. To infer insertion dates from LTR divergence of MLERV1 and FcERV_γ6 elements, we used previously estimated lineage-specific neutral substitution rates of 2.7×10^{-9} yr^{-1} [69] and 1.8×10^{-9} yr^{-1} [44] for the vespertilionid and felid lineages respectively. Since no substitution rate has yet been estimated for the pangolin lineage, we used the 'average' mammal neutral substitution rate of 2.2×10^{-9} yr^{-1} [70] to infer the age of MPERV1 insertions.

Phylogenetic analysis

The maximum-likelihood phylogenies presented for LTR sequences were built using PhyML3.1 and the support for each node is determined with an approximate likelihood ratio test [108]. The multiple alignment of LTR sequences was constructed using MUSCLE and PRANK with default nucleotide parameters and manually adjusted using SEAVIEW [105,106,109]. Nucleotide substitution model was chosen using AIC criterion in jmodeltest2.1.6 [110] (GTR + Γ). Dendroscope 3 was used for tree visualization [111].

Selection and integrity analysis on coding domains

ERV coding regions were predicted using HMMER3 in all 6 reading frames [112] to delineate Gag_MA (matrix), Gag_p30 (capsid), RVP (protease), RVT_1 (reverse transcriptase), RnaseH, rve (integrase) and TLV_coat (envelope transmembrane domain) domains in MLERV1, FcERV_γ6 and MPERV1 proviruses. A multiple codon alignment was generated for each set of coding domains using MUSCLE and manually adjusted with SEAVIEW [106]. The program codeml from the PAML4.8 package [77] was used to estimate dN/dS ratio with branch

model = 2. A maximum likelihood phylogeny of the LTR sequences was used as the guide tree in codeml. To test for purifying selection on each coding domain, we calculated the control lnL value by running codeml with ω fixed to 1. Then likelihood ratio test was performed as suggested by PAML to test if ω is significantly different from 1 [78].

Coding region integrity was assessed by calculating the frequency of stop codon or frameshift indels per codon. We calculated the total length of each domain in every subfamily, and counted the occurrence of stop codon and frameshift indels. The mean frequency and 95% confidence interval is calculated using the Poisson distribution [113].

Supporting Information

S1 Table. Excel table containing estimated age, ortholog and presence/absence of envelope of each provirus.
(XLSX)

S1 Fig. Unrooted ML tree of all related cat and tiger LTRs.
(PDF)

S2 Fig. The unrooted ML tree of all related proviruses in three lineages built with both 5' and 3' LTRs.
(PDF)

S3 Fig. Rooted RT domain ML tree of all related proviruses in three lineages with FcERV_γ6–46 used as outgroup.
(PDF)

S4 Fig. Integrity analysis of coding domains. Y axis represents frequency of stop codon or frameshift indel per codon in each domain, four different subfamilies are illustrated with different colors. Error bars represent 95% confidence interval.
(PDF)

S5 Fig. Pangolin soloLTR ML tree. All members of MPERV1 are enclosed within an arch, and the aLRT support of MPERV1 is labeled at the branch node.
(PDF)

S6 Fig. The ML tree with all *E. fuscus* soloLTRs and selected LTRs from other bats. Three subfamilies of MLERV1 are illustrated. SoloLTRs in *E. fuscus* cluster with either MLERV1_1 or MLERV1_2.
(PDF)

S1 File. Compressed bed files of all identified ERVs.
(ZIP)

Acknowledgments

We thank Edward Chuong, Rachel L. Cosby, Aurelie Kapusta, Ray Malfavon-Borja, Claudia Marquez, John McCormick and Ellen Pritham for helpful discussion.

Author Contributions

Conceived and designed the experiments: XZ CF. Performed the experiments: XZ. Analyzed the data: XZ CF. Contributed reagents/materials/analysis tools: XZ. Wrote the paper: XZ CF.

References

1. Parrish CR, Holmes EC, Morens DM, Park E-C, Burke DS, Calisher CH, et al. Cross-species virus transmission and the emergence of new epidemic diseases. Microbiol Mol Biol Rev. 2008; 72: 457–470. doi: 10.1128/MMBR.00004-08 PMID: 18772285
2. Wolfe ND, Dunavan CP, Diamond J. Origins of major human infectious diseases. Nature. 2007; 447: 279–283. doi: 10.1038/nature05775 PMID: 17507975
3. Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF, et al. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. Nature. 1999; 397: 436–441. doi: 10.1038/17130 PMID: 9989410
4. Lemey P, Pybus OG, Wang B, Saksena NK, Salemi M, Vandamme A-M. Tracing the origin and history of the HIV-2 epidemic. Proc Natl Acad Sci USA. 2003; 100: 6588–6592. doi: 10.1073/pnas.0936469100 PMID: 12743376
5. Coffin JM, Hughes SH, Varmus HE. Retroviruses. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 1997.
6. Troyer JL, VandeWoude S, Pecon-Slattery J, McIntosh C, Franklin S, Antunes A, et al. FIV cross-species transmission: an evolutionary prospective. Vet Immunol Immunopathol. 2008; 123: 159–166. doi: 10.1016/j.vetimm.2008.01.023 PMID: 18299153
7. Locatelli S, Peeters M. Cross-species transmission of simian retroviruses: how and why they could lead to the emergence of new diseases in the human population. AIDS. 2012; 26: 659–673. doi: 10.1097/QAD.0b013e328350fb68 PMID: 22441170
8. Minardi da Cruz JC, Singh DK, Lamara A, Chebloune Y. Small ruminant lentiviruses (SRLVs) break the species barrier to acquire new host range. Viruses. 2013; 5: 1867–1884. doi: 10.3390/v5071867 PMID: 23881276
9. Denner J. Transspecies transmissions of retroviruses: new cases. Virology. 2007; 369: 229–233. doi: 10.1016/j.virol.2007.07.026 PMID: 17870141
10. Niewiadomska AM, Gifford RJ. The extraordinary evolutionary history of the reticuloendotheliosis viruses. Plos Biol. 2013; 11: e1001642. doi: 10.1371/journal.pbio.1001642 PMID: 24013706
11. Gifford R, Tristem M. The evolution, distribution and diversity of endogenous retroviruses. Virus Genes. 2003; 26: 291–315. PMID: 12876457
12. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001; 409: 860–921. doi: 10.1038/35057062 PMID: 11237011
13. Mayer J, Blomberg J, Seal RL. A revised nomenclature for transcribed human endogenous retroviral loci. Mobile DNA. 2011; 2: 7. doi: 10.1186/1759-8753-2-7 PMID: 21542922
14. Magiorkinis G, Blanco-Melo D, Belshaw R. The decline of human endogenous retroviruses: extinction and survival. Retrovirology. 2015; 12: 8. doi: 10.1186/s12977-015-0136-x PMID: 25640971
15. Katzourakis A, Rambaut A, Pybus OG. The evolutionary dynamics of endogenous retroviruses. Trends in Microbiology. 2005; 13: 463–468. doi: 10.1016/j.tim.2005.08.004 PMID: 16109487
16. Feschotte C, Gilbert C. Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet. 2012; 13: 283–296. doi: 10.1038/nrg3199 PMID: 22421730
17. Gifford RJ. Viral evolution in deep time: lentiviruses and mammals. Trends Genet. 2012; 28: 89–100. doi: 10.1016/j.tig.2011.11.003 PMID: 22197521
18. Henzy JE, Gifford RJ, Johnson WE, Coffin JM. A novel recombinant retrovirus in the genomes of modern birds combines features of avian and mammalian retroviruses. Journal of Virology. 2014; 88: 2398–2405. doi: 10.1128/JVI.02863-13 PMID: 24352464
19. Hayward A, Grabherr M, Jern P. Broad-scale phylogenomics provides insights into retrovirus-host evolution. Proc Natl Acad Sci USA. 2013. doi: 10.1073/pnas.1315419110
20. Hayward A, Cornwallis CK, Jern P. Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. Proceedings of the National Academy of Sciences. 2015; 112: 464–469. doi: 10.1073/pnas.1414980112
21. Weiss RA. The discovery of endogenous retroviruses. Retrovirology. 2006; 3: 67. doi: 10.1186/1742-4690-3-67 PMID: 17018135
22. Mang R, Maas J, van der Kuyl AC, Goudsmit J. Papio cynocephalus endogenous retrovirus among old world monkeys: evidence for coevolution and ancient cross-species transmissions. Journal of Virology. 2000; 74: 1578–1586. PMID: 10627573
23. van der Kuyl AC, Dekker JT, Goudsmit J. Distribution of baboon endogenous virus among species of African monkeys suggests multiple ancient cross-species transmissions in shared habitats. Journal of Virology. 1995; 69: 7877–7887. PMID: 7494300

24. Yohn CT, Jiang Z, McGrath SD, Hayden KE, Khaitovich P, Johnson ME, et al. Lineage-specific expansions of retroviral insertions within the genomes of African great apes but not humans and orangutans. Plos Biol. 2005; 3: e110. doi: 10.1371/journal.pbio.0030110 PMID: 15737067

25. Polavarapu N, Bowen NJ, McDonald JF. Identification, characterization and comparative genomics of chimpanzee endogenous retroviruses. Genome Biol. 2006; 7: R51. doi: 10.1186/gb-2006-7-6-r51 PMID: 16805923

26. Gilbert C, Maxfield DG, Goodman SM, Feschotte C. Parallel germline infiltration of a lentivirus in two Malagasy lemurs. PLoS Genet. 2009; 5: e1000425. doi: 10.1371/journal.pgen.1000425 PMID: 19300488

27. Wang Y, Liška F, Gosele C, Šedová L, Křen V, Křenová D, et al. A novel active endogenous retrovirus family contributes to genome variability in rat inbred strains. Genome Research. 2010; 20: 19–27. doi: 10.1101/gr.100073.109 PMID: 19887576

28. Escalera-Zamudio M, Mendoza MLZ, Heeger F, Loza-Rubio E, Rojas-Anaya E, Méndez-Ojeda ML, et al. A novel endogenous betaretrovirus in the common vampire bat (Desmodus rotundus) suggests multiple independent infection and cross-species transmission events. Journal of Virology. 2015; 89: 5180–5184. doi: 10.1128/JVI.03452-14 PMID: 25717107

29. MARTIN J, Herniou E, Cook J, O'Neill RW, Tristem M. Interclass transmission and phyletic host tracking in murine leukemia virus-related retroviruses. Journal of Virology. 1999; 73: 2442–2449. PMID: 9971829

30. van der Kuyl AC, Dekker JT, Goudsmit J. Discovery of a new endogenous type C retrovirus (FcEV) in cats: evidence for RD-114 being an FcEV(Gag-Pol)/baboon endogenous virus BaEV(Env) recombinant. Journal of Virology. 1999; 73: 7994–8002. PMID: 10482547

31. Henzy JE, Johnson WE. Pushing the endogenous envelope. Philosophical Transactions of the Royal Society B: Biological Sciences. 2013; 368: 20120506. doi: 10.1098/rstb.2012.0506

32. Tarlinton R, Meers J, Young P. Biology and evolution of the endogenous koala retrovirus. Cell Mol Life Sci. 2008; 65: 3413–3421. doi: 10.1007/s00018-008-8499-y PMID: 18818870

33. Magiorkinis G, Gifford RJ, Katzourakis A, De Ranter J, Belshaw R. Env-less endogenous retroviruses are genomic superspreaders. Proc Natl Acad Sci USA. National Acad Sciences; 2012; 109: 7385–7390. doi: 10.1073/pnas.1200913109 PMID: 22529376

34. Mata H, Gongora J, Eizirik E, Alves BM, Soares MA, Ravazzolo AP. Identification and characterization of diverse groups of endogenous retroviruses in felids. Retrovirology. 2015; 12: 26. doi: 10.1186/s12977-015-0152-x PMID: 25808580

35. Calisher CH, Childs JE, Field HE, Holmes KV, Schountz T. Bats: Important Reservoir Hosts of Emerging Viruses. Clinical Microbiology Reviews. 2006; 19: 531–545. doi: 10.1128/CMR.00017-06 PMID: 16847084

36. Wong S, Lau S, Woo P, Yuen K-Y. Bats as a continuing source of emerging infections in humans. Rev Med Virol. 2007; 17: 67–91. doi: 10.1002/rmv.520 PMID: 17042030

37. Omatsu T, Watanabe S, Akashi H, Yoshikawa Y. Biological characters of bats in relation to natural reservoir of emerging viruses. Comparative Immunology, Microbiology and Infectious Diseases. 2007; 30: 357–374. doi: 10.1016/j.cimid.2007.05.006 PMID: 17706776

38. Luis AD, Hayman DTS, O'Shea TJ, Cryan PM, Gilbert AT, Pulliam JRC, et al. A comparison of bats and rodents as reservoirs of zoonotic viruses: are bats special? Proceedings of the Royal Society B: Biological Sciences. 2013; 280: 20122753. doi: 10.1098/rspb.2012.2753 PMID: 23378666

39. Brook CE, Dobson AP. Bats as "special" reservoirs for emerging zoonotic pathogens. Trends in Microbiology. 2015; 23: 172–180. doi: 10.1016/j.tim.2014.12.004 PMID: 25572882

40. Moratelli R, Calisher CH. Bats and zoonotic viruses: can we confidently link bats with emerging deadly viruses? Mem Inst Oswaldo Cruz. 2015. doi: 10.1590/0074-02760150048

41. Smith I, Wang L-F. Bats and their virome: an important source of emerging viruses capable of infecting humans. Curr Opin Virol. 2013; 3: 84–91. doi: 10.1016/j.coviro.2012.11.006 PMID: 23265969

42. Zhuo X, Rho M, Feschotte C. Genome-wide characterization of endogenous retroviruses in the bat Myotis lucifugus reveals recent and diverse infections. Journal of Virology. 2013; 87: 8493–8501. doi: 10.1128/JVI.00892-13 PMID: 23720713

43. Pontius JU, Mullikin JC, Smith DR, Agencourt Sequencing Team, Lindblad-Toh K, Gnerre S, et al. Initial sequence and comparative analysis of the cat genome. Genome Research. 2007; 17: 1675–1689. doi: 10.1101/gr.6380007 PMID: 17975172

44. Cho YS, Hu L, Hou H, Lee H, Xu J, Kwon S, et al. The tiger genome and comparative analysis with lion and snow leopard genomes. Nat Commun. 2013; 4: 2433. doi: 10.1038/ncomms3433 PMID: 24045858

45. Seim I, Fang X, Xiong Z, Lobanov AV, Huang Z, Ma S, et al. Genome analysis reveals insights into physiology and longevity of the Brandt's bat Myotis brandtii. Nat Commun. 2013; 4: 2212. doi: 10.1038/ncomms3212 PMID: 23962925

46. Zhang G, Cowled C, Shi Z, Huang Z, Bishop-Lilly KA, Fang X, et al. Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. Science. 2013; 339: 456–460. doi: 10.1126/science.1230835 PMID: 23258410

47. Yuhki N, Mullikin JC, Beck T, Stephens R, O'Brien SJ. Sequences, Annotation and Single Nucleotide Polymorphism of the Major Histocompatibility Complex in the Domestic Cat. Ellegren H, editor. PLoS ONE. 2008; 3: e2674. doi: 10.1371/journal.pone.0002674.t006 PMID: 18629345

48. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA. 2015; 6: 11. doi: 10.1186/s13100-015-0041-9 PMID: 26045719

49. Song N, Jo H, Choi M, Kim J-H, Seo HG, Cha S-Y, et al. Identification and classification of feline endogenous retroviruses in the cat genome using degenerate PCR and in silico data analysis. J Gen Virol. 2013. doi: 10.1099/vir.0.051862–0

50. Martoglio B, Graf R, Dobberstein B. Signal peptide fragments of preprolactin and HIV-1 p-gp160 interact with calmodulin. The EMBO Journal. 1997; 16: 6636–6645. doi: 10.1093/emboj/16.22.6636 PMID: 9362478

51. Corbet S, Muller-Trutwin MC, Versmisse P, Delarue S, Ayouba A, Lewis J, et al. env sequences of simian immunodeficiency viruses from chimpanzees in Cameroon are strongly related to those of human immunodeficiency virus group N from the same geographic area. Journal of Virology. 2000; 74: 529–534. doi: 10.1128/JVI.74.1.529–534.2000 PMID: 10590144

52. Huet T, Cheynier R, Meyerhans A, Roelants G, Wain-Hobson S. Genetic organization of a chimpanzee lentivirus related to HIV-1. Nature. 1990; 345: 356–359. doi: 10.1038/345356a0 PMID: 2188136

53. Pancino G, Ellerbrok H, Sitbon M, Sonigo P. Conserved framework of envelope glycoproteins among lentiviruses. Curr Top Microbiol Immunol. 1994; 188: 77–105. PMID: 7924431

54. Meredith RW, Janecka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, et al. Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification. Science. 2011; 334: 521–524. doi: 10.1126/science.1211028 PMID: 21940861

55. Hedges SB, Dudley J, Kumar S. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics. 2006; 22: 2971–2972. doi: 10.1093/bioinformatics/btl505 PMID: 17021158

56. Shedlock AM, Okada N. SINE insertions: powerful tools for molecular systematics. Bioessays. 2000; 22: 148–160. doi: 10.1002/(SICI)1521-1878(200002)22:2<148::AID-BIES6>3.0.CO;2-Z PMID: 10655034

57. Ray DA, Xing J, Salem A-H, Batzer MA. SINEs of a nearly perfect character. Syst Biol. 2006; 55: 928–935. doi: 10.1080/10635150600865419 PMID: 17345674

58. Maeda N, Kim HS. Three independent insertions of retrovirus-like sequences in the haptoglobin gene cluster of primates. Genomics. 1990; 8: 671–683. PMID: 2177446

59. Johnson WE, Coffin JM. Constructing primate phylogenies from ancient retrovirus sequences. Proc Natl Acad Sci USA. 1999; 96: 10254–10260. PMID: 10468595

60. Bashir A, Ye C, Price AL, Bafna V. Orthologous repeats and mammalian phylogenetic inference. Genome Research. 2005; 15: 998–1006. doi: 10.1101/gr.3493405 PMID: 15998912

61. Johnson WE. The Late Miocene Radiation of Modern Felidae: A Genetic Assessment. Science. 2006; 311: 73–77. doi: 10.1126/science.1122277 PMID: 16400146

62. Stadelmann B, Lin LK, Kunz TH, Ruedi M. Molecular phylogeny of New World Myotis (Chiroptera, Vespertilionidae) inferred from mitochondrial and nuclear DNA genes. Molecular Phylogenetics and Evolution. 2007; 43: 32–48. doi: 10.1016/j.ympev.2006.06.019 PMID: 17049280

63. Miller-Butterworth CM, Murphy WJ, O'Brien SJ, Jacobs DS, Springer MS, Teeling EC. A family matter: conclusive resolution of the taxonomic position of the long-fingered bats, miniopterus. Molecular Biology and Evolution. 2007; 24: 1553–1561. doi: 10.1093/molbev/msm076 PMID: 17449895

64. Lack JB, Roehrs ZP, Stanley CE Jr, Ruedi M, Van Den Bussche RA. Molecular phylogenetics of Myotis indicate familial-level divergence for the genus Cistugo (Chiroptera). Journal of Mammalogy. 2010; 91: 976–992. doi: 10.1644/09-MAMM-A-192.1

65. Agnarsson I, Zambrana-Torrelio CM, Flores-Saldana NP, May-Collado LJ. A time-calibrated species-level phylogeny of bats (Chiroptera, Mammalia). PLoS Curr. 2011; 3: RRN1212. doi: 10.1371/currents.RRN1212 PMID: 21327164

66. Dangel AW, Baker BJ, Mendoza AR, Yu CY. Complement component C4 gene intron 9 as a phylogenetic marker for primates: long terminal repeats of the endogenous retrovirus ERV-K(C4) are a molecular clock of evolution. Immunogenetics. 1995; 42: 41–52. PMID: 7797267

67. Vitte C, Panaud O, Quesneville H. LTR retrotransposons in rice (Oryza sativa, L.): recent burst amplifications followed by rapid DNA loss. BMC Genomics. BioMed Central Ltd; 2007; 8: 218. doi: 10.1186/1471-2164-8-218 PMID: 17617907

68. Yoder JA, Walsh CP, Bestor TH. Cytosine methylation and the ecology of intragenomic parasites. Trends Genet. 1997; 13: 335–340. doi: 10.1016/S0168-9525(97)01181-5 PMID: 9260521

69. Pace JK, Gilbert C, Clark MS, Feschotte C. Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. Proc Natl Acad Sci USA. 2008; 105: 17023–17028. doi: 10.1073/pnas.0806548105 PMID: 18936483

70. Kumar S, Subramanian S. Mutation rates in mammalian genomes. Proc Natl Acad Sci USA. 2002; 99: 803–808. doi: 10.1073/pnas.022629899 PMID: 11792858

71. Hughes JF. Human Endogenous Retroviral Elements as Indicators of Ectopic Recombination Events in the Primate Genome. Genetics. 2005; 171: 1183–1194. doi: 10.1534/genetics.105.043976 PMID: 16157677

72. Kijima TE, Innan H. On the Estimation of the Insertion Time of LTR Retrotransposable Elements. Molecular Biology and Evolution. 2010; 27: 896–904. doi: 10.1093/molbev/msp295 PMID: 19955475

73. Slattery JP, Franchini G, Gessain A. Genomic evolution, patterns of global dissemination, and interspecies transmission of human and simian T-cell leukemia/lymphotropic viruses. Genome Research. 1999; 9: 525–540. PMID: 10400920

74. Fernández-Medina RD, Ribeiro JMC, Carareto CMA, Velasque L, Struchiner CJ. Losing identity: structural diversity of transposable elements belonging to different classes in the genome of Anopheles gambiae. BMC Genomics. 2012; 13: 272. doi: 10.1186/1471-2164-13-272 PMID: 22726298

75. Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A, et al. Long-term reinfection of the human genome by endogenous retroviruses. Proc Natl Acad Sci USA. 2004; 101: 4894–4899. doi: 10.1073/pnas.0307800101 PMID: 15044706

76. Belshaw R, Katzourakis A, Paces J, Burt A, Tristem M. High copy number in human endogenous retrovirus families is associated with copying mechanisms in addition to reinfection. Molecular Biology and Evolution. 2005; 22: 814–817. doi: 10.1093/molbev/msi088 PMID: 15659556

77. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution. 2007; 24: 1586–1591. doi: 10.1093/molbev/msm088 PMID: 17483113

78. Yang Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Molecular Biology and Evolution. 1998; 15: 568–573. PMID: 9580986

79. Ribet D, Harper F, Dupressoir A, Dewannieux M, Pierron G, Heidmann T. An infectious progenitor for the murine IAP retrotransposon: Emergence of an intracellular genetic parasite from an ancient retrovirus. Genome Research. 2008; 18: 597–609. doi: 10.1101/gr.073486.107 PMID: 18256233

80. Meyerson NR, Sawyer SL. Two-stepping through time: mammals and viruses. Trends in Microbiology. 2011; 19: 286–294. doi: 10.1016/j.tim.2011.03.006 PMID: 21531564

81. Daugherty MD, Malik HS. Rules of engagement: molecular insights from host-virus arms races. Annu Rev Genet. 2012; 46: 677–700. doi: 10.1146/annurev-genet-110711-155522 PMID: 23145935

82. Demogines A, Abraham J, Choe H, Farzan M, Sawyer SL. Dual host-virus arms races shape an essential housekeeping protein. Plos Biol. 2013; 11: e1001571. doi: 10.1371/journal.pbio.1001571 PMID: 23723737

83. Gaudin TJ, Emry RJ, Wible JR. The phylogeny of living and extinct pangolins (Mammalia, Pholidota) and associated taxa: a morphology based analysis. Journal of mammalian evolution. 2009. doi: 10.1007/s10914-009-9119-9

84. Ancillotto L, Serangeli MT, Russo D. Curiosity killed the bat: domestic cats as bat predators. Mammalian Biology-Zeitschrift für . . .. 2013. doi: 10.1016/j.mambio.2013.01.003

85. Phillips S, Coburn D, James R. An observation of cat predation upon an eastern blossom bat Syconycteris australis. Australian Mammalogy. 2001.

86. Scrimgeour J, Beath A, Swanney M. Cat predation of short-tailed bats (Mystacina tuberculata rhyocobia) in Rangataua Forest, Mount Ruapehu, Central North Island, New Zealand. New Zealand Journal of Zoology. 2012; 39: 257–260. doi: 10.1080/03014223.2011.649770

87. Woods M, McDonald RA, Harris S. Predation of wildlife by domestic cats Felis catus in Great Britain. Mammal review. 2003.

88. Grzimek B, Kleiman DG, Schlager N, Geist V, Olendorf D, McDade MC, et al. Grzimek's Animal Life Encyclopedia: Mammals I-V. Gale / Cengage Learning; 2003.

89. Dacheux L, Larrous F, Mailles A. European bat lyssavirus transmission among cats, Europe. Emerging infectious . . .. 2009. doi: 10.3201/eid1502.080637

90. Dewannieux M, Dupressoir A, Harper F, Pierron G, Heidmann T. Identification of autonomous IAP LTR retrotransposons mobile in mammalian cells. Nat Genet. 2004; 36: 534–539. doi: 10.1038/ng1353 PMID: 15107856

91. Mietz JA, Grossman Z, Lueders KK, Kuff EL. Nucleotide sequence of a complete mouse intracisternal A-particle genome: relationship to known aspects of particle assembly and function. Journal of Virology. 1987; 61: 3020–3029. PMID: 3041022

92. David VA, Menotti-Raymond M, Wallace AC, Roelke M, Kehler J, Leighty R, et al. Endogenous Retrovirus Insertion in the KIT Oncogene Determines White and White spotting in Domestic Cats. G3 (Bethesda). 2014. doi: 10.1534/g3.114.013425

93. Ndung'u T, Weiss RA. On HIV diversity. AIDS. 2012; 26: 1255–1260. doi: 10.1097/QAD.0b013e32835461b5 PMID: 22706010

94. Moosman PR Jr, Thomas HH. Diet of the widespread insectivorous bats Eptesicus fuscus and Myotis lucifugus relative to climate and richness of bat communities. Journal of . . .. 2012. doi: 10.1644/11-MAMM-A-274.1

95. Hutson AM, Mickleburgh SP, Racey PA. Microchiropteran Bats. IUCN; 2001.

96. Katzourakis A, Magiorkinis G, Lim AG, Gupta S, Belshaw R, Gifford R. Larger Mammalian Body Size Leads to Lower Retroviral Activity. PLoS Pathog. Public Library of Science; 2014; 10: e1004214. doi: 10.1371/journal.ppat.1004214 PMID: 25033295

97. Hayman DTS, Bowen RA, Cryan PM, McCracken GF, O'Shea TJ, Peel AJ, et al. Ecology of zoonotic infectious diseases in bats: current knowledge and future directions. Zoonoses Public Health. 2013; 60: 2–21. doi: 10.1111/zph.12000 PMID: 22958281

98. O'Shea TJ, Cryan PM, Cunningham AA, Fooks AR, Hayman DTS, Luis AD, et al. Bat flight and zoonotic viruses. Emerging Infect Dis. 2014; 20: 741–745. doi: 10.3201/eid2005.130539 PMID: 24750692

99. Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA. Database indexing for production MegaBLAST searches. Bioinformatics. 2008; 24: 1757–1764. doi: 10.1093/bioinformatics/btn322 PMID: 18567917

100. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics. 2008; 9: 18. doi: 10.1186/1471-2105-9-18 PMID: 18194517

101. Steinbiss S, Willhoeft U, Gremme G, Kurtz S. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. Nucleic Acids Research. 2009; 37: 7002–7013. doi: 10.1093/nar/gkp759 PMID: 19786494

102. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26: 841–842. doi: 10.1093/bioinformatics/btq033 PMID: 20110278

103. Smit A, Hubley R, Green P. 2013–2015. RepeatMasker Open-4.0. [Internet].

104. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010; 59: 307–321. doi: 10.1093/sysbio/syq010 PMID: 20525638

105. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004; 32: 1792–1797. doi: 10.1093/nar/gkh340 PMID: 15034147

106. Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular Biology and Evolution. 2010; 27: 221–224. doi: 10.1093/molbev/msp259 PMID: 19854763

107. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980; 16: 111–120. PMID: 7463489

108. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 2003; 52: 696–704. PMID: 14530136

109. Löytynoja A, Goldman N. An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci USA. 2005; 102: 10557–10562. doi: 10.1073/pnas.0409137102 PMID: 16000407

110. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 2012; 9: 772. doi: 10.1038/nmeth.2109 PMID: 22847109

111. Huson DH, Scornavacca C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol. 2012; 61: 1061–1067. doi: 10.1093/sysbio/sys062 PMID: 22780991

112. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011; 7: e1002195. doi: 10.1371/journal.pcbi.1002195 PMID: 22039361

113. Garwood F. Fiducial limits for the Poisson distribution. Biometrika. 1936. doi: 10.2307/2333958

CHAPTER 4

IDENTIFICATION OF INFECTIOUS OR HYPOMETHYLATED ERVS IN MAMMALIAN GENOME

4.1 Abstract

Transposable elements (TEs) comprise about half of a typical mammal genome. Among all TEs, endogenous retroviruses (ERVs) are unique in that they are derived from infectious retroviruses and their evolutionary history are featured by frequent cross-species transmission. More intriguingly, there are several recent studies suggesting that ERVs are more likely to be co-opted by host than other TEs. However, it is still unclear how their unique amplification mechanism and tendency of being adapted would affect ERV evolution. On the other hand, most TEs in mammalian genomes including ERVs are heavily methylated at CpG sites, which mutates at a much higher rate than other sites via spontaneous deamination. Here we established a new method and statistical framework to analyze ERV evolution dynamics in vertebrates by separating CpG site mutations from non-CpG site mutations. With this method we can easily screen interesting ERV families (actively moving or potentially adapted) in any reference genome without experimental data. By applying our method to human genome we distinguished hypomethylated (LTR12C, MER57E3) ERV families which can act as enhancer or promoter. we also separated HERV underwent reinfection (HERVK-HML2). We subsequently applied our method to mouse genome which hosts many more active moving ERVs, and currently hyperactive ERVs stand out in our analysis. Finally, we applied it to 40-plus vertebrate genomes and uncovered several potentially infectious new ERVs. We also found many more ERVs or LTR retrotransposons that might undergo reinfection or germline hypomethylation in several reptilian (alligator, sea turtle) genomes. Our finding illustrated the variability among different ERV families with in a genome, and variability among different vertebrates in their ERV dynamics.

4.2 Introduction

Transposable elements (TEs) constitute about half of a typical mammal genome. They are classified into class I retrotransposon and class II DNA transposon. With exception of vesper bats (family *vespertilionidae*),[1] the repetitive landscape of all known mammals are mostly constituted by retrotransposons.[2] Retrotransposon can be further divided into non-LTR retrotransposon and LTR retrotransposons.

None-LTR retrotransposons include autonomous long interspersed elements (LINEs) and nonautonomous short interspersed elements (SINEs) and make up about 80% of repetitive sequence in both human and mouse genomes. They evolve continuously in most of mammalians but rarely transmit cross species.

Endogenous retroviruses (ERVs) account for a sizeable fraction of vertebrate genomes, for instance, 5-10% of diverse mammalian genomes thus far analyzed.[3] In contrast to LINEs and SINEs, ERV evolution in mammals often features continuous cross-species transmission (CST) between distinct hosts.[4–7] ERV's distinct evolutionary history is probably stemmed from their infectious retroviral origin. Retroviruses are unique among animal viruses in that they are obligated to integrate their genome in the host cell chromosome as provirus for their replication.[8] When retroviruses infect the germline, the integrated proviruses become inheritable vertically, which opens the door for their spread and fixation in the host population as mendelian alleles.[9]

Cytosine methylation plays a critical role in gene regulation. It is well known that TEs are heavily methylated at CpG sites of their nucleotide sequence in vertebrates.[10] Because of the frequent spontaneous deamination of methylated cytosines,[11] the evolution of TE sequences, which generally follows a universal neutral rate after their integration in the genome,[12] is characterized by an elevated transition rate at CpG sites.[13] Therefore, if we estimate transition rate at CpG sites and at non-CpG sites among copies derived from a single ancestor (copies belongs to the same subfamily), we should find the former rate much higher.

Indeed, Xing et al. measured CpG sites and non-CpG sites substitution density of different Alu subfamilies in human genome and found their dynamics can be explained by non-linear CpG decay and a universal CpG/non-CpG mutation ratio.[14]

However, spontaneous deamination only occurs to hypermethylated region in host genome. As to hypomethylated region, the difference between CpG and non-CpG transition rates would be much smaller. In addition, it has also been found that hypomethylation is also associated with regulatory elements in vertebrates (enhancers and promoters) (more refs here).[15] Therefore, if a significant fraction of a particular TE subfamily are co-opted as

regulatory elements, the substitution rate difference between CpG and non-CpG of this subfamily could be smaller. On the other hand, during TE replication process, retrotransposons replicated their copies through reverse transcription, which is known to be error prone and should not be affected by cytosine methylation. We suspect that despite the human *Alu* elements could be described by the smooth nonlinear decay model,[14] other TEs especially some ERV families that have complex evolutionary history or were adapted as regulatory network may have a different mutation spectra. Now with many more genomes sequenced and TE annotated, we decided to conduct a comprehensive analysis to explore mutation spectra of all TE subfamilies in different hosts.

Here we characterized substitution density at CpG sites and non-CpG sites of all TE subfamilies in 44 vertebrate genomes. In agreement with Xing et al.,[14] we found that the relationship between CpG substitution and non-CpG substitution of SINEs and LINEs of different ages can be largely explained by the decay of CpG sites in most of genomes. However, among different ERV subfamilies we found substantial variation that cannot be explained by the age associated CpG decay alone. Our finding illustrated that unlike non-LTR retrotransposons, which are generally hypermethylated and evolve uniformly, different ERV subfamilies evolve differently: they replicate through different paths and are differentially methylated. We also identified several distinct ERV subfamilies that are hypomethylated or accumulated excessive mutations during active mobilization.

4.3 Result

4.3.1 HERV is distinct from LINE and SINE in their mutation pattern

We begin our analysis with human genome. We constructed a multiple sequences alignment for each TE subfamily annotated in the human genome, and estimated their consensus sequences applying majority rule and used it as the ancestral sequence of family (details in method section). To compare the mutation rate difference between CpG sites and non-CpG sites, we only counted C-to-T transition for both CpG and non-CpG sites for each consensus-sequence pair. Since it has been documented that retrotransposons are the primary target of APOBEC deaminases which specifically induce C-to-U transitions in antisense single-stranded DNA and cause an excess of G-to-A transition in the sense strand of retrotransposon sequences[16 17],[18] we intentionally excluded G-to-A transition in our calculation to avoid potential APOBEC-driven mutations.

The copy number of TE families varies greatly within human genome. There are some *Alu* subfamilies with millions of copies scattered in our genome, and there are other

subfamilies with only a few copies. To detect any systematic bias caused by copy number difference in our method, we repeatedly sampled 10, 30, 100, 300, and 1000 full length copies from AluSz6 subfamily 100 times to simulate TE families with different sizes and calculated C to T transition rate at both CpG and non-CpG sites with the method we described above (Figure 4.1). As expected, we found that the variation of low copy number subfamilies are larger than of high copy number subfamilies in the simulation. However, our simulation also illustrated that the C-T transition rate at non-CpG sites are inflated (shifted right in the plot) when the family size is low (Figure 4.1a). We suspect that our ancestor sequence estimation was impaired in small subfamilies by elevated mutation rate at CpG sites (CpG sites in ancestor sequence could be wrongly predicted as CpA by the majority rule consensus). Therefore, we excluded predicted CpA sites from estimated consensus sequence and repeated the sample simulation. Indeed, the distribution of C-T transition at both CpG and non-CpG sites of different sized subfamilies centered together in our plot with estimated ancestral CpA sites excluded (Figure 4.1b). Our simulation with different sized TE subfamilies indicates that by excluding predicted CpA sites in consensus we can unbiasedly compare C-T transition of different TE subfamilies. All the subsequence analysis excluded CpA sites in consensus, and we restricted our analysis to TE subfamilies with at least 30 full-length copies to minimize fluctuation caused by small sample size.

We next generated a scatterplot of C-to-T transition rates at CpG and non-CpG sites for each TE subfamily in the human genome (Figure 4.2). As we expected, the distribution of *Alu* elements in the plot indicates an elevated transition rate at CpG sites and it recapitulated the results previously described by Xing et al.[14] The distribution of LINE-1 (L1) elements and DNA transposons were very similar to that of *Alu* elements. By contrast, the distribution of Human Endogenous Retrovirus (HERV) subfamilies show a much more dispersed distribution. In particular, there are multiple HERV subfamilies located at the lower right side of the Alu distribution characterized by low transition rates at CpG sites and high transition rate at non-CpG sites, which are described as low CpG/non-CpG ratio by us (Figure 4.2). To statistically distinguish HERV subfamilies with low CpG/non-CpG ratio from other TEs, we modified the function previously defined by Xing et al. for Alu elements[14] to describe the relationship between CpG transition and non-CpG transition of all non-ERV TE families, and estimated the 95% predication interval of the function using the delta method.[19] We also estimated the exact 95% confidence interval of HERV mutation rate assuming mutation as a Poisson process (corrected with FDR).[20,21] Using the statistical threshold defined above, we identified 13 HERV subfamilies with significantly

lower CpG/nonCpG ratio. Using the same statistics, there were no HERV subfamily with significantly higher CpG/nonCpG ratio.

4.3.2 Germline hypomethylation of 5 out of the 13 HERV subfamilies with low CpG/non-CpG ratio

All the TE substitutions we observed there among TE copies in human genome arose either before integration during TE amplification, or after integration along with our genome. We found 13 HERV subfamilies with low CpG/non-CpG ratio, and we want to further investigate whether the low CpG/non-CpG ratio is contributed by mutations before integration or mutations after integration. Sequence difference between paralogs within a genome includes both before and after integration mutations, but the difference between orthologus copies from different species can only arose after TE integration. Therefore, we extracted orthologus TE pairs of human and chimpanzee of the 13 HERV subfamilies with low CpG/non-CpG ratio along with an *Alu* subfamily of comparable age (AluSc8), and calculated after integration CpG/non-CpG ratio (see methods for ratio calculation) between orthologs (Figure 4.3). We found the after integration CpG/non-CpG ratio of five (LTR12C, LTR12E, LTR6A, LTR6B, MER57E3) of the 13 HERV subfamilies are significantly lower than the AluSc8 subfamily, while the ratio of the other 8 of them are not lower than AluSc8. That includes LTR5_Hs, LTR7, and LTR7Y, or which CpG/non-CpG ratio is actually higher than AluSc8.

The high mutation rate at CpG sites is driven by cytosine hypermethylation. We next sought to examine whether the relatively low CpG/non-CpG ratio characterizing these HERV subfamilies identified above might be caused by an exceptional resistance of these elements to cytosine methylation in the human germline. To examine this, we analyzed the level of CpG methylation of all the 13 HERV subfamilies in human mature sperm and spermatogonial stem cells using a single base-pair resolution map of cytosine methylation previously generated by genome bisulfite sequencing.[22] As a comparison, we also analyzed the level of CpG methylation of an *Alu* subfamily of comparable age (AluSc8) in the same dataset. The results showed that 5 out of the 13 HERV subfamilies (LTR12C, LTR12E, LTR6A, LTR6B and MER57E3) display lower methylation level than the *Alu* subfamily or the other ERV subfamilies (Figure 4.4). The methylation level of these 5 HERV subfamilies is consistent with the observation that their human-chimp CpG/non-CpG ratio is lower than hypermethylated TEs, since methylation driven transition only occurs to them after their integration.

4.3.3 MER57E3 may act as zinc-finger promoter in primates

We found MER57E3, a primate specific ERV subfamily, shows strikingly low methylation level during human spermatogenesis (Figure 4.4). To better understand why this subfamily is so drastically hypomethylated, we further investigated the location of MER57E3 elements in human genome. Surprisingly, out of 273 MER57E3 copies we found in hg38 reference genome, 143 of them locate in the proximal promoter regions (within 1 kb from the transcription start site) of refseq genes (Figure 4.5). 124 of them are close to promoter of zinc-finger (ZNF) genes. The association between MER57E3 with ZNF genes is consistent with the observation that majority of MER47E3 copies are found in chromosome 19, which hosts huge clusters of ZNF genes in our genome (Figure 4.5).[23] Consistent with previous finding,[24–26] The hypomethylation and association with ZNF gene promoter suggests MER57E3 may be wired as cis-regulatory element in ZNF regulatory network. Intriguingly, it has been suggested that these ZNF genes in human chromosome 19 act as repressors of ERVs.[27] If the MER57E3 adaptation hypothesis is true, the co-option of one ERV sequence to suppress other ERVs becomes another case of "fighting fire with fire" in the arms race between host and parasitic elements.[28]

4.3.4 Alternative promoter function of LTR12

Different LTR12 subfamilies are associated with HERV9. Recently there are multiple account reporting LTR12 especially LTR12C providing promoters for different protein coding genes and lncRNAs in normal tissue, tumor and cell lines.[29] It has been shown that LTR12 drives expression of ZNF80 and ADH1C,[30,31] it can also activate downstream -globin expression by long range interaction.[32] In K562 cells, LTR12C is the most enriched TE drives vlincRNA expression.[33] It also regulates expression by long range interaction in other ENCODE reference cell lines.[34] In hepatocellular carcinoma tumors, LTR12C is again highly activated and drives abnormal ncRNA expression.[35] LTR12 might also be adapted to protect genome integrity in human male germline by driving multiple transcripts expression (GTAp63 and TNFRSF10B) and inducing apoptosis in response to DNA damage.[36,37] There proapoptosis transcripts are silenced in testis cancer cells but can be restored by HDAC inhibitors.[36] It is promising that not only in testis cancer, their expression in other cancer cells can also be restored by those drugs.[38] And that suggests LTR12 is promising target candidate of epigenetic cancer therapy. Without using any experimental data, we identified LTR12 as an unusual TE in human genome here, and confirmed that it is relatively hypomethylated during spermatogenesis by examining published WGBS data.

Thus, we conclude that our method has the power to identify atypical elements that have experienced exceptionally low level of CpG methylation in the germline, and efficiently screen for candidate TE that might be adapted for regulatory function.

4.3.5 HERV families replicated through reinfection

There are eight other HERV subfamilies of which the low CpG/non-CpG ratio cannot be explained by CpG hypomethylation: their CpG/non-CpG ratio is high if only human-chimpanzee orthologus pairs are used for calculation, and they are hypermethylated during spermatogenesis. It suggests that their low CpG/non-CpG ratio is not merely the result of being hypomethylated in the human germline. These results indicate that these elements have been subject to CpG methylation since their integration in the genome and that the low CpG/non-CpG mutation ratio characterizing the intra-subfamily divergence of these ERVs is not caused by a deficit of transitions at CpG sites, Rather, it is best explained by the relatively large number of mutations accumulated during replication prior to chromosomal integration, most likely reflecting a process of reinfection.

Consistent with the idea, seven of the eight outlier HERV families have been previously described as HERV families having proliferated primarily via reinfection based on the signature of purifying selection that acted on their envelope domain.[39,40] These include the HERVK(HML2) subfamily (and its associated LTR5_Hs), which has long been suspected to have spread primarily if not exclusively via reinfection,[39,41–43] and HERV-H (LTR7 and LTR7Y), which amplified via both reinfection and retrotransposition.[40] It is interesting that the orthologus CpG/non-CpG ratio of both HERVK(HML2) and HERV-H are significantly higher than that of the generic non-LTR retrotransposon (AluSc8), suggesting their methylation level is even higher than common retrotransposons in human genome. Our analysis of HERVs implies ERV reinfection may lead to excessive replication cycle and excessive substitution especially in non-LTR context.

4.3.6 Hyperactive ERVs in mouse genome behave differently from HERVs

The ERV activity has declined substantially in human genome.[44] Currently only one HERV family might still be capable of moving (HERVK-HML2).[41] In contrast, multiple ERV families have been actively moving in mouse lineage (MuERV-L, IAP, MusD/ETn, etc.), including some ERV families invaded mouse genome recently (MuLV and MMTV).[45] With a method that can distinguish germline-hypomethylated or infectious ERVs in human genome, we applied it to mouse genome which hosts many more active and diverse ERV

families than human genome does.

We separated and plotted non-CpG and CpG substitution density of different mouse TEs, and applied the same statistical framework we used for human ERVs to mouse ERVs (Figure 4.6). With the same criteria, we identified 28 mouse ERV subfamilies of which have a CpG/non-CpG ratio is significantly lower than of non-LTR retrotransposons. All of them are mouse specific ERV subfamilies and do not have orthologs in other available species genome to separately calculate CpG/non-CpG substitution ratio after integration like we did for HERVs shared by human and chimpanzee. Nonetheless we estimated their germline methylation using WGBS dataset from mouse spermatogenesis.[22] To our surprise, these mouse ERVs with low CpG/non-CpG ratio are all hypermethylated during spermatogenesis. The hypermethylation of them implies high mutation rate at CpG sites after proviral integration, leaving excessive accumulation of non-CpG substitutions before proviral integration as the best hypothesis (Figure 4.7).

Intriguingly, different from HERVs in human genome, those ERVs accumulated excessive non-CpG substitutions before integration do not correlate with known infectious ERVs in mouse genome. There are some infectious mouse ERV showing low CpG/non-CpG ratio including MuLV and GLN (Figure 4.6). Besides, we identified VL30, a nonautonomous element can become infectious by packing itself within MuLV particles,[46] with a low CpG/non-CpG ratio. Our ability to separate VL30 illustrated the advantage of our method: we can identify interesting nonautonomous elements without coding capacity. However, there are some classic noninfecting mouse ERVs revealed by our method as low CpG/non-CpG ratio, and there are some well-known ERVs with infectious capability identified as high CpG/non-CpG ratio. For example, mouse intracisternal A-type particle (IAP) is a successful ERV family without envelope gene and infectious capability. But their counterpart IAPE (for IAP-related elements containing an envelope domain) have envelope and thought to be infectious.[47,48] In contrast, we found IAP subfamilies with a low CpG/non-CpG ratio, and IAPE subfamilies with a high CpG/non-CpG ratio. Aside from IAPs, a large number of ERV subfamilies with low CpG/non-CpG ratio are ETn, which is noninfectious but hyperactive (Table 4.1). We believe the discrepancy between identified ERVs with low CpG/non-CpG ratio and known infectious ERV in mouse is because we are separating ERVs family based on how mutations were accumulated. For instance, despite the capability to encode an envelope protein and infect other cells, if IAP related elements containing an envelope (IAPE) did not amplify extensively as exogenous retrovirus then their mutation pattern may be more similar to non-LTR retrotransposon. On the

other hand, combining the young age and hyperactivity of IAPs, it is possible that there is a significant portion of mutations among IAP elements are accumulated during reverse transcription that drives the overall CpG/non-CpG ratio low. Replication through "random template model" may also contribute to the low CpG/non-CpG ratio.[49]

4.3.7 Application to other vertebrate genomes

Having validated the utility of our method in human and mouse genomes, we expanded our analysis to 42 other vertebrate genomes having a draft genome assembly and a reasonably comprehensive TE library. These included 42 mammals, three birds and five non-avian reptile species (crocodile and turtles). For brevity, we highlight here some of the most interesting observations gleaned from several of these species.

The analysis of the pig genome recovered another notoriously infectious family, PERV, which has been extensively characterized for its infectious capacity and the potential risk that this poses for xenotransplantation of pig organs into humans (Figure 4.8).[50] Reassuringly, while many ERV families have been identified in the pig genome, porcine endogenous retrovirus (PERV) is one of only two our approach detected as candidate infectious element. The other one is LTR6_ss, which represent a family of lineage specific solitary LTR without identifiable internal sequence. It is possible that full-length, infectious proviruses from this family still circulate among domesticated pigs.

The opossum genome was one of the most significantly enriched for ERVs with low CpG/non-CpG ratio (Figure 4.9). Interestingly, the opossum is also remarkable among mammals for its vast and diverse ERV population[51] as well as its massive lineage-specific expansion of Krüppel associated box (KRAB) zinc-finger repressors, possibly as an evolutionary response to pervasive ERV infiltration.[52] Our results further support this hypothesis as they suggest that the opossum has a rich collection of ERVs with the signature of recent infectious activity.

The rhesus macaque genome has also been noted for harboring several recently expanded ERV families likely introduced via cross-species transmission.[53] One of these families, known as MacERV4 and closely related to Simian Retrovirus 1 (SRV1),[54] is characterized by a very low CpG/nonCpG ratio in our analysis (Figure 4.10). To further characterize the amplification dynamics of this family in the rhesus genome, we estimated dN/dS employing PAML dN/dS analysis to assess selective constraint acting on its envelope domain during its propagation.[55] The results reveal a clear signal of purifying selection (dN/dS = 0.38***), consistent with the idea that MacERV4 spread primarily via reinfection.

Outside of mammals, we found many more LTR retrotransposons with low CpG/nonCpG ratio in three reptiles (crocodile, alligator, and turtle). On contrast, the distribution of non-LTR retrotransposons in these species resembles their mammalian counterpart. We still know little about ERV and LTR retrotransposon evolution in reptile system and our data implies they may be very different from mammalian ERVs.

4.4 Discussion

In this study, we described a new method to investigate ERV evolution. We showed that it is possible to make inference on the replication mode of ERVs by comparing the pattern of divergence at CpG and non-CpG sites within each subfamily. Compared to other types of TE such as non-LTR retrotransposons and DNA transposons, we observed that many ERV families display an atypical pattern of sequence divergence characterized by a relatively low CpG/non-CpG mutation ratio. Closer inspection of human ERVs revealed that this atypical substitution pattern can, for some families (LTR12c and MER57E3), be explained by unusually low levels of CpG methylation in the germline. Their hypomethylation also suggests their possible adaptation for host cellular regulatory functions. But, for other families, we found that this pattern does not reflect hypomethylation but rather the accumulation of mutations acquired prior to endogenization, likely during infectious cycles of replication. This model is corroborated by the fact that ERVs known to have spread via reinfection, such as HERVK(HML2). Applying our method to mouse genome, we found more ERV subfamilies with low CpG/non-CpG ratio in mouse (28 subfamilies) than in human (5 subfamilies). Different from HERVs, none of these mouse ERVs are currently hypomethylated during spermatogenesis, leaving the accumulation of excessive non-CpG substitutions before integration as the only feasible explanation. Some of them are known infectious ERVs like MuLV and nonautonomous element VL30. But among those low CpG/non-CpG ERV subfamilies some of them are not known to be infectious. On the other hand, IAPE, a family with a functional envelope domain and supposed to be infectious, does not have excessive non-CpG substitutions.

We believe the difference we observed in human and mouse can be explained by the ERV age difference between these two hosts. In human, most of ERVs inserted our genome millions of years ago and have accumulated sufficient number of after integration substitutions. However, there are many more mouse ERVs inserted recently and do not have many after integration substitutions. Therefore, their CpG/non-CpG ratio is much more sensitive. In fact, we observed that all of these mouse ERVs with low CpG/non-CpG ratio but not

known to be infectious are highly active in mouse genome.[45] Therefore, for relatively aged ERVs (HERVs in human genome), excessive non-CpG substitutions can be best explained by reinfection; for recent and current active ERVs (mouse ERVs), rampant intracellular retrotransposition coupled with "random template model" replication may be enough to produce enough excessive non-CpG substitutions.

Non-LTR retrotransposons like *Alu* and B2 elements are very active and achieved millions of copies. Why their hyperactivity will not result in excessive non-CpG mutations? We believe that is because within non-LTR retrotransposon subfamilies they replicated via "strict master model." Therefore, every copy is just a few replication cycles and substitutions away from their ancestral master copy. On contrast, we believe most of ERVs replicated via "random template model," by which heterogeneity increase substantially within subfamilies.

Building on these observations, we established a statistical framework to identify ERVs bearing the signature of reinfection and applied our method to analyze 53 vertebrate genomes. Our analysis suggests that PERV, which is known to produce infectious viral particles, is one of at least two ERV families in the pig genome that has spread mostly via reinfection. The method also identified at least one ERV family in the rhesus macaque genome that has probably undergone recent episodes of reinfection. But among all the mammal species examined, it is the opossum that stands out for possibly hosting highest number of ERV families with a signature of recent infectious activities. Of course, at this point, we cannot exclude the possibility that the atypical substitution pattern observed for these ERVs simply reflect their escape from CpG methylation in the germ line. Further analyses, both experimental and computational, such as selection analysis of envelope domains, will be necessary to confirm the hypothesis that these elements have primarily spread by reinfection. Nonetheless, we contend that our approach is useful to quickly screen genomes for retroviral-derived sequences that have amplified via reinfection, even those bearing no coding sequences, such as solo LTRs and nonautonomous elements. As many more genome sequences become available for a wide range of vertebrates, along with their own diverse ERV repertoire, our approach will accelerate the identification of ERV families with infectious potential and further our ability to delineate the paleohistory and epidemiological dynamics of these elements in their host population.

4.5 Method

4.5.1 CpG and non-CpG mutation density calculation

We downloaded RepeatMasker alignment files of 57 Species from repeatmasker.org and developed genomeRM_CpG.pl (https://github.com/xzhuo/CpGrate) to process the downloaded RepeatMasker alignment files. It builds a multiple sequence alignment for each repeat subfamily, estimate a new consensus based on the alignment using majority rule. For HIV-1 strains, the multiple sequence alignement was downloaded directly from HIV databases (https://www.hiv.lanl.gov) and the consensus was estimated by majority rule. Then the new consensus is used as the ancestral sequence to compare with each sequence in the alignment. For each comparison, the numbers of cytosine (C), guanine (G), and CpG sites in the ancestral sequence, the number of C to thymine (T) transition, G to adenine (A) transition, CpG to CpA transition, and CpG to TpG transition from ancestral state to current sequence are counted and summed. In our analysis, only full length TE copies (missing fragment $<$ 50 bp at both ends) are used to build the alignment, and only TE subfamilies with more than 30 copies in the alignment are included in the calculation. We also excluded TpG in ancestral sequences from non-CpG sites because some of them are actually CpG in real ancestral sequence and including them in non-CpG sites inflated their substitution rate (Figure 4.1). The same calculation is applied to HIV-1 multiple strain alignment file to calculate CpG/nonCpG mutation rate of exogenous retroviruses.

We further processed and plotted the data with nonlinear_regression_calculation.py (https://github.com/xzhuo/CpGrate). CpG substitution density (D_{cg}) is calculated as number of mutation from CpG to TpG divided by number of all CpG sites in ancestral state. As comparison, nonCpG substitution density (D_{ncg})is the number of C to T mutation divided by number of C in the ancestral sequence.

4.5.2 Curve fitting and candidate ERV subfamily selection

We modified the function in Xing et al. to describe the CpG substitution density with non-CpG substitution density f(a,r,x):

$$D_{cg} = \frac{0.5}{1+a} * \left[1 - \left(1 - \frac{4}{3} * D_{\text{ncg}}\right)^{3*r/4}\right] \tag{4.1}$$

Where D_{cg} and D_{ncg} stands for CpG substitution density and non-CpG substitution density, and a and r are two parameters used to describe their relationship.

Then we estimated parameters a and r using non-linear regression by fitting non-LTR retrotransposons data. Lower 95% prediction interval of the curve is calculated using delta method:

$$\text{Lower 95\% PI} = \hat{D_{\text{cg}}} - t_{0.95,\text{DF}} * \sqrt{\left[\frac{\partial D_{\text{cg}}}{\partial a} \quad \frac{\partial D_{\text{cg}}}{\partial r}\right] * \text{COV} * \left[\frac{\frac{\partial D_{\text{cg}}}{\partial a}}{\frac{\partial D_{\text{cg}}}{\partial r}}\right] + \frac{\text{SS}}{\text{DF}}} \tag{4.2}$$

PI: prediction interval

$\hat{D}$cg: calculated CpG substitution density with fitted a and r.

t: critical t value

COV: covariance matrix of fitted a and r

SS: sum of square

DF: degree of freedom

Poisson exact 95% critical interval is estimated for all LTR subfamilies at both non-LTR and LTR distance. We assume there is no interaction between CpG and non-CpG sites and the critical interval ellipse is axis aligned. We calculated the probability of error ellipse above the lower 95% prediction interval of the curve by one tail test with distribution described above and corrected using false discovery rate (FDR). All LTR subfamilies with Q<0.05 are considered as positive, illustrated, and used for subsequent analysis.

4.5.3 Orthologus TE CpG and non-CpG mutation density calculation

Orthologus TE pairwise alignment between human and chimpanzee is extracted from chimp-human whole genome alignment file available in UCSC genome browser.[56,57]Since it is impossible to derive ancestral sequence using only pairwise alignment of ortholog, we counted number of shared C ($N_{C\text{-}C}$ or $N_{CG\text{-}CG}$) in the pairwise alignment and number of heterozygotes sites with (C/T) ($N_{C\text{-}T}$ or $N_{CG\text{-}TG}$) in both CpG and non-CpG context. We then calculated $N_{C\text{-}T}$ / $N_{C\text{-}C}$ and $N_{CG\text{-}TG}$ / $N_{CG\text{-}CG}$ as approximate of non-CpG substitution density (Dncg) and CpG substitution density (Dcg).

To test if Dcg/Dncg of given ERV subfamilies are different from Alu element, we conducted following Likelihood ratio test:

The poisson probability density function is:

$$f(x;\lambda) = \frac{\lambda^x}{x!}e^{\lambda} \tag{4.3}$$

The log likelihood function is:

$$L = -\ln [f(N_{1C-T}; \mu_1 N_{1C-C}) + f(N_{1CG-TG}; a\mu_1 N_{1CG-CG}) + f(N_{2C-T}; \mu_2 N_{2C-C}) + f(N_{2CG-TG}; b\mu_2 N_{2CG-CG})] \tag{4.4}$$

$N_{1C\text{-}C}$: number of preserved C in non-CpG context in the orthologus Alu alignment;

$N_{1C\text{-}T}$: number of C-T transition in non-CpG context in the orthologus Alu alignment;

$N_{1CG\text{-}CG}$: number of preserved C in CpG context in the orthologus Alu alignment;

$N_{1CG\text{-}TG}$: number of C-T transition in CpG context in the orthologus Alu alignment;

$N_{2C\text{-}C}$: number of preserved C in non-CpG context in the orthologus TE alignment;

$N_{2C\text{-}T}$: number of C-T transition in non-CpG context in the orthologus TE alignment;

$N_{2CG\text{-}CG}$: number of preserved C in CpG context in the orthologus TE alignment;

$N_{2CG\text{-}TG}$: number of C-T transition in CpG context in the orthologus TE alignment;

There are 4 parameters in the equation:

μ_1: substitution density at non-CpG sites (Dncg) in Alu;

μ_2: substitution density at non-CpG sites (Dncg) in the given TE;

a: CpG/non-CpG substitution ratio (Dcg/Dncg) in Alu;

b: CpG/non-CpG substitution ratio (Dcg/Dncg) in the given TE.

Under null hypothesis, we assume a = b. Under alternative hypothesis, we have $a \neq b$. Maximum likelihood estimation was conducted with bbmle package in R and Likelihood ratio test is subsequently performed.

4.5.4 CpG methylation from whole genome bisulfite sequencing data

The whole genome bisulfite sequencing (WGBS) data is obtained through cooperation.[22] All the reads are mapped to hg19 or mm10. We intersected TE with WGBS data using bedtools to produce TE methylation data set,[58] then we plotted TE methylation level with violin plot using ggplot2 in R.[59]

4.6 Bibliography

[1] Ray, D. A.; Feschotte, C.; Pagan, H. J. T.; Smith, J. D.; Pritham, E. J.; Arensburger, P.; Atkinson, P. W.; Craig, N. L. *Genome Res.* **2008**, *18*, 717–728.

[2] Huang, C. R. L.; Burns, K. H.; Boeke, J. D. *Annu. Rev. Genet.* **2012**, *46*, 651–675.

[3] Katzourakis, A.; Magiorkinis, G.; Lim, A. G.; Gupta, S.; Belshaw, R.; Gifford, R. *PLoS Pathog.* **2014**, *10*, e1004214.

[4] Magiorkinis, G.; Gifford, R. J.; Katzourakis, A.; De Ranter, J.; Belshaw, R. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 7385–7390.

[5] Zhuo, X.; Feschotte, C. *PLoS Pathog.* **2015**, *11*, e1005279.

[6] Hayward, A.; Grabherr, M.; Jern, P. *Proc. Natl. Acad. Sci. U.S.A.* **2013**,

[7] Hayward, A.; Cornwallis, C. K.; Jern, P. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 464–469.

[8] Coffin, J. M.; Hughes, S. H.; Varmus, H. E. *Retroviruses*; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 1997.

[9] Dewannieux, M.; Heidmann, T. *Curr. Opin. Virol.* **2013**, *3*, 646–656.

[10] Law, J. A.; Jacobsen, S. E. *Nat. Rev. Genet.* **2010**, *11*, 204–220.

[11] Coulondre, C.; Miller, J. H.; Farabaugh, P. J.; Gilbert, W. *Nature* **1978**, *274*, 775–780.

[12] Brunet, T. D. P.; Doolittle, W. F. *Genome Biol. Evol.* **2015**, *7*, 2445–2457.

[13] Meunier, J.; Khelifi, A.; Navratil, V.; Duret, L. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 5471–5476.

[14] Xing, J.; Hedges, D. J.; Han, K.; Wang, H.; Cordaux, R.; Batzer, M. A. *J. Mol. Biol.* **2004**, *344*, 675–682.

[15] Smith, Z. D.; Meissner, A. *Nat. Rev. Genet.* **2013**, *14*, 204–220.

[16] Jern, P.; Stoye, J. P.; Coffin, J. M. *PLoS Genet.* **2007**, *3*, e183.

[17] Esnault, C.; Heidmann, O.; Delebecque, F.; Dewannieux, M.; Ribet, D.; Hance, A. J.; Heidmann, T.; Schwartz, O. *Nature* **2005**, *433*, 430–433.

[18] Knisbacher, B. A.; Gerber, D.; Levanon, E. Y. *Trends Genet. : TIG* **2015**,

[19] Oehlert, G. W. *Am. Stat.* **1992**, *46*, 27–29.

[20] Clopper, C. J.; Pearson, E. S. *Biometrika* **1934**, *26*, 404–413.

[21] Nei, M.; Kumar, S. *Molecular Evolution and Phylogenetics*; Oxford University Press, 2000.

[22] Hammoud, S. S.; Low, D. H. P.; Yi, C.; Carrell, D. T.; Guccione, E.; Cairns, B. R. *Cell Stem Cell* **2014**, *15*, 239–253.

[23] Grimwood, J. et al. *Nature* **2004**, *428*, 529–535.

[24] Feschotte, C. *Nat. Rev. Genet.* **2008**, *9*, 397–405.

[25] Sundaram, V.; Cheng, Y.; Ma, Z.; Li, D.; Xing, X.; Edge, P.; Snyder, M. P.; Wang, T. *Genome Res.* **2014**, *24*, 1963–1976.

[26] Chuong, E. B.; Rumi, M. A. K.; Soares, M. J.; Baker, J. C. *Nat. Genet.* **2013**,

[27] Lukic, S.; Nicolas, J.-C.; Levine, A. J. *Cell Death Differ.* **2014**, *21*, 381–387.

[28] Malfavon-Borja, R.; Feschotte, C. *J. Virol.* **2015**, *89*, 4047–4050.

[29] Babaian, A.; Mager, D. L. *Mob. DNA* **2016**, *7*, 24.

[30] Di Cristofano, A.; Strazzullo, M.; Longo, L.; La Mantia, G. *Nucleic Acids Res.* **1995**, *23*, 2823–2830.

[31] Chen, H.-J.; Carr, K.; Jerome, R. E.; Edenberg, H. J. *DNA Cell Biol.* **2002**, *21*,

793–801.

[32] Pi, W.; Zhu, X.; Wu, M.; Wang, Y.; Fulzele, S.; Eroglu, A.; Ling, J.; Tuan, D. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 12992–12997.

[33] St Laurent, G.; Shtokalo, D.; Dong, B.; Tackett, M. R.; Fan, X.; Lazorthes, S.; Nicolas, E.; Sang, N.; Triche, T. J.; McCaffrey, T. A.; Xiao, W.; Kapranov, P. *Genome Biol.* **2013**, *14*, R73.

[34] Sokol, M.; Jessen, K. M.; Pedersen, F. S. *Retrovirology* **2015**, *12*, 32.

[35] Hashimoto, K. et al. *Genome Res.* **2015**, *25*, 1812–1824.

[36] Beyer, U.; Moll-Rocek, J.; Moll, U. M.; Dobbelstein, M. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 3624–3629.

[37] Beyer, U.; Krönung, S. K.; Leha, A.; Walter, L.; Dobbelstein, M. *Cell Death Differ.* **2016**, *23*, 64–75.

[38] Krönung, S. K.; Beyer, U.; Chiaramonte, M. L.; Dolfini, D.; Mantovani, R.; Dobbelstein, M. *Oncotarget* **2016**, *7*, 33484–33497.

[39] Belshaw, R.; Pereira, V.; Katzourakis, A.; Talbot, G.; Paces, J.; Burt, A.; Tristem, M. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 4894–4899.

[40] Belshaw, R.; Katzourakis, A.; Paces, J.; Burt, A.; Tristem, M. *Mol. Biol. Evol.* **2005**, *22*, 814–817.

[41] Grow, E. J.; Flynn, R. A.; Chavez, S. L.; Bayless, N. L.; Wossidlo, M.; Wesche, D. J.; Martin, L.; Ware, C. B.; Blish, C. A.; Chang, H. Y.; Reijo Pera, R. A.; Wysocka, J. *Nature* **2015**,

[42] Wildschutte, J. H.; Williams, Z. H.; Montesion, M.; Subramanian, R. P.; Kidd, J. M.; Coffin, J. M. *Proc. Natl. Acad. Sci. U.S.A.* **2016**,

[43] Dewannieux, M.; Harper, F.; Richaud, A.; Letzelter, C.; Ribet, D.; Pierron, G.; Heidmann, T. *Genome Res.* **2006**, *16*, 1548–1556.

[44] Magiorkinis, G.; Blanco-Melo, D.; Belshaw, R. *Retrovirology* **2015**, *12*, 8.

[45] Stocking, C.; Kozak, C. A. *Cell Mol. Life Sci.* **2008**, *65*, 3383–3398.

[46] French, N. S.; Norton, J. D. *Biochim. Biophys. Acta* **1997**, *1352*, 33–47.

[47] Ribet, D.; Harper, F.; Dupressoir, A.; Dewannieux, M.; Pierron, G.; Heidmann, T. *Genome Res.* **2008**, *18*, 597–609.

[48] Dewannieux, M.; Vernochet, C.; Ribet, D.; Bartosch, B.; Cosset, F.-L.; Heidmann, T. *PLoS Pathog.* **2011**, *7*, e1002309.

[49] Nascimento, F. F.; Rodrigo, A. G. *PLoS ONE* **2016**, *11*, e0162454.

[50] Kimsa, M. C.; Strzalka-Mrozik, B.; Kimsa, M. W.; Gola, J.; Nicholson, P.; Lopata, K.; Mazurek, U. *Viruses* **2014**, *6*, 2062–2083.

[51] Gentles, A. J.; Wakefield, M. J.; Kohany, O.; Gu, W.; Batzer, M. A.; Pollock, D. D.; Jurka, J. *Genome Res.* **2007**, *17*, 992–1004.

[52] Thomas, J. H.; Schneider, S. *Genome Res.* **2011**, *21*, 1800–1812.

[53] Han, K.; Konkel, M. K.; Xing, J.; Wang, H.; Lee, J.; Meyer, T. J.; Huang, C. T.; Sandifer, E.; Hebert, K.; Barnes, E. W.; Hubley, R.; Miller, W.; Smit, A. F. A.; Ullmer, B.; Batzer, M. A. *Science* **2007**, *316*, 238–240.

[54] van der Kuyl, A. C.; Mang, R.; Dekker, J. T.; Goudsmit, J. *J. Virol.* **1997**, *71*, 3666–3676.

[55] Yang, Z. *Mol. Biol. Evol.* **2007**, *24*, 1586–1591.

[56] Schwartz, S.; Kent, W. J.; Smit, A.; Zhang, Z.; Baertsch, R.; Hardison, R. C.; Haussler, D.; Miller, W. *Genome Res.* **2003**, *13*, 103–107.

[57] Kent, W. J.; Baertsch, R.; Hinrichs, A.; Miller, W.; Haussler, D. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 11484–11489.

[58] Quinlan, A. R.; Hall, I. M. *Bioinformatics* **2010**, *26*, 841–842.

[59] Wickham, H. *ggplot2*; Elegant Graphics for Data Analysis; Springer Science & Business Media: New York, NY, 2009.

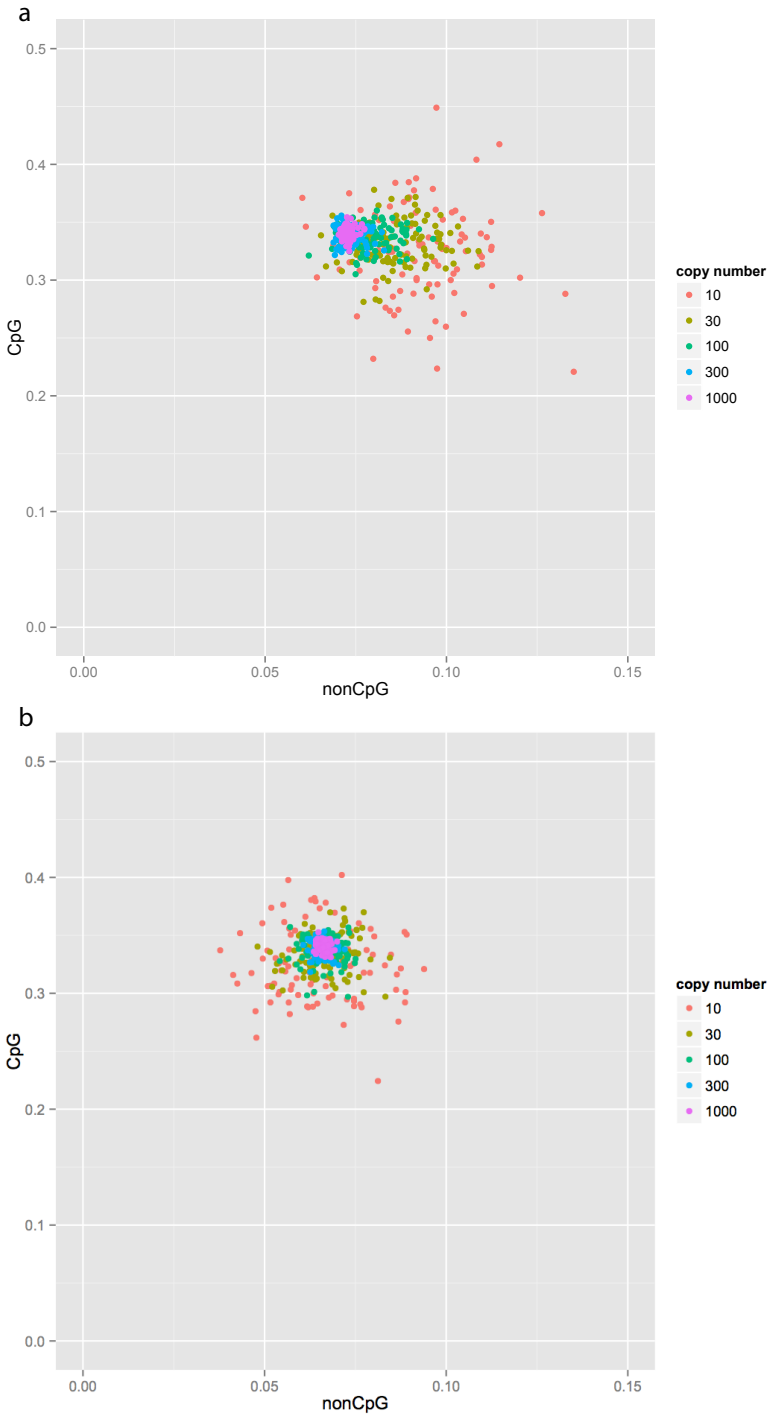

Figure 4.1. CpG/non-CpG mutation ratio is not systematically affected by copy number of TE family if CpA sites are excluded from calculation. (a). Simulated CpG/nonCpG mutation rate of different copy number using AluSz6. The substitution rate of non-CpG site is overestimated for families with low copy number in the simulation. (b) CpG and non-CpG rate no longer biased if CpA sites are excluded from calculation.

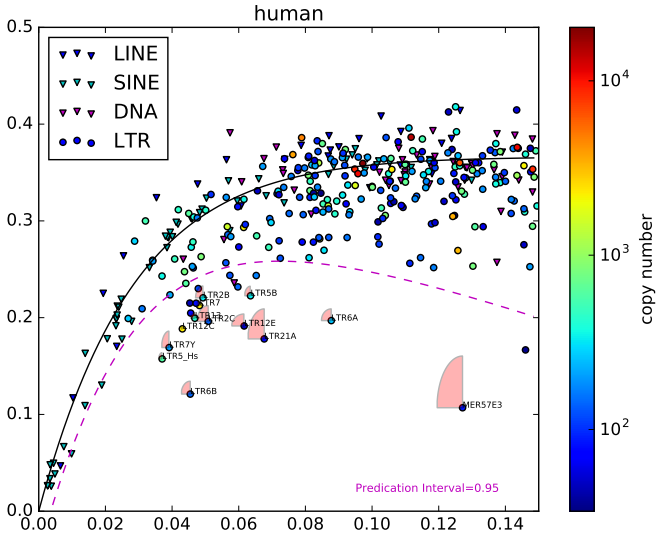

Figure 4.2. The distribution of nonCpG and CpG density of other TEs are fitted with a non-linear curve (see method). Ninty-five percent Prediction interval and Ninty-five percent critical interval are both used and plotted to distinguish ERVs with significal CpG/nonCpG ratio.

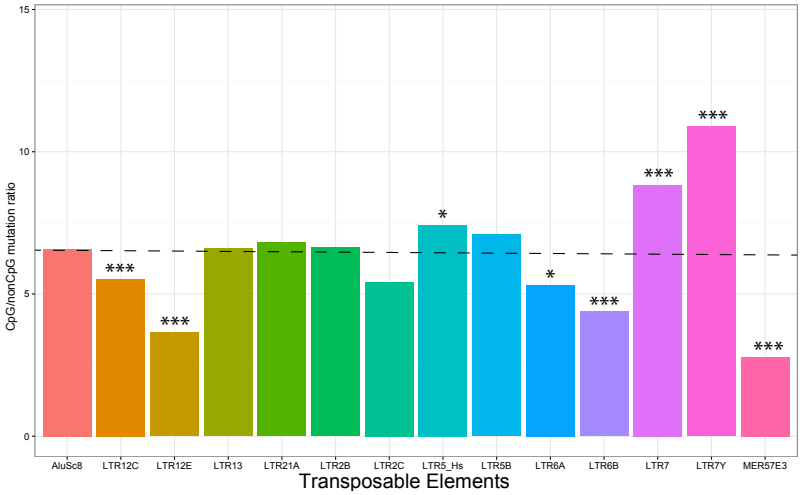

Figure 4.3. We found five ERVs that are hypomethylated during spermatogenesis have low CpG/non-CpG substitution ratio if we only include mutations after integration. However, the other eight ERV candidates have high CpG/non-CpG ratio with the calculation.

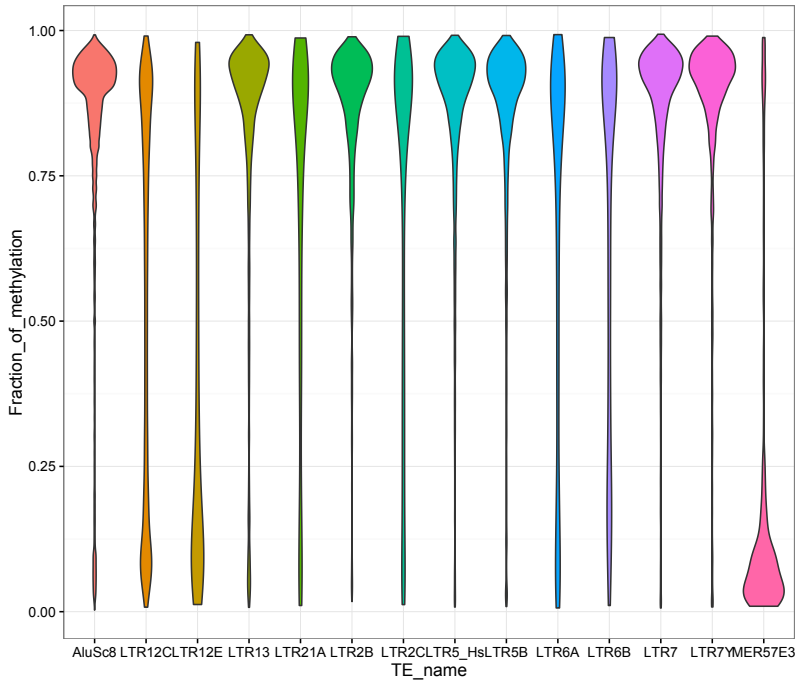

Figure 4.4. Violin plot of methylation level of 13 HERV candidates in spermatogonial sterm cells.

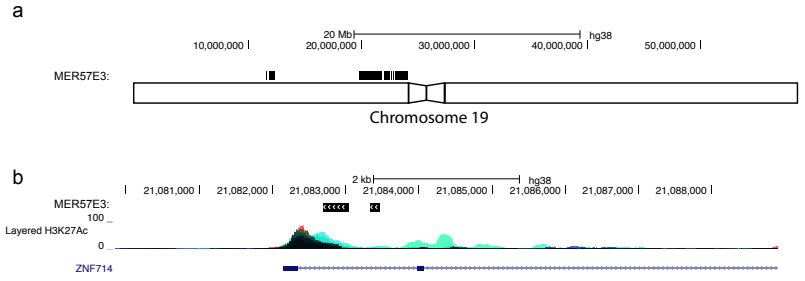

Figure 4.5. Association of MER57E3 with promoters. (a) Cluster of MER57E3 on chromosome 19. (b) MER57E3 is associated with ZNF714 promoter.

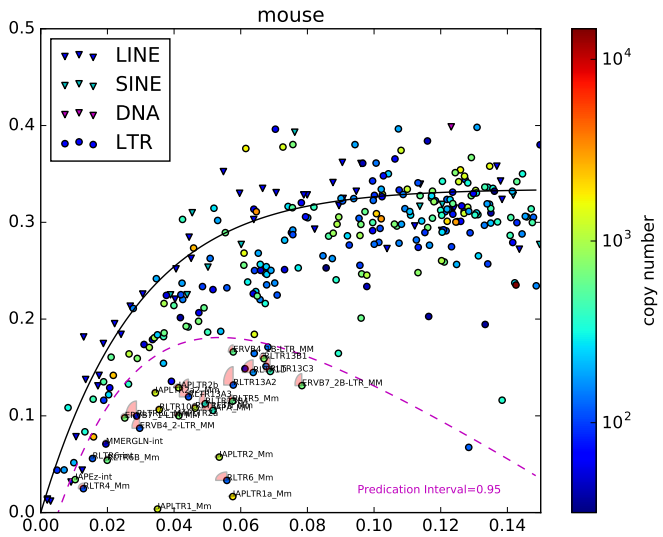

Figure 4.6. CpG mutation pattern of ERVs in mouse genome.

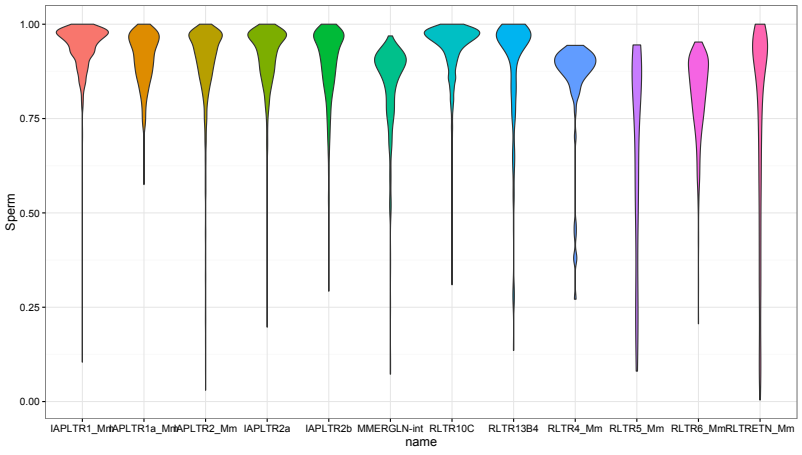

Figure 4.7. Violin plot of methylation level of 12 ERV candidates in spermatogonial sterm cells.

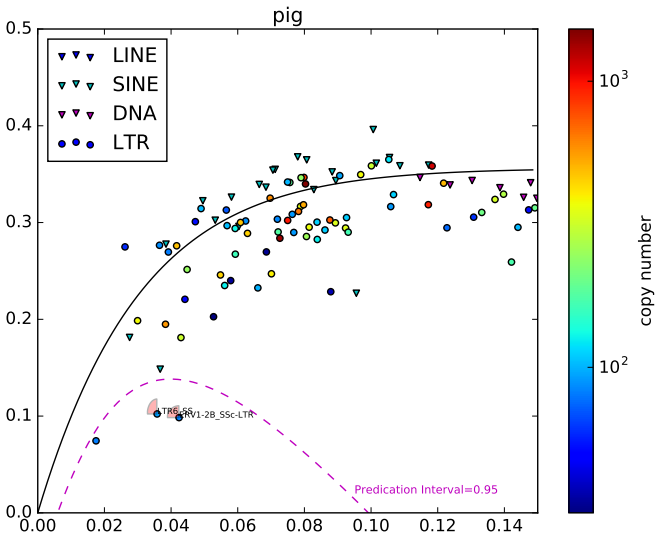

Figure 4.8. CpG mutation pattern of ERVs in pig genome.

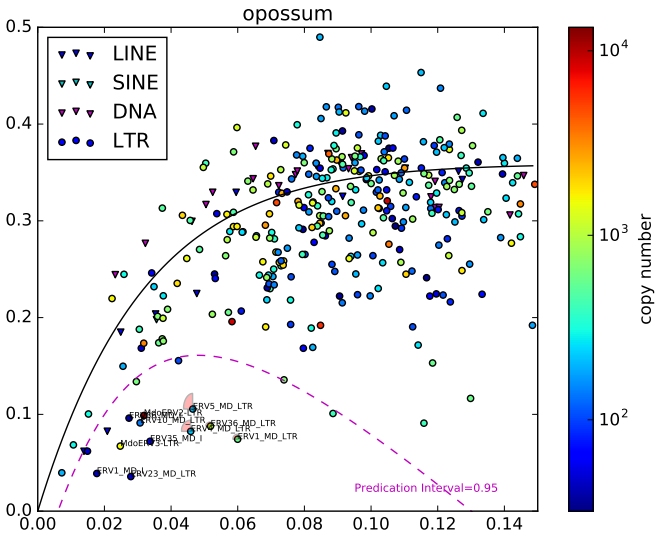

Figure 4.9. CpG mutation pattern of ERVs in opossum genome.

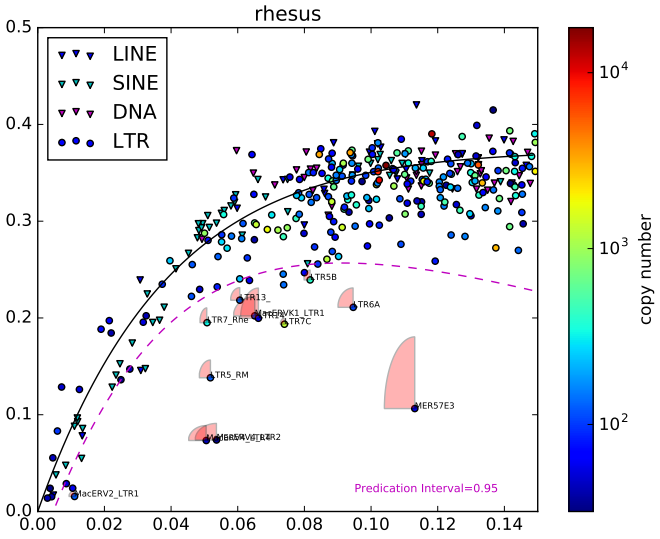

Figure 4.10. CpG mutation pattern of ERVs in rhesus genome.

Table 4.1. Common ERVs in the mouse genome.

Family	envelope	infectious	autonomous	repbase name	Copy No.
IAP	No	No	Yes	IAPEz, IAPLTR1_Mm	High
IAPE	Yes	Yes	Yes	IAPEy, IAPEY2-IAPEY5	Low
MuLV	Yes	Yes	Yes	MuLV, RLTR4_Mm	Low
VL30	No	Yes	No	RLTR6_Mm	High
MusD	No	No	Yes	ERVB7_1	High
GLN	Yes	Yes	Yes	MMERGLN	Low
Etn	No	No	No	RLTR13, RLTRETN_Mm	High
MERVL	No	No	Yes	MT2_Mm	High
MMTV	Yes	Yes	Yes	MMTV	Low

CHAPTER 5

CONCLUSION

Here we investigated ERV evolution in the little brown bat genome, described an intriguing case of cross-species retroviral transmission, and developed a new method to study ERV evolution.

I identified 362 full length proviruses in assembled little brown bat genome using LTRharvest, LTRdigest, and MGEScan-LTR.[1–3] Then I used a homology-based method to scan the whole genome and mined virtually all the fragments that are derived from ERVs in the bat genome. Taken together, ERVs contributed to about 5% of the little brown bat genome. Despite rampant activity of DNA transposons within the vesper bat lineage, the amount of ERV sequence is comparable with that of other mammals such as human, dog, or mouse. Thus, it seems that ERV abundance is not directly affected by the plethora of circulating zoonotic viruses in this reservoir species. Despite this, I still found hundreds of lineage-specific, recently integrated ERVs in the little brown bat genome. Some copies even display complete and apparently intact coding capacity, including envelope genes, suggestive of recent and potentially current infectious capability.

Building on the catalogue of ERVs I assembled in the little brown bat genome, I subsequently searched for highly similar ERV sequences in other mammalian genomes in order to look for potential cases of cross-species transmission. These searches led to the identification of genomic sequences highly similar to the bat MLERV1 element in cat, tiger, and pangolin genomes. Thus, this ERV family was found to be present in other mammals but with a very patchy distribution, indicative of multiple independent introduction in these species lineages. Furthermore, the level of interspecific sequence similarity between these elements was much higher than what would be expected from a scenario invoking their vertical inheritance from a common ancestor. Together these results strongly suggested a case of cross-species retroviral transmission among these species or some of their ancestors.

I further dated the endogenization of this retrovirus in these species lineages using two independent methods, which led to an estimated introduction between 10 and 20 million

years ago. The data also suggested that the ERV has continued to proliferate actively in both the cat and bat lineages for millions of years until very recently, generating hundreds of insertions unique to each species. I used a phylogenomics approach to further investigate the mode and tempo of this ERV's amplification in the cat and bat lineages. The results suggested that the ancestral retrovirus underwent a single endogenization event in the cat lineage and subsequently lost its infectious capacity and amplified predominantly if not exclusively by retrotransposition. The evolutionary history and dynamics of the family was more complex in the bat lineage, where we could infer there were at least three different independent endogenization events, and potentially many more. In fact, it appears that several MLERV1 elements remain infectious for extended period in the bat lineage, and we identified several copies that may still be capable of reinfection. Together this study illustrated how a nearly identical retrovirus can be endogenized in widely diverged mammal species, but follow a different fate and amplification dynamics in each of the species lineages.

In the last part of my dissertation, I developed a method to explore more systematically the mode of proliferation of ERVs in different species. In particular, we wanted to assess the extent by which ERVs have spread via reinfection or retrotransposition in a large sample of vertebrate species. Conventionally, these two different modes of proliferation are distinguished by examining whether the envelope coding domain, which is required for infection, is present and has evolved under functional constraint while the ERV proliferated in the genome,.[4,5] However, this approach suffers from several limitations that make it extremely difficult to apply at a genome-wide scale for a large collection of species. To overcome these shortcomings, I developed a new and relatively simple approach to investigate ERV evolution based on the prediction that ERVs amplified via reinfection would show a different mutation pattern than those amplified via intracellular retrotransposition. Specifically, I hypothesized that reinfecting ERVs, as they diverge from each other, would display a relatively less pronounced transition bias at CpG sites compared to nonCpG sites. We validated this prediction using ERVs known to have undergone reinfection in the human and mouse lineages. We then developed a statistical framework and computational pipeline to automate the analysis and apply it to profile a wide range of vertebrate genomes. We successfully recovered the infamous infectous porcine endogenous retrovirus (PERV) in pig genome, and our finding suggesting that opossum is loaded with reinfecting ERVs.

However, the method we developed here relies on other TEs to define CpG substitution pattern of hypermethylated retrotransposon. It has been demonstrated that most of TEs are hypermethylated in germline, but it is still possible that some of them are not. Besides,

the sensitivity of our method will be compromised if the number of available non-ERV TE families is low. Therefore, this method can be improved by employing better statistics.

My dissertation research demonstrated that worth of characterizing ERVs. We unveiled an intriguing case of ancient cross-species transmission as well as many ancient infections. By examining substitution pattern of ERVs, we can even distinguish how they propagated in the host genome. Retroviral zoonosis is a potential threat to public health and wild animal conservation.[6,7] Here I described a past retroviral zoonosis among different mammalian orders 10-20 million years ago and the method developed in my dissertation can be used to find potential infectious ERVs and help prevent possible future retroviral zoonosis.

5.1 Bibliography

[1] Ellinghaus, D.; Kurtz, S.; Willhoeft, U. *BMC Bioinformatics* **2008**, *9*, 18.

[2] Steinbiss, S.; Willhoeft, U.; Gremme, G.; Kurtz, S. *Nucleic Acids Res.* **2009**, *37*, 7002–7013.

[3] Rho, M.; Choi, J.-H.; Kim, S.; Lynch, M.; Tang, H. *BMC Genomics* **2007**, *8*, 90.

[4] Belshaw, R.; Pereira, V.; Katzourakis, A.; Talbot, G.; Paces, J.; Burt, A.; Tristem, M. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 4894–4899.

[5] Belshaw, R.; Katzourakis, A.; Paces, J.; Burt, A.; Tristem, M. *Mol. Biol. Evol.* **2005**, *22*, 814–817.

[6] Locatelli, S.; Peeters, M. *AIDS.* **2012**, *26*, 659–673.

[7] Xu, W.; Eiden, M. V. *Annu. Rev. Virol.* **2015**, *2*, 119–134.