The role of mobile elements in recent primate genomes

Wanxiangfu Tang, M.Sc.

Department of Biological Sciences

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

Faculty of Mathematics and Science, Brock University
St. Catharines, Ontario

Abstract

Mobile elements (MEs), which constitute ~50% of the primate genomes, have contributed to both genome evolution and gene function as demonstrated by ample evidence discovered over the last few decades. The three studies in this thesis aims to provide a better understanding of the evolutionary profile and function of MEs in the primate genomes by taking a computational comparative genomics approach.

The first study represents a comprehensive analysis of the differential ME transposition among primates via identification of species-specific MEs (SS-MEs) in eight primate genomes from the families of *Hominidae* and *Cercopithecidae* using a comparative genomics approach. In total, 230,855 SS-MEs are identified, which reveal striking differences in retrotransposition level in the eight primate genomes. The second study represents a more focused analysis for the identification of a new type of MEs, which we term "retro-DNA" for non-LTR retrotransposons derived from DNA transposons, in the recent primate genomes. By investigating biallelic DNA transposons that have both the insertion and pre-integration alleles in ten primate genomes, a total of 1,750 retro-DNA elements representing 750 unique insertion events are reported for the first time. The third study provides an analysis of the mechanism underlying the differential SINE transposition in the primate genomes. In this study, Alu profiles are compared and the Alu master copies are identified in six primate genomes in the *Hominidae* and *Cercopithecidae* groups. The results show that each lineage of the primates and each species owns a unique Alu profile exclusively defined by the AluY transposition activity, which is determined by the number of Alu master copies and their relative activity.

Overall, work in this thesis provides new insights about MEs and their impact on the recent primate genomes by revealing differential ME transposition as an important mechanism in generating genome diversity among primate lineages and species through discovering a new type of MEs and preliminary analysis of the mechanism underlying the differential ME transposition among primates. Furthermore, taking advantage of the recently available primate genomes and transcriptomes data, the work in this thesis demonstrates the great potential of the comparative genomic approach in studying MEs in primate genomes.

Acknowledgement

I would first like to thank my supervisor, Professor Ping Liang, for his continuous support throughout the years. Without his help, both academically and financially, this thesis would never come to reality. I would also like to thank Drs. Alan Castle and Doug Bruce, who are in my supervisory committee, for their kind advice on the projects.

I would like to thank my wife, Guang Ling, for everything during my pursuit for Ph.D., and my parents, for their love and support, from thousands of miles away. I would also like to thank my peers and friends for all the help and inspiration, particularly Radesh, Jina and Musa.

Table of contents

List of tables

List of figures

List of Abbreviations

BLAST, Basic Local Alignment Search Tool

BLAT, BLAST-Like Alignment Tool

CDS, CoDing Sequences

ERV, Endogenous RetroViruses

LCA, Last Common Ancestor

LINE, Long INterspersed Elements

LTR, Long Terminal Repeat

ME, Mobile Element

MITE, Miniature Inverted-repeat Transposable Element

MR SS-ME, Most Recent Species-Specific Mobile Element

MUSCLE, MUltiple Sequence Comparison by Log-Expectation

NCBI, National Center for Biotechnology Information

NGS, next-generation sequencing

NHPRTR, Non-Human Primate Reference Transcriptome Resource

ORF, Open Reading Frame

PERL, Practical Extraction and Reporting Language

SINE, Short-INterspersed Elements

SRA, Sequence Read Archive

SS-ME, Species-Specific Mobile Element

SVA, SINE-R/VNTR/Alu

TE, Transposable Element

TIR, Terminal Inverted Repeat

TPRT, Target-Prime Reverse Transcription

TSD, Target Site Duplication

UCSC, University of California at Santa Cruz

UTR, UnTranslated Region

Chapter 1 General Introduction

Mobile elements (MEs), which are DNA elements that can either move or copy in the genomes, have been proven to have immense impact on both genome evolution and gene function. These MEs, which make up to ~50% of primate genomes including the human genome, are known to affect the host genomes through many different mechanisms, such as generating insertion mutations, genomic instability, new genes or splicing variants, and alteration in gene expression. This chapter provides background information on MEs which are relevant to the subsequent chapters. Chapter 2 presents a study that identifies and characterizes species-specific MEs (SS-MEs) in eight primate genomes from the families of *Hominidae* and *Cercopithecidae*, focusing on retrotransposons. Chapter 3 presents a study that reports a new type of non-LTR retrotransposon, named as retro-DNA, which represent DNA transposons by sequence but non-LTR retrotransposons by retrotransposition mechanism, in the recent primate genomes. Chapter 4 presents a study that examines Alu profile and identifies Alu master copies in six primate genomes from the *Hominidae* and *Cercopithecidae* families to better understand the differential Alu transposition in primate genomes. Chapter 5 contains general discussions for the aforementioned data chapters.

1.1 Mobile elements in the primate genomes

MEs are defined as genomic DNA sequences, which can mobilize in the host genomes, either by changing their own positions or by making new copies and inserting into other locations. As shown in Fig. 1.1, MEs are very successful in the genomes of higher eukaryotic species such as primates, as they can be found abundantly in these genomes; MEs' contribution to the primate genomes can range from 46.8% in the green monkey genome to 50.7% in the baboon genome (Carbone et al. 2014; Chimpanzee Sequencing and Analysis 2005; Cordaux and Batzer 2009; Deininger et al. 2003; Lander et al. 2001; Locke et al. 2011; Rhesus Macaque Genome Sequencing and Analysis et al. 2007; Scally et al. 2012; Yan et al. 2011).

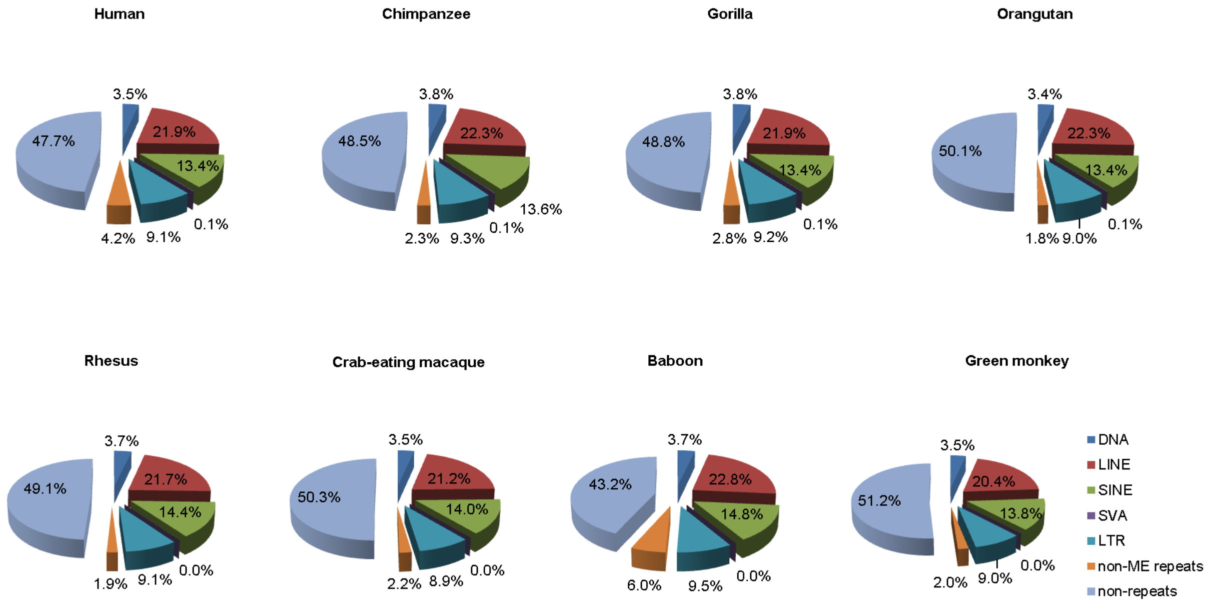


Figure 1.1 Mobile elements composition by type in the primate reference genomes

This is an unpublished figure originally prepared for author's publication "Comparative genomics analysis reveals high levels of differential retrotransposition among primates from the Hominidae and the Cercopithecidae families, Genome Biology and Evolution, evz234, https://doi.org/10.1093/gbe/evz234"

The percentage of MEs in each genome is calculated using the most updated versions of primate reference genomes, excluding gap sequences. The colour scheme for different ME types is the same for each panel.

Because of their abundances, MEs are highly repetitive in the host genomes and therefore are creating difficulties for genome assembly and annotation. The gap regions in these genomes are usually biased towards the repeat sequence regions involving MEs. With further improvements in these genomes, especially in the current gap regions, the percentage of MEs is expected to increase slightly.

By the mechanism of their transposition, MEs can be divided into two major classes: DNA transposons and retrotransposons (Stewart et al. 2011). DNA transposons, which constituent approximately 3.6% of the primate genomes, are able to excise themselves out from their original locations and move to new sites in the genome in the form of DNA, leading to no direct change of their copy numbers in the genome during the process (Pace Ii and Feschotte 2007). In comparison, retrotransposons mobilize in genomes via an RNA-based duplication process called retrotransposition, in which a retrotransposon is first transcribed into RNA and then reverse transcribed into DNA as a new copy inserting into a new location in the genome (Herron 2004; Kazazian 2004). Retrotransposons' high success made them the major classes of MEs in the primate genomes, constituting on average 45% of the genomes.

1.2 DNA transposons

DNA transposons or class II MEs, were initially known as "jumping genes" because of their ability to move in the host genome (Deininger et al. 2003). By using a transpose encoded by the autonomous copies, DNA transposons can excise themselves out from the original locations as double-stranded DNA and insert into new sites elsewhere in the genome in a "cut-and-paste" style which doesn't result in any direct changes in their copy numbers (Feschotte and Pritham 2007; Pace Ii and Feschotte 2007). Used by the ten out of the total twelve DNA transposon superfamilies, this mechanism is considered as the canonical transposition mechanism for DNA transposons. However, two other superfamilies, Helitron and Mavericks, transpose through non-canonical mechanisms by utilizing a single-stranded DNA as the intermediate, which leads to a "copy-and-paste" style (Feschotte and Pritham 2007; Kapitonov and Jurka 2001; Pritham et al. 2007).

Despite their early success in the primate evolution, the DNA transposons have been considered inactive in the current primate genomes, therefore, they received very little research attention. Lander and colleagues in their initial human genome analysis concluded that there was no evidence for DNA transposon activity during the past 50 million years (My) (Lander et al. 2001), while a later study suggested that DNA transposons had been highly active during the early part of primate evolution till ~37 My ago (Pace Ii and Feschotte 2007).

1.3 Retrotransposons

Depending on the presence or absence of long terminal repeats (LTRs), the retrotransposons can be further divided into LTR retrotransposons and non-LTR retrotransposons, respectively (Cordaux and Batzer 2009; Deininger et al. 2003). In primates, the LTR retrotransposons mainly consist of endogenous retrovirus (ERVs), which are results of infecting virus integrating into the host genomes during different stages of primate evolution (Kazazian 2004). The Short-INterspersed Elements (SINEs), the Long INterspersed Elements (LINEs), and the chimeric elements, SINE-R/VNTR/Alu (SVA), as well as processed pseudogenes, collectively represent the non-LTR retrotransposons in the primate genomes. A canonical non-LTR retrotransposon has a 3' poly (A) tail, and a pair of short repeats at the ends of the insertion sequence called target site duplications (TSDs) (Allet 1979; Grindley 1978). As shown in Fig. 1.2, TSDs are a result and hallmark of the LINE-1 (L1) driven target-primed reverse transcription (TPRT) mechanism (Goodier 2016). The presence of TSDs is a hallmark of all ME transposition with different ME types having a unique TSD characteristics mostly by length.

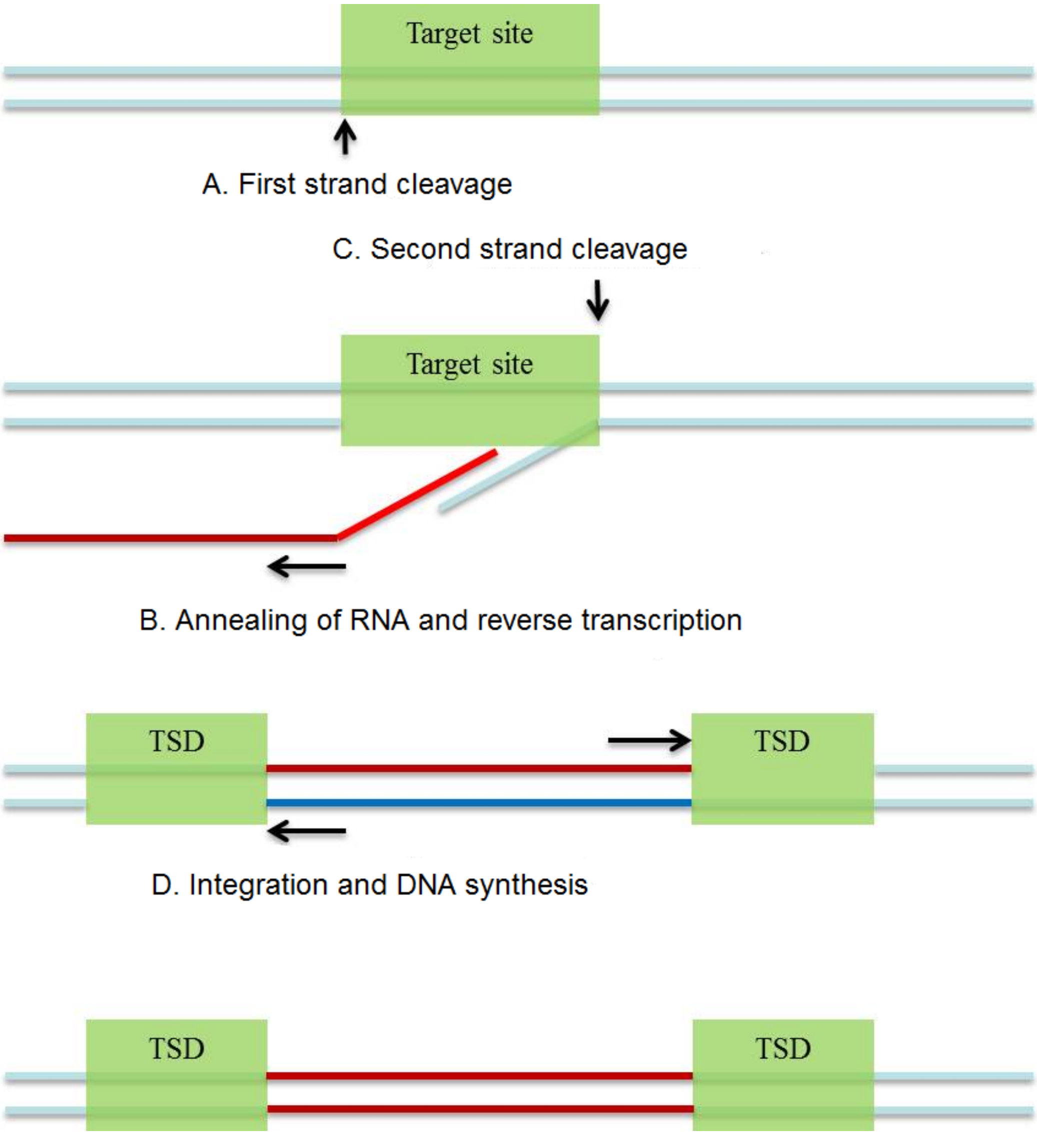

Target site
A. First strand cleavage
C. Second strand cleavage
Target site
B. Annealing of RNA and reverse transcription
TSD
TSD
D. Integration and DNA synthesis
TSD
TSD

Figure 1.2 A schematic diagram of target-primed reverse transcription (TPRT)

This is reprinted from the author's MSc thesis "The identification and characterization of inter- and intra-species genetic diversity derived from retrotransposons in humans, Brock 2012"

A: Cleavage of first DNA strand at the target site by the retrotransposon endonuclease (EN); B: The retrotransposon RNA anneals at the nick site and starts reverse transcription by the retrotransposon reverse transcriptase (RT); C: Cleavage of second DNA strand. D: Integration at the double-strand break and removal of RNA and completion of DNA synthesis, leading to the insertion of a new copy of the retrotransposon at the target site and generation of target site duplications (TSDs).

1.3.1 LINE-1 elements

LINE-1s (L1s), being the only subfamily of autonomous non-LTR retrotransposons in the primate genomes, provide the TPRT machinery for all other non-LTR retrotransposons, which are considered non-autonomous for transposition (Cost and Boeke 1998; Goodier 2016; Jurka 1997; Mita and Boeke 2016; Tang et al. 2018; Xing et al. 2006). A typical autonomous L1 copy, which is ~6,000 bp long, consists of an internal RNA polymerase II promoter, two open reading frames (ORF1 and ORF2) and a polyadenylation signal followed by a polyA tail (Kazazian and Goodier 2002). The ORF1 gene encodes an RNA-binding protein and ORF2 encodes a protein with endonuclease and reverse transcriptase activity (Goodier 2016; Kazazian and Goodier 2002). Several studies have shown that Alus, L1s, and SVAs have an identical core sequence motif of "TT/AAAA" for the insertion sites, confirming that all non-LTR retrotransposition use the same TPRP mechanism (Cost and Boeke 1998; Jurka 1997; Tang et al. 2018; Wang et al. 2006).

1.3.2 Alu elements

Alu elements are a family of primate-specific SINEs, which have contributed to ~11% primate genomes, second only to the ~18% contribution by LINE-1s (L1s). Their success in the primate genomes is even more impressive, considering that the Alu family is one of the shortest ME families, averaging only ~300bp in length. The higher percentage of Alu elements in the primate genomes is primarily due to their extremely high copy numbers, averaging ~1.2 million copies per genome (Ahmed et al. 2013; Deininger 2011). A typical Alu element consists of two diverged dimers, which are believed to have derived from the 7SL RNA gene during a very early stage of primate evolution. The 3' end of Alu usually has a long consecutive "A"s, which is referred to as the poly-A tail. Alu elements carry an internal RNA polymerase II promoter and, therefore, have the ability to express them as RNAs. The expressed Alu transcripts can hijack L1's TPRT machinery for retrotransposition. However, despite both using the same mechanism, there seems to be a difference between L1 retrotransposition and Alu transposition; while L1s depend both ORF1p and ORF2p to retrotranspose, Alus seem to only rely on the presence of ORF2p protein to retrotranspose (Dewannieux et al. 2003; Goodier 2016; Moran et al. 1996; Wallace et al. 2008). According to the data in L1base (Penzkofer et al. 2017) and our recent observation (Nanayakkara et al., manuscript in preparation), the primate genomes usually have only a handful of functional L1s with the ability to code intact ORF1p and ORF2p proteins. Meanwhile, there are more L1s with intact ORF2p protein-coding capacity but have lost the capability to encode intact ORF1p protein for being subject to a higher level of mutations (Goodier 2016; Penzkofer et al. 2017). This may explain the fact that Alus have been able to amplify in most of the primate genomes more efficiently than L1s by having much larger copy numbers (Tang and Liang 2019). In particular, the baboon genome showed an extremely high level of Alu expansion in its recent evolution through a large number of highly active baboon-specific Alu subfamilies (Jordan et al. 2018; Steely et al. 2018; Tang and Liang 2019).

1.4 The impact of MEs

MEs, despite once being mistakenly considered as "junk DNA", have shown the ability to contribute to host genome evolution through a variety of mechanisms. Such mechanisms include, but are not limited to, generation of insertional mutations and causing genomic instability, creation of new genes and splicing isoforms, exon shuffling, and regulation of gene expression (Bourque et al. 2018; Callinan et al. 2005; Chuong et al. 2016; Han et al. 2004; Han et al. 2005; Han et al. 2007; Konkel and Batzer 2010; Mita and Boeke 2016; Quinn and Bubb 2014; Sen et al. 2006; Symer et al. 2002; Szak et al. 2003; Trizzino et al. 2017; Wheelan et al. 2005). MEs are also known to be associated with genetic disorders in human via both germline and somatic insertions, including haemophilia, cystic fibrosis, Apert syndrome, neurofibromatosis, and colon cancers (Anwar et al. 2017; Goodier 2016).

Chapter 2 Comparative genomics analysis reveals high levels of differential retrotransposition among primates from the *Hominidae* and the *Cercopithecidae* families

(The content of this chapter is mostly copied from the published article: “Wanxiangfu Tang, Ping Liang, Comparative genomics analysis reveals high levels of differential retrotransposition among primates from the *Hominidae* and the *Cercopithecidae* families, Genome Biology and Evolution, evz234, https://doi.org/10.1093/gbe/evz234” with some minor changes for table formats and figure reorganization (renumbered after combining with supplementary figures)

The candidate is the main author of this article and was responsible for generating most of the data included in the article. The manuscript was drafted by the candidate and edited by the corresponding author, Dr. Liang, and other collaborative authors to its final form.

In addition to the above publication, part of the results from this work has also be used to generate the following two collaborative publications, which are not included in this thesis:

1. Lee S, Tang W, Liang P, Han K. A comprehensive analysis of chimpanzee (Pan troglodytes)-specific LINE-1 retrotransposons. Gene 693: 46-51, 2019.

2. Lee W, Choi M, Kim S, Tang W, Kim DH, Kim HS, Liang P, Han K. A comprehensive analysis of the Baboon-specific full-length LINE-1 retrotransposons. Genes & Genomics 41:831–837, 2019.)

2.1 Abstract

Mobile elements (MEs), making ~50% of primate genomes, are known to be responsible for generating inter- and intra-species genomic variations and play important roles in genome evolution and gene function. Using a bioinformatics comparative genomics approach, we performed analyses of species-specific MEs (SS-MEs) in eight primate genomes from the families of *Hominidae* and *Cercopithecidae*, focusing on retrotransposons. We identified a total of 230,855 SS-MEs, with which we performed normalization based on evolutionary distances, and we also analyzed the most recent SS-MEs in these genomes. Comparative analysis of SS-MEs reveals striking differences in ME transposition among these primate genomes. Interesting highlights of our results include: 1) the baboon genome has the highest number of SS-MEs with a strong bias for SINEs, while the crab-eating macaque genome has a sustained extremely low transposition for all ME classes, suggesting the existence of a genome-wide mechanism suppressing ME transposition; 2) While SS-SINEs represent the dominant class in general, the orangutan genome stands out by having SS-LINEs as the dominant class; 3) The human genome stands out among the eight genomes by having the largest number of recent highly active ME subfamilies, suggesting a greater impact of ME transposition on its recent evolution; 4) At least 33% of the SS-MEs locate to genic regions, including protein-coding regions, presenting significant potentials for impacting gene function. Our study, as the first of its kind, demonstrates that mobile elements evolve quite differently among these primates, suggesting differential ME transposition as an important mechanism in primate evolution.

2.2 Introduction

Transposable elements or mobile elements ("MEs" hereafter) are defined as genomic DNA sequences, which can change their positions or making copies and inserting into other locations in the genomes. MEs are quite abundant in genomes of higher species such as primates and plants; their contribution to the primate genomes ranges from 46.8% in the green monkey genome to 50.7% in the baboon genome (Carbone et al. 2014; Chimpanzee Sequencing and Analysis 2005; Cordaux and Batzer 2009; Deininger et al. 2003; Lander et al. 2001; Locke et al. 2011; Rhesus Macaque Genome Sequencing and Analysis et al. 2007; Scally et al. 2012; Yan et al. 2011). This percentage is expected to increase slightly in these genomes from further improvements of the genome sequences and repeat annotation, especially for the non-human primate genomes.

There are two major types of MEs, DNA transposons and retrotransposons, by the mechanism of their transposition (Stewart et al. 2011). DNA transposons move in the genome in a "cut and paste" style, for which they were initially called "jumping genes" (Deininger et al. 2003; McClintock 1950). It means that they are able to excise themselves out from their original locations and move to new sites in the genome in the form of DNA, leading to no direct change of their copy numbers in the genome during the process (Pace Ii and Feschotte 2007). DNA transposons constituent approximately 3.6% of the primate genomes. In comparison, retrotransposons mobilize in genomes via an RNA-based duplication process called retrotransposition, in which a retrotransposon is first transcribed into RNA and then reverse transcribed into DNA as a new copy inserting into a new location in the genome (Herron 2004; Kazazian 2004). Therefore, retrotransposons move in the genome through a "copy and paste" style, which leads to a direct increase in their copy numbers. Retrotransposons' high success in the primate genomes made them as the major classes of MEs, constituting on average 45% of the genomes. Depending on the presence or absence of long terminal repeats (LTRs), the retrotransposons can be further divided into LTR retrotransposons and non-LTR retrotransposons, respectively (Cordaux and Batzer 2009; Deininger et al. 2003). In primates, the LTR retrotransposons mainly consist of endogenous retrovirus (ERVs), which are results of endogenous virus integrating into the host genomes during different stages of primate evolution (Kazazian 2004). The Short-INterspersed Elements (SINEs), the Long INterspersed Elements (LINEs), and the chimeric elements, SINE-R/VNTR/Alu (SVA), as well as processed pseudogenes, collectively represent the non-LTR retrotransposons in the primate genomes. A canonical non-LTR retrotransposon has a 3' poly (A) tail and a pair of short repeats at the ends of the insertion sequence called target site duplications (TSDs) (Allet 1979; Grindley 1978). TSDs are a result and hallmark of the L1 driven target-primed reverse transcription (TPRT) mechanism (Goodier 2016).

Despite once being considered "junk DNA", researchers have obtained ample evidence, mostly during the last two decades, that MEs make significant contributions to genome evolution and they can impact gene function via a variety of mechanisms. These mechanisms include, but are not limited to, generation of insertional mutations and causing genomic instability, creation of new genes and splicing isoforms, exon shuffling, and regulation of gene expression (Bourque et al. 2018; Callinan et al. 2005; Chuong et al. 2016; Han et al. 2004; Han et al. 2005; Han et al. 2007; Konkel and Batzer 2010; Mita and Boeke 2016; Quinn and Bubb 2014; Sen et al. 2006; Symer et al. 2002; Szak et al. 2003; Trizzino et al. 2017; Wheelan et al. 2005). MEs also contribute to genetic diseases in human via both germline and somatic insertions (Anwar et al. 2017; Goodier 2016) .

Furthermore, MEs have intimate associations with other repetitive elements such as microsatellite repeats and tandem repeats in plants (Ramsay et al. 1999) or may have involved in the genesis of these repetitive elements (Wilder and Hollocher 2001). It was shown more recently that MEs contribute to at least 23% of all minisatellites and satellites in the human genome (Ahmed and Liang 2012).

MEs have been accumulating along with primate evolution. Although the majority of MEs are “fixed” in the primate genomes meaning they are shared by all primate genomes, certain MEs are uniquely owned by a particular species or lineage. A recent study has suggested that regulatory regions derived from primate and human lineage-specific MEs can be transcriptionally activated in a heterologous regulatory environment to alter histone modifications and DNA methylation, as well as expression of nearby genes in both germline and somatic cells (Ward et al. 2013). This observation suggests that lineage- and species-specific MEs can provide novel regulatory sites in the genome, which can potentially regulate nearby genes’ expression, and ultimately lead to in lineage- and species-specific phenotypic differences. For example, it was recently shown that lineage-specific ERV elements in the primate genomes can act as IFN-inducible enhancers in mammalian immune defenses (Chuong et al. 2016).

Past and ongoing studies on MEs in primate genomes have been mainly focused on the human genome, examining mostly the youngest and active members that contribute to genetic variations among individuals (Battilana et al. 2006; Ewing and Kazazian 2011; Jha et al. 2009; Ray et al. 2005; Seleme et al. 2006; Stewart et al. 2011; Wang et al. 2006). For example, studies have shown that certain members from *L1*, *Alu*, *SVA*, and *HERV* families are still active in the human genome and they are responsible for generating population-specific or polymorphic MEs (Ahmed et al. 2013; Beck et al. 2010; Benit et al. 2003; Mills et al. 2007; Thomas et al. 2018;

Wang et al. 2005). Besides these, limited analyses of species-specific mobile elements have also been performed in a few primate genomes. The first of such study was done by Mills and colleagues, who analyzed species-specific MEs in both the human and chimpanzee genomes based on earlier versions of the genomic sequences (GRCh35/hg17 and CGSC1.1/panTrol1.1), which led to the identification of a total of 7,786 and 2,933 MEs that are uniquely owned by human and chimpanzee, respectively (Mills et al. 2006). However, these early studies of species-specific MEs were limited by the low quality of available genome sequences and unavailability of other primate genome sequences. Recently, we have provided a comprehensive compilation of MEs that are uniquely present in the human genome by making use of the most recent genome sequences for human and many other closely related primates and a robust multi-way comparative genomic approach, leading to the identification of 14,870 human-specific MEs, which contribute to 14.2 Mbp net genome sequence increase (Tang et al. 2018). Other studies focused on species-specific MEs target on either one particular ME type and/or a few primate genomes. For example, Navarro & Galante performed comparative analysis of retrogenes (processed pseudogenes) in seven primate genomes (Navarro and Galante 2015), while Steely et al., recently ascertained 28,114 baboon-specific Alu elements by comparing the genomic sequences of baboon to both rhesus macaque and human genome (Steely et al. 2018).

Despite these many small-scale studies, a large-scale systematic comparative analysis of ME transposition among primates is still lacking. In this study, we adopted our robust multi-way comparative genomic approach used for identifying human-specific MEs to analyze species-specific MEs in eight primate genomes, representing the *Hominidae* family and the *Cercopithecidae* family of the primates. Our analysis identified a total of 230,855 species-specific MEs (SS-MEs) in these genomes, which collectively contribute to ~82 Mbp genome sequences, revealing significant differential ME transposition among primate species.

2.3 Material and methods

2.3.1 Sources of primate genome sequences

For our study, we chose to include four members from each of the *Hominidae* and *Cercopithecidae* primate families. All genome sequences in fasta format and the corresponding RepeatMasker annotation files were downloaded from the UCSC genomic website (http://genome.ucsc.edu) onto our local servers for in-house analysis. In all cases except for gorilla, the most recent genome versions available on the UCSC genome browser site at the time of the study were used. The four *Hominidae* genomes include the human genome (GRCh38/UCSC hg38), chimpanzee genome (May 2016, CSAC Pan_troglodytes-3.0/panTro5), gorilla genome (Dec 2014, NCBI project 31265/gorGor4.1), and orangutan genome (Jul. 2007, WUSTL version Pongo_albelii-2.0.2/ponAbe2). For the gorilla genome, there is a newer version (Mar. 2016, GSMRT3/gorGor5) available, but not assigned into chromosomes, making it difficult to be used for our purpose. The four *Cercopithecidae* genomes include green monkey genome (Mar. 2014 VGC Chlorocebus_sabeus-1.1/chlSab2), crab-eating macaque genome (Jun. 2013 WashU Macaca_fascicularis_5.0/macFas5), rhesus monkey genome (Nov. 2015 BCM Mmul_8.0.1/rheMac8), and baboon (Anubis) genome (Mar. 2012 Baylor Panu_2.0/papAnu2). The information regarding the sequencing platforms and the genome assembly quality is provided in Table 2.1.

Table 2.1 Summary information on sequencing methods, coverage, and assembly quality matrices for the eight primate reference genomes

Genome	UCSC version	Sequencing technology	coverage	# of scaffolds*	Scaffold N50 (bp)	Gap length (Mb)
Human	hg38	BAC/WGS	N/A	456	145138636	174
Chimpanzee	panTro5	Sanger/Illumina/PacBio	6x Sanger/55x Illumina/9x PacBio	44449	135926727	99
Gorilla	gorGor4	capillary sequencing/Illumina	80x	40692	146757320	146
Orangutan	ponAbe2	Illumina	6x	54	135191526	353
Rhesus	rheMac8	Illumina	47.4x	284728	144306982	94
Crab-eating macaque	macFas5	Illumina	68x	7601	152835861	143
Baboon	papAnu2	Sanger/454 FLX/Illumina	2.5x Sanger/4.5x 454/85x Illumina	63250	139646187	55
Green monkey	chlSab2	454 Titanium/Illumina HiSeq/ABI	95x	2004	101219884	37

*, excluding alternative assemblies

2.3.2 Identification of species-specific mobile element sequences (SS-MEs)

We used a computational comparative genomic approach as previously described (Tang et al. 2018) to identify SS-MEs. In this approach, the presence or absence status of a mobile element in the orthologous regions of other genomes is determined by focusing on both whole genome alignment using liftOver and local sequence alignment using BLAT (Hinrichs et al. 2006; Kent 2002).

2.3.2.1 LiftOver overchain file generation

A total of 56 liftOver chain files were needed for comparative analysis of the eight genomes used in this study. These files contain information linking the orthologous positions in a pair of genomes based on lastZ alignment (Harris 2007). Twenty-two of these were available and downloaded from the UCSC genome browser site, while the remaining 34 liftOver chain files, mostly for linking between non-human primate genomes, were generated on a local server using a modified version of UCSC pipeline RunLastzChain (http://genome.ucsc.edu).

2.3.2.2 Pre-processing of MEs

Our starting lists of MEs in each primate genome were those annotated using RepeatMasker. Since RepeatMasker reports fragments of MEs interrupted by other sequences and internal inversions or deletions as individual ME entries, we performed a pre-process to integrate these fragments back to ME sequences representing the original transposition events as previously described (Tang et al. 2018). This step is critical for obtaining more accurate counting of the transposition events, and more importantly for obtaining correct flanking sequences to identify SS-MEs and their TSDs.

2.3.2.3 Identification of SS-MEs

As previously described (Tang et al. 2018), our strategy for identifying SS-MEs is to examine ME insertions and their two flanking regions (after integration) in a genome and compare with the sequences of the corresponding orthologous regions in all genomes with detectable orthologous sequences. If a ME is determined with high confidence that its absence from the orthologous regions of all other genomes is not due to the presence of a gap, then it is considered to be species-specific in this genome. It means that a SS-ME can be identified as one being absent from the orthologous regions in other genomes or from the absence of an orthologous sequence in other genomes (i.e., SS-ME in a species-specific region). Briefly, we used two tools, BLAT and liftOver (http://genomes.ucsc.edu), for determining the orthologous sequences and the species-specific status of MEs using the aforementioned integrated RepeatMasker ME list as input. Only the ME copies that are supported to be unique to a species by both tools were included in the final list of SS-MEs.

2.3.2.4 Normalization of SS-MEs counts

A rooted neighbor-joining (NJ) phylogenetic tree of the eight primate genomes, plus marmoset as an outgroup, was constructed based on the coding sequences (CDS) of the *ACTB* genes using Clustal (Chenna et al. 2003) for multiple sequence alignment and NJ tree generation and displayed using FigTree (https://github.com/rambaut/figtree/). The GenBank accessions for the nine *ACTB* sequences used in the analysis include NM_001101.5 (hs_ACTB/human), NM_001009945.1 (pt_ACTB/chimpanzee), 019030619.1 (gg_ACTB/gorilla), NM_001133354.1 (po_ACTB/orangutan), NM_001285025.1 (mf_ACTB/crab-eating macaque), NM_001033084 (rm_ACTB/Rhesus monkey), XM_003895688.3 (poa_ACTB/baboon), NM_001330273.1 (cs_ACTB/green monkey), and XM_008983711 (cj_ACTB/marmoset). The closest pairwise evolutionary distance for each species among the eight genomes were obtained based on the total branch length between the two closest species provided on the phylogenetic tree. The distance of the genomes with the shortest among the eight genomes is used as the base distance for normalizing the SS-ME counts for all other genomes using a formula of (normalized SS-ME count=raw count x (base distance/genome distance)), where the base distance is always 0.0043 (for distance between rhesus and crab-eating macaque) and the genome distance is the shortest distance of the genome to be normalized. This formula is based on an assumed positive linear relationship between the numbers of SS-MEs and evolutionary distances of the genomes.

2.3.3 Identification of TSDs, transductions, and insertion mediated-deletions (IMD)

The TSDs, as well as transductions and IMDs for all SS-MEs, were identified using in-house Perl scripts as described previously (Tang et al. 2018). For those with TSDs successfully identified, a 30-bp sequence centered at each insertion site in the predicted pre-integration alleles were extracted after removing the ME sequence and one copy of the TSDs from the ME alleles. Entries with identified TSDs and extra sequences between the ME and either copy of the TSDs are considered potential candidates for ME insertion-mediated transductions and were subject to further validation as previously described (Tang et al. 2018). For entries without TSDs, if there are extra sequences at the pre-integration site in the out-group genomes, they were considered candidates for IMDs, which were subject to further validation.

2.3.4 Identification of most recent SS-MEs and survey of age profile for SS-MEs

The raw list of SS-MEs in each genome was used to identify a subset of MEs that represent the most recent ME copies based on sequence divergence level by running an all-against-all sequence alignment among all SS-MEs in a genome using BLAT (minScore ≥100; minIdentity ≥ 97%). Those showing a 100% sequence identity with another copy of SS-ME (non-self-match) are considered as the most recent SS-MEs. For human and chimpanzee genomes, the numbers of SS-MEs were binned by the percentage of sequence similarity for plotting the age profiles for all SS-MEs and each ME class from each genome. The percentage of sequence similarity was calculated using an in-house PERL script based on the blat output considering the gaps and mismatches in the aligned block(s).

2.3.5 Analysis of SS-MEs' association with genes in the primate genomes

We used the genomic coordinates of genes broken down to individual exons based on GENCODE gene annotation (Harrow et al. 2012) and NCBI RefSeq data (Pruitt et al. 2007) for the human genome while only the ENSEMBL gene annotation data (Zerbino et al. 2018) were used for the non-human primate genomes. The sequences of each of genome were divided into a non-redundant list of categorized regions in gene context, including coding sequence (CDS), non-coding RNA, 5'-UTR, 3'-UTR, promoter (1 kb), intron, and intergenic regions using an in-house PERL script as previously described (Tang et al. 2018). This order of genic region categories as listed above was used to set the priority from high to low in handling overlapping regions between splice forms of the same gene or different genes. For example, if a region is a CDS for one transcript/gene and is a UTR or intron for another, then this region would be categorized as CDS.

2.3.6 Computational analyses

Data analysis and figure plotting were performed using a combination of Linux shell scripts, R and Microsoft Excel. Most of the genome sequence analyses were performed on Compute Canada high-performance computing facilities (http://computecanada.ca).

2.4 Results

2.4.1 The overall ME profiles in the eight primate genomes

The initial ME lists used in this study were based on the RepeatMasker annotations obtained from the UCSC Genome Browser, and we performed integration of fragmented MEs to represent original transposition events to improve the accuracy in identifying SS-MEs and the TSDs. As shown in Table 2.2, the consolidation led to an average reduction of 940,000 ME counts per genome. Among the eight genomes after consolidation, the chimpanzee genome has the largest number of MEs (3,609,255) and the green monkey and crab-eating macaque genomes have very similar and the least number of MEs at 3,327,187 and 3,327,372, respectively (Table 2.2). By copy number from low to high among the genomes, SINEs as the most successful MEs have 1,631,6262 copies in crab-eating macaque to 1,706,611 copies in rhesus genome; LINEs have 875,720 copies in crab-eating macaque to 1,000,667 copies in chimpanzee; LTRs have from 460,094 copies in crab-eating macaque to 499,454 copies in chimpanzee; DNA transposons have 359,802 copies in crab-eating macaque to 421,580 copies in chimpanzee; SAVs that are uniquely found in the *Hominidae* group with 2,328 copies in the orangutan to 4,931 and 4,933 copies in chimpanzee and human, respectively (Table 2.2). By the percentage of the genome, LINEs as the most successful contribute to the genome from 20.4% in green monkey to 22.8% in baboon; SINEs contribute from 13.4% in human and gorilla to 14.8% in baboon; SVAs, as the youngest ME class, contribute ~0.1% in all hominid genomes; very small numbers of macSVA are found in the monkey genomes, which seem to have a separate origin from the hominid SVAs and they were excluded from further analysis; LTRs contribute from 8.9% in crab-eating macaque to 9.5% in baboon; DNA transposons contribute from 3.4% in orangutan to 3.8% in gorilla (Table 2.2). Collectively, MEs from these five major classes constitute from 46.8% (green monkey) to 50.7% (baboon) to the genomes (Table 2.2). All retrotransposons together contribute from 43.3% in the green monkey genome to 47.1% in the baboon genome (Table 2.2). DNA transposons were excluded from further analyses in this study due to their smaller percentages and the very low activity levels in these genomes.

Table 2.2 The copy numbers and sizes of mobile elements (MEs) in eight primate genomes.

Reference version	ME type	DNA	LINE	SINE	SVA	LTR	Total
NCBI 38/UCSC hg38 (December 2013) (non-gap size: 2,937,641,526)	raw counts	483,994	1,516,226	1,779,233	5,397	720,177	**4,505,027**
	integrated counts	399,590	969,873	1,689,416	4,933	496,946	**3,560,758**
	total size	102,664,356	643,469,259	394,684,907	4,228,693	267,988,862	**1,413,036,077**
	% genome	3.5	21.9	13.4	0.1	9.1	**48.1**
Chimpanzee/UCSC panTro5 (May 2016) (non-gap size: 2,870,696,247)	raw counts	510,250	1,551,601	1,771,039	5,358	723,412	**4,561,660**
	integrated counts	421,580	1,000,667	1,682,623	4,931	499,454	**3,609,255**
	total size	107,832,154	641,198,795	391,733,671	4,294,837	267,300,551	**1,412,360,008**
	% genome	3.8	22.3	13.6	0.1	9.3	**49.2**
Gorilla/UCSC gorGor4 (Dec 2014) (non-gap size: 2,790,653,262)	raw counts	503,480	1,533,883	1,722,434	5,492	707,051	**4,472,340**
	integrated counts	418,454	1,000,110	1,638,587	4,809	494,156	**3,556,116**
	total size	106,573,049	611,178,732	373,516,073	2,632,794	257,170,637	**1,351,071,285**
	% genome	3.8	21.9	13.4	0.1	9.2	**48.4**
Orangutan/UCSC ponAbe2 (July 2007) (non-gap size: 2,725,322,026)	raw counts	429,467	1,428,157	1,689,629	2,771	671,620	**4,221,644**
	integrated counts	347,471	907,077	1,602,634	2,328	470,734	**3,330,244**
	total size	93,420,030	607,029,348	364,089,696	2,707,548	244,033,406	**1,311,280,028**
	% genome	3.4	22.3	13.4	0.1	9.0	**48.1**
Rhesus/UCSC rheMac8 (Nov 2015) (non-gap size: 2,763,835,834)	raw counts	486,991	1,477,648	1,796,021	152	695,510	**4,456,322**
	integrated counts	401,546	948,851	1,706,611	137	480,535	**3,537,680**
	total size	102,546,356	600,701,619	399,175,118	53,160	251,233,008	**1,353,709,261**
	% genome	3.7	21.7	14.4	0.0	9.1	**49.0**
Crab-eating macaque/UCSC macFas5 Jun 2013)	raw counts	443,909	1,414,592	1,721,680	145	664,942	**4,245,268**
	integrated counts	359,802	875,720	1,631,626	130	460,094	**3,327,372**

(non-gap size: 2,734,297,941)	total size	94,910,440	579,886,936	382,173,139	50,390	244,265,185	**1,301,286,090**
	% genome	3.5	21.2	14.0	0.0	8.9	**47.6**
Baboon/UCSC papAnu2 (Mar 2012) (non-gap size: 2,682,265,895)	raw counts	459,662	1,471,152	1,801,595	169	701,611	**4,434,189**
	integrated counts	369,684	899,503	1,648,361	141	467,533	**3,385,222**
	total size	97,943,467	610,860,150	397,160,589	53,859	255,227,094	**1,361,245,159**
	% genome	3.7	22.8	14.8	0.0	9.5	**50.7**
Green monkey/UCSC chlSab2 (Mar 2014) (non-gap size: 2,708,021,715)	raw counts	445,724	1,426,343	1,709,337	145	676,130	**4,257,679**
	integrated counts	361,048	886,492	1,616,578	133	462,936	**3,327,187**
	total size	95,097,218	553,320,897	373,621,718	48,192	244,275,512	**1,266,363,537**
	% genome	3.5	20.4	13.8	0	9	**46.8**

2.4.2 Differential level of species-specific MEs (SS-MEs) in primate genomes

To assess the detailed differential ME transposition among the primate genomes, we first examined SS-MEs that are defined as being uniquely present in only one of the examined genomes. Our analysis of SS-MEs was based on the consolidated ME lists as discussed in the previous section, and it was performed using a multi-way comparative genomics approach extended from our previously described method in identifying human-specific MEs (Tang et al. 2018). By comparing each of the eight genomes to the seven other genomes, we identified a total of 228,450 SS-MEs, consisting of 150,260 SINEs, 61,216 LINEs, 5,230 SVAs, and 11,744 LTRs (Table 2.3). The list of SS-MEs for the human genome is the same as what was in our previous work (Tang et al. 2018) and is provided here for comparative analysis.

Table 2.3 The total and average numbers of normalized species-specific mobile elements (SS-MEs) and most recent SS-MEs for all eight genomes and for two families of primates

ME class	SS-ME type	SINE	LINE	SVA /macSVA	LTR	Total
Hominidae total	raw SS-MEs	35,410	37,030	5,225	6,076	83,741
	adjusted SS-MEs	21,740	17,073	3,613	3,140	45,566
	most recent SS_MEs	9,361	20,245	1,861	539	32,006
Cercopithecidae total	raw SS-MEs	114,850	24,186	5	5,668	144,709
	adjusted SS-MEs	94,810	17,437	3	4,356	117,111
	most recent SS_MEs	37,713	13,932	0	623	52,268
All genome total	raw SS-MEs	150,260	61,216	5,230	11,744	228,450
	adjusted SS-MEs	116,550	34,510	3,617	7,496	162,172
	most recent SS_MEs	47,074	34,177	1,861	1,162	84,274
Hominidae average	raw SS-MEs	8,853	9,258	1,306	1,519	20,935
	adjusted SS-MEs	5,435	5,757	839	785	12,816
	most recent SS_MEs	2,340	5,061	465	135	8,002
Cercopithecidae average	raw SS-MEs	28,713	6,047	1	1,417	36,177
	adjusted SS-MEs	23,702	4,485	1	1,089	29,278
	most recent SS_MEs	9,428	3,483	-	156	13,067
All genome average	raw SS-MEs	18,783	7,652	654	1,468	28,556
	adjusted SS-MEs	14,569	4,314	452	937	20,271
	most recent SS_MEs	5,884	4,272	233	145	10,534

As seen in Table 2.4 and Fig. 2.1A, the total numbers of SS-MEs are drastically different across the eight primate genomes with the baboon genome having the largest number (66,418), which is more than 20 times higher than that of the crab-eating macaque genome with the smallest number of SS-MEs (3,273). Certainly, these differences in the raw list of SS-MEs are directly tied to the different evolutionary distances among the genomes, making these numbers not suitable to represent the relative retrotransposition level in these genomes. However, the extremely low level of SS-MEs in the crab-eating macaque genome seems to be striking by being merely 1/8 of the SS-MEs in the rhesus genome, which is the mutually closest genome, making the two numbers directly comparable to each other (3,273 vs. 26,433). Similarly, the differences between the human and chimpanzee genomes are also substantial by the total number of SS-MEs (14,891 vs. 21,421) or by specific ME types. For example, the chimpanzee has almost 4 times more SS-LTRs than human (1,924 vs. 530) and two times of SS-LINEs (7,288 vs 3,946), while the numbers of SS-SVAs are more or less similar (1,597 vs. 1,571) (Table 2.4).

Table 2.4 Species-specific mobile elements (SS-MEs) and most recent SS-MEs in eight primate genomes

Genome	ME class	SINE	LINE	SVA	LTR	Total
Human	raw	8,844	3,946	1,571	530	14,891
	normalized	7,175	3,201	1,275	430	12,081
	MR SS-MEs	4,775	2,736	658	110	8,279
Chimpanzee	raw	10,612	7,288	1,597	1,924	21,421
	normalized	8,610	5,913	1,296	1,561	17,379
	MR SS-MEs	2,309	3,595	564	175	6,643
Gorilla	raw	6,324	4,085	877	689	11,975
	normalized	3,399	2,197*	471	370	6,437
	MR SS-MEs	2,105	2,197	397	147	4,846
Orangutan	raw	9,630	21,711	1,180	2,933	35,454
	normalized	2,556	11,717*	313	779	15,365
	MR SS-MEs	172	11,717	242	107	12,238
Rhesus	raw	22,069	3,016	2	1,346	26,433
	normalized	22,069	3,016	2	1,346	26,433
	MR SS-MEs	4,083	1,217	0	107	5,407
Crab-eating macaque	raw	2,257	782	0	234	3,273
	normalized	2,257	782	0	234	3,273
	MR SS-MEs	416	411	0	50	877
Baboon	raw	56,247	8,407	0	1,764	66,418
	normalized	54,969	8,216	0	1,724	64,909
	MR SS-MEs	25292	6376	0	268	31,936
Green monkey	raw	34,277	11,981	3	2,324	48,585
	normalized	15,515	5,928*	1	1,052	22,496
	MR SS-MEs	7,922	5,928	0	198	14,048

*: Normalized numbers were lower but manually adjusted to be the same as the most recent SS-MEs

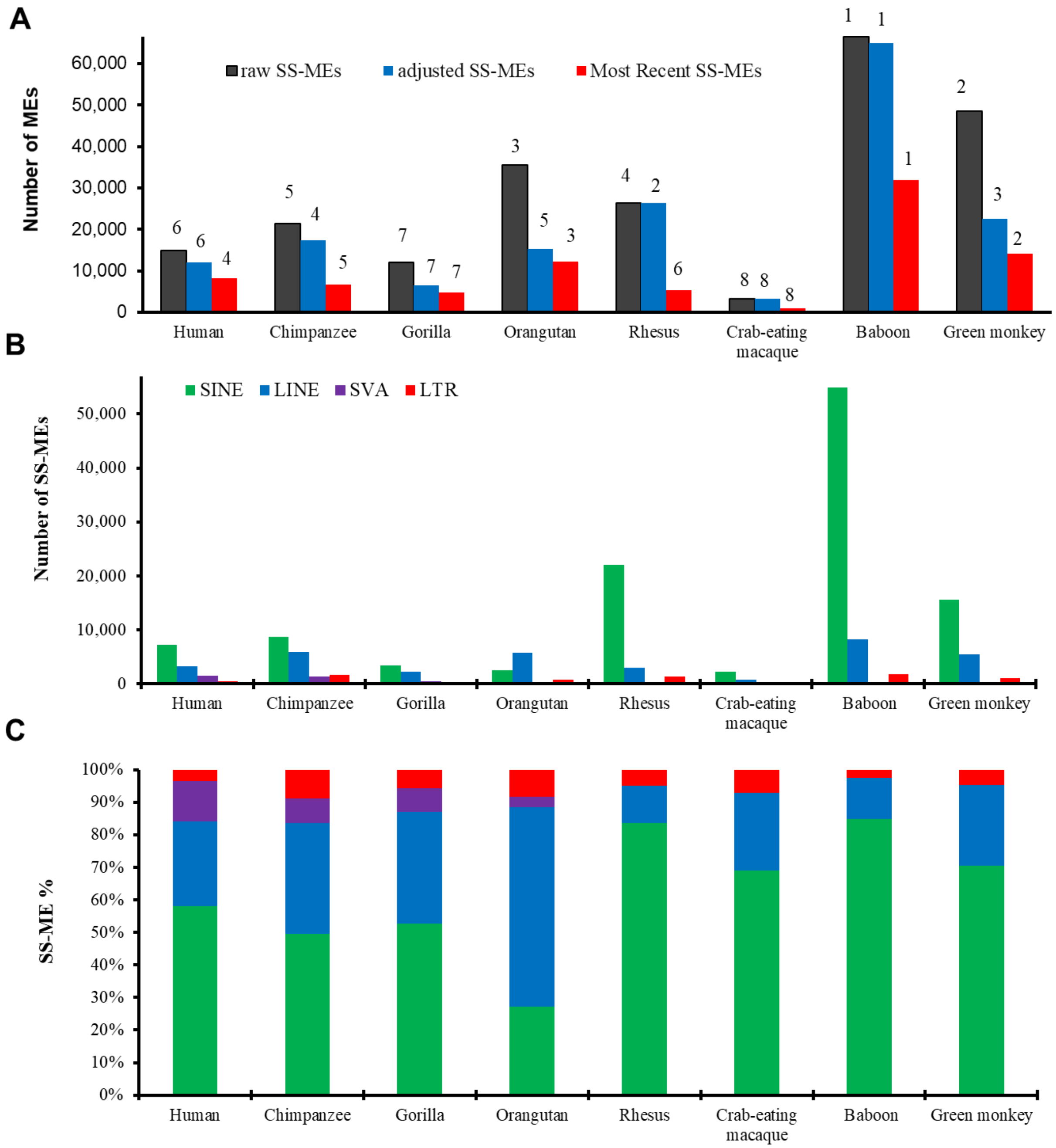
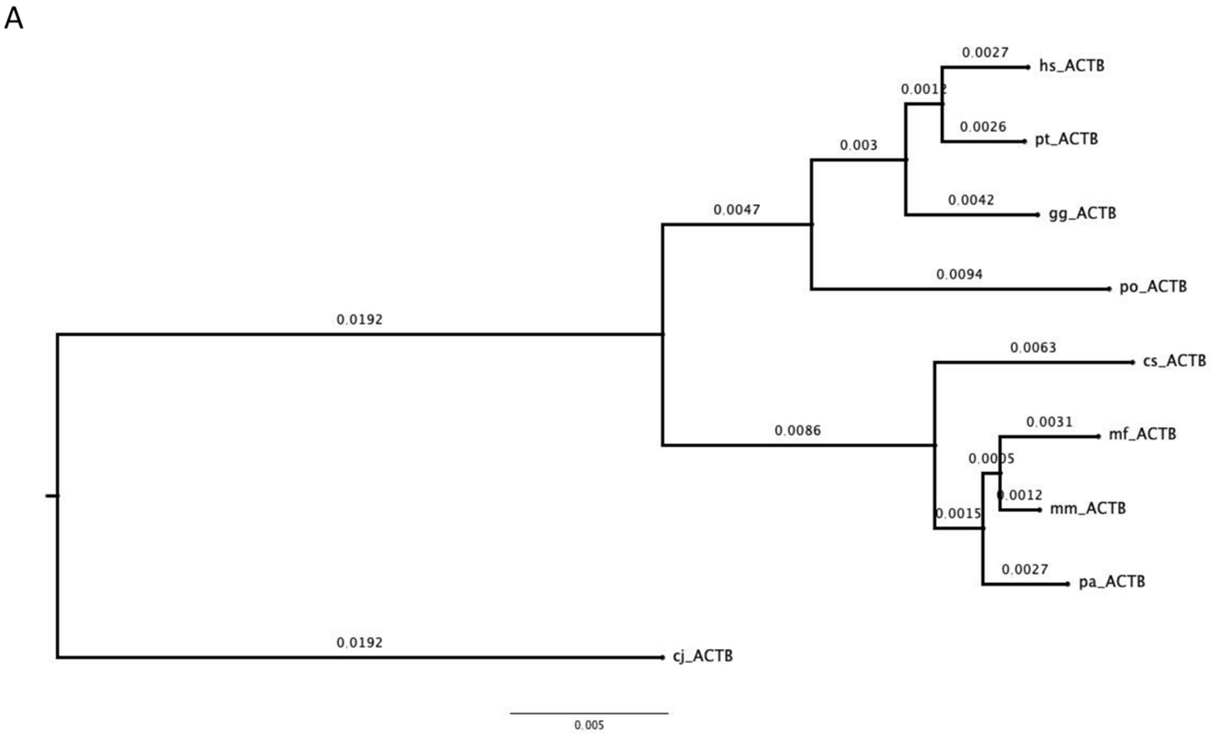


Figure 2.1 Comparisons of the species-specific mobile element (SS-MEs) across eight primate genomes

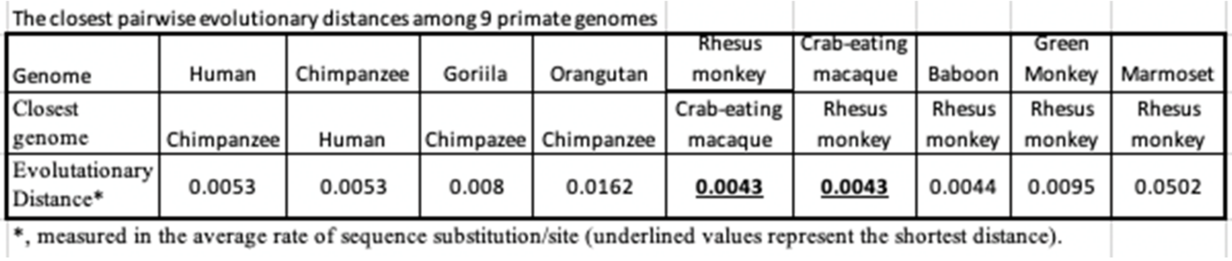
A. Bar plots showing the total numbers of raw, normalized, and most recent SS-MEs in each genome. The numbers at the top of the bars represent the ranking among the eight genomes with 1 being the highest and 8 being the lowest for the total numbers of MEs in the corresponding ME category. B. Bar plots showing the normalized numbers of SS-MEs for each ME class in each genome; C. Stacked bar plots showing the percentage of normalized SS-MEs by ME class in each genome. The color scheme for C is the same as in B.

It is worth mentioning here that while a few factors associated with the variable quality of the genome assemblies and ME annotation, etc. may have some impact on the numbers of SS-MEs as further discussed later, they do not seem to be the main contributor to the large degrees of the SS-ME differences among the genomes based on several lines of evidence. First, the quality of the genome assemblies as measured by scaffold N50 is variable but comparable (within 30%, Table 2.1), and as one would expect, the total numbers of MEs (after integration) in these genomes are quite similar to each other (Table 2.2) with variation below 7% (data not shown), further confirming the qualities of genome assemblies and ME annotation are comparable across these genomes. Second, there is a lack of correlation between the scaffold N50 and the total number of SS-MEs. For example, the green monkey genome has the lowest scaffold N50, but has the 3rd largest number of SS-MEs, while the crab-eating macaque genome with the highest scaffold N50 has a dramatically low number of SS-MEs (Table 2.4, Fig. 2.1A). Therefore, we are confident that the differences of SS-MEs we observed are mainly a result of differential ME transposition in these genomes rather than as artifacts from variations of genome assembly and ME annotation quality.

Since the numbers of SS-MEs identified using our method are expected to be directly impacted by the evolutionary distance among the species involved in the analysis, meaning that in general the larger the evolutionary distance of a genome from the rest genomes is, the more SS-MEs are expected to be identified, we performed normalization to these numbers to make them more comparable. It was done by adjusting the numbers of SS-MEs of a genome based on its shortest pairwise evolutionary distance from the seven other genomes calculated based on a phylogenetic tree constructed using the beta actin (*ACTB*) coding sequences (CDS) collected from NCBI (Fig. 2.2). As shown in Table 2.4, after normalization, the numbers of SS-MEs decreased for all genomes except for the two macaque genomes, which have the closest mutual distance among all eight genomes and were used as the baseline for normalization. While the overall pattern of ranking based on the total numbers of normalized SS-MEs is largely the same as for raw SS-MEs, the orangutan and rhesus genomes had the largest changes in ranking based on normalized SS-MEs with the former dropped from 3rd to 5th due to its largest distance from the other genomes and the latter moved up by two from the 4th to the 2nd due its shortest evolutionary distance, while the chimpanzee genome moved up by 1 position (Fig. 2.1A). The rest four genomes remained their ranking same as for raw SS-MEs, and more specifically, the baboon, crab-eating macaque, and gorilla genomes remain as the one with the largest, the least, and 2nd least number of SS-MEs, respectively, while the human genome remains as the 6th. Further analyses from this point on were based on normalized SS-MEs unless otherwise specified.



B



The closest pairwise evolutionary distances among 9 primate genomes

Genome	Human	Chimpanzee	Goriila	Orangutan	Rhesus monkey	Crab-eating macaque	Baboon	Green Monkey	Marmoset
Closest genome	Chimpanzee	Human	Chimpanzee	Chimpanzee	Crab-eating macaque	Rhesus monkey	Rhesus monkey	Rhesus monkey	Rhesus monkey
Evolutationary Distance*	0.0053	0.0053	0.008	0.0162	**0.0043**	**0.0043**	0.0044	0.0095	0.0502

*, measured in the average rate of sequence substitution/site (underlined values represent the shortest distance).

Figure 2.2 Calculation of the shortest pairwise phylogenetic distances among the eight genomes

A. A rooted neighbor-joining phylogenetic tree of the eight primate genomes, plus marmoset as an outgroup, constructed based on the coding sequences (CDS) of the ACTB genes. The GenBank accessions for the 9 sequences used in the tree are NM_001101.5 (hs_ACTB/human), NM_001009945.1 (pt_ACTB/chimpanzee), XM_019030619.1 (gg_ACTB/gorilla), NM_001133354.1 (po_ACTB/orangutan), NM_001285025.1 (mf_ACTB/crab-eating macaque), NM_001033084 (mm_ACTB/Rhesus monkey), XM_003895688.3 (pa_ACTB/baboon), NM_001330273.1 (cs_ACTB/green monkey), and XM_008983711 (cj_ACTB/marmoset). The numeric values on the branches represent the relative evolutionary distance as the average rate of sequence substitutions per site. B. The table listing the closest species for each genome within the eight genomes based on the shortest distance on the phylogenetic tree and the specific values for the minimal distances in A, which were used as the basis to obtain the normalized species-specific mobile elements (SS-MEs) in Table 2.4.

Based on the normalized SS-MEs, we examined differential ME transposition among these genomes in details. First, we compared the composition of SS-MEs by ME class across genomes. Overall, SS-SINEs represent the largest class of SS-MEs in all genomes except for the orangutan genome. In the *Hominidae* genomes, the numbers of SS-SINEs are larger than the numbers of SS-LINEs for three of the four genomes. This difference is much larger in the *Cercopithecidae* genomes, especially in the baboon genome, which has 54,969 SS-SINEs constituting ~85% of all SS-MEs in the genome and being more than two times higher than the 2nd highest genome (rhesus, 22,069) and more than 3 times higher than all genome average (14,569) (Table 2.3 & 2.4, Fig. 2.1B & C). This observation is in good agreement with the results of two very recent studies reporting dramatically elevated recent Alu insertions in the baboon genome due to a larger number of baboon-specific Alu subfamilies (Rogers et al. 2019; Steely et al. 2018). The orangutan genome is also very unique in SS-ME composition by being the only genome having a larger number of SS-LINEs than that of SS-SINEs in the same genome (11,717 vs 2,556) (Table 2.4, Fig. 2.1A). In contrast, the number of SS-SINEs in orangutan is significantly lower than that of all other genomes (2,556 vs. ≥3,399) except for crab-eating genome, which has the lowest number of SS-SINEs (2,257). For SS-LTRs, the crab-eating macaque genome has the least number (234), while the baboon genome has the largest number (1,724), followed by chimpanzee (1,561), rhesus (1,346), green monkey (1,052), orangutan (779), human (430), and gorilla genome (370). For SS-SVAs, the human genome had the largest number (1,533), followed by chimpanzee (1,296), gorilla (471) and orangutan (313) seemly in negative correlation with the evolutionary ages. While between 100 and 200 MacSVAs are present in the *Cercopithecidae* genomes, no more than 3 or zero SS-MacSVAs are detected (Table 2.2 & 2.4), and thus they were excluded from further analysis.

It is worth noting that for all genomes except for crab-eating macaque have one or more ME class being very successful (e.g. baboon for SINE and LTR, orangutan for LINE, and human for SVA) or moderately successful (e.g. rhesus and green monkey genomes for SINE and chimpanzee for LTR), the extreme low number of SS-MEs applies to all ME classes in the crab-eating macaque genome (Table 2.4, fig 2B). This strongly suggests the existence of a universal molecular mechanism, which suppresses the activity of all ME classes in this genome.

Between the two primate families, there also seem to have some differences in their SS-ME profiles with the *Cercopithecidae* family having more than 4 times of SS-SINEs than the *Hominidae* family (23,702/genome vs. 5,435/genome), but with a lower number of SS-LINEs (4,485/genome) than the *Hominidae* family (5,757/genome), leading to an overall higher level of SS-MEs than the latter (29,278/genome vs. 12,816/genome) (Table 2.3). The slightly higher level of SS-LTRs in the *Cercopithecidae* family (1,089 vs. 785 for *Hominidae*) also contribute to these differences. Interestingly, while the level of ME accumulation seems to be more or less similar (within 1 order of differences) among the *Hominidae* genomes, it differs dramatically (more than 1 order) among the members of the *Cercopithecidae* family by having members with both the lowest and highest number of SS-MEs among the eight genomes (Table 2.4, Fig. 2.1A).

Besides comparison of SS-MEs by the numbers, we also compared the composition of SS-MEs by the percentages of ME class across the genomes. As shown in Fig. 2.1C, the uniqueness of the SS-ME composition for each of the genomes is very evident with no two genomes being identical. The orangutan genome stands out by having an extremely large portion of SS-LINEs and a very small portion of SS-SINEs. The ME composition is more similar among the *Cercopithecidae* genomes despite the huge differences by the number of SS-MEs as seen in Fig. 2.1B.

2.4.3 Differential level of the most recent SS-MEs in primate genomes

In addition to normalizing the SS-MEs by the evolutionary distances of the species, we also collected a subset of SS-MEs as most recent SS-MEs, which were involved as either as the parent or daughter copies in most recent transposition events. They are identified as SS-MEs sharing 100% sequence similarity (≥100 bp of the ME sequence) with another SS-ME copy in the same genome not associated with segmental duplication. By requiring 100% sequence similarity, we are focusing on the SS-MEs resulted from the narrowest window (compared to if a lower stringency, e.g. 98% sequence similarity, was used) of species evolution towards the current genomes, making it sufficiently distinct from the entire period of species evolution as reflected by the normalized SS-MEs. Since the same criteria were applied to all genomes, the numbers of these most recent SS-MEs can be used to measure and compare the more recent and current ME transposition activity across genomes without being biased by variable species evolutionary distances. Certainly, this method can also be subject to biases from variable mutation rate across the species. It is also worth to point out that many MEs outside of SS-MEs were found to have 100% sequence similarity with another ME copy in the same genome, seemly most due to segmental duplication and more recent MEs that are shared between closely related species (data not shown). Even though these non-SS-MEs may represent products of ME transposition events very close to the separation of the species from their perspective closest relatives among the eight genomes, they are not the targets for our study for not being SS-MEs.

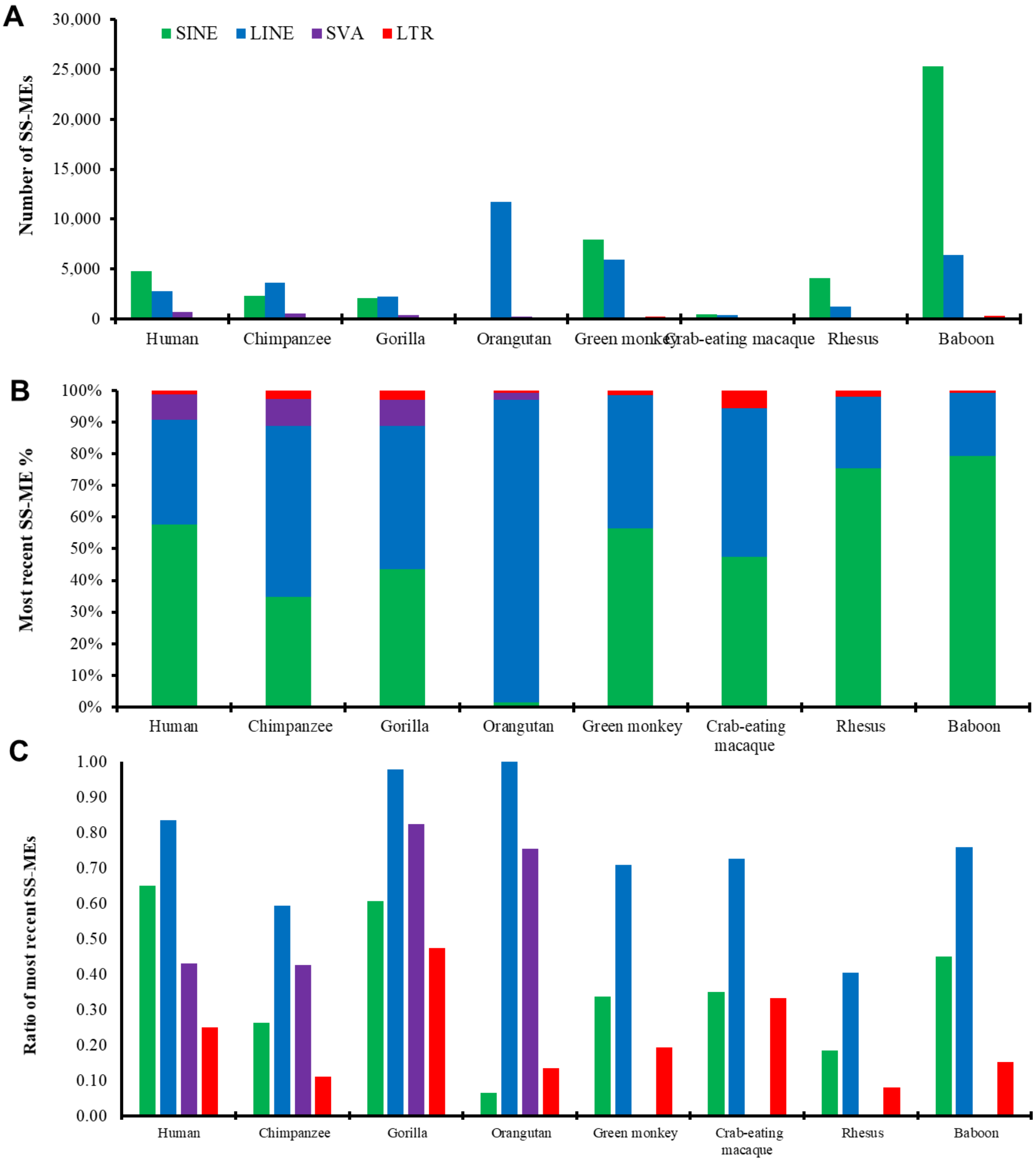


Figure 2.3 The compositions of the most recent species-specific mobile elements (SS-MEs) by ME class in the eight primate genomes

A. The number of the most recent SS-MEs for each ME class in each genome; B. The percentage of most recent SS-MEs by ME class in each genome. C. The ratio of most recent SS-MEs to the normalized SS-MEs by ME class based on copy number. The color scheme is the same for all panels.

The overall trend for the total number of most recent SS-MEs among the genomes is similar to that of normalized SS-MEs (Table 2.4, Fig. 2.1A). Like for the raw and normalized SS-MEs, the baboon genome keeps its 1st position as having the highest number of most recent SS-MEs (31,936), while the crab-eating macaque genome has the lowest number (877), and gorilla genome has the 2nd least number (4,846), making the ranking of these three genomes being the same by all three sets of SS-ME numbers (Table 2.4, Fig. 2.1A). The overall patterns of the most recent SS-ME profiles by ME class in number and percentage are also more or less similar to these of the normalized SS-MEs (Fig. 2.3A vs. Fig. 2.1B for numbers and Fig. 2.3B vs. Fig. 2.1C for percentage). The fact that the crab-eating macaque genome has the lowest number of most recent SS-MEs as in the case of SS-MEs (Fig. 2.1A) indicates a sustained extremely low level of ME transposition activity in this genome. Further, the fact that the composition of the most recent SS-MEs by ME class in this genome is similar to the other monkey genomes (Fig. 2.3B) as in the case of SS-MEs (Fig. 2.1C) indicates that the suppression of transposition applies to all ME classes examined in the crab-eating macaque genome.

Despite the similarity in the overall trend between the most recent SS-MEs and normalized SS-MEs, a few interesting differences were also observed. In striking contrast with the crab-eating macaque, the baboon genome seems to maintain a sustained high level of ME transposition activity leading to the largest number of SS-MEs and most recent SS-MEs both with a strong bias for SS-SINEs (Fig. 2.1A, Fig. 2.3A and Fig. 2.1B). The rhesus genome had the largest drop in ranking from the 2nd for normalized SS-MEs to the 6th position by the number of most recent SS-MEs, while the orangutan and human genomes had the largest increase from the 5th to the 3rd and from the 6th to the 4th, respectively. It is also worth noting that between human and chimpanzee, which are mutually the closest among the eight genomes, the ranking moved up

2 positions for human, but moved down 1 position for chimpanzee. While the chimpanzee genome has a much larger number of SS-MEs than human genome (17,379 vs. 12,081), the situation is opposite for the most recent SS-MEs with human having a much larger number of the most recent SS-SINEs than chimpanzee (8,279 vs. 6,643) (Table 2.4, Fig. 2.1A). Another interesting difference is the much stronger dominance of LINEs in the most recent SS-MEs (~99%) (Fig. 2.3B) than in the SS-MEs (~85%) (Fig. 2.1C) in the orangutan genome. By number, orangutan genome has the largest number of most recent SS-LINEs, being more than two times higher than the genome averages (11,7171 vs. 4,722) (Table 2.3 & 2.4). In contrast to the most recent SS-LINEs, the most recent SS-SINEs in the orangutan genome is extremely low, lower even than that in the crab-eating macaque genome (172 vs 416) (Table 2.4). These data indicate that the ME transposition profile in most recent genomes has changed from the less recent period, revealing a temporal difference in ME transposition in these genomes.

We also examined the ratios of the most recent SS-MEs in the SS-MEs (normalized) and compared across the genomes by ME class as a way to assess the relative very recent ME transposition activity across the genomes. As seen in Fig. 2.3C, each genome has its unique ratio profile by ME class although the overall pattern is more or less similar among the genomes excluding the differences for SVA between the two primate families. Among the ME classes, LINE showed a more consistent pattern by having the highest ratio among all ME classes in each genome. This is also true in the baboon genome, despite SINE being much more successful than LINE in this genome by copy number (Fig. 2.3A). As a matter of fact, for the genomes of gorilla, orangutan, and green monkey, the numbers of most recent SS-LINEs are higher than the normalized SS-LINEs, a situation not seen for any other ME type (Table 2.4). These results indicate that the high success of SINEs and other non-LTR retrotransposons always requires the support of LINE activity, or L1 to be more specific.

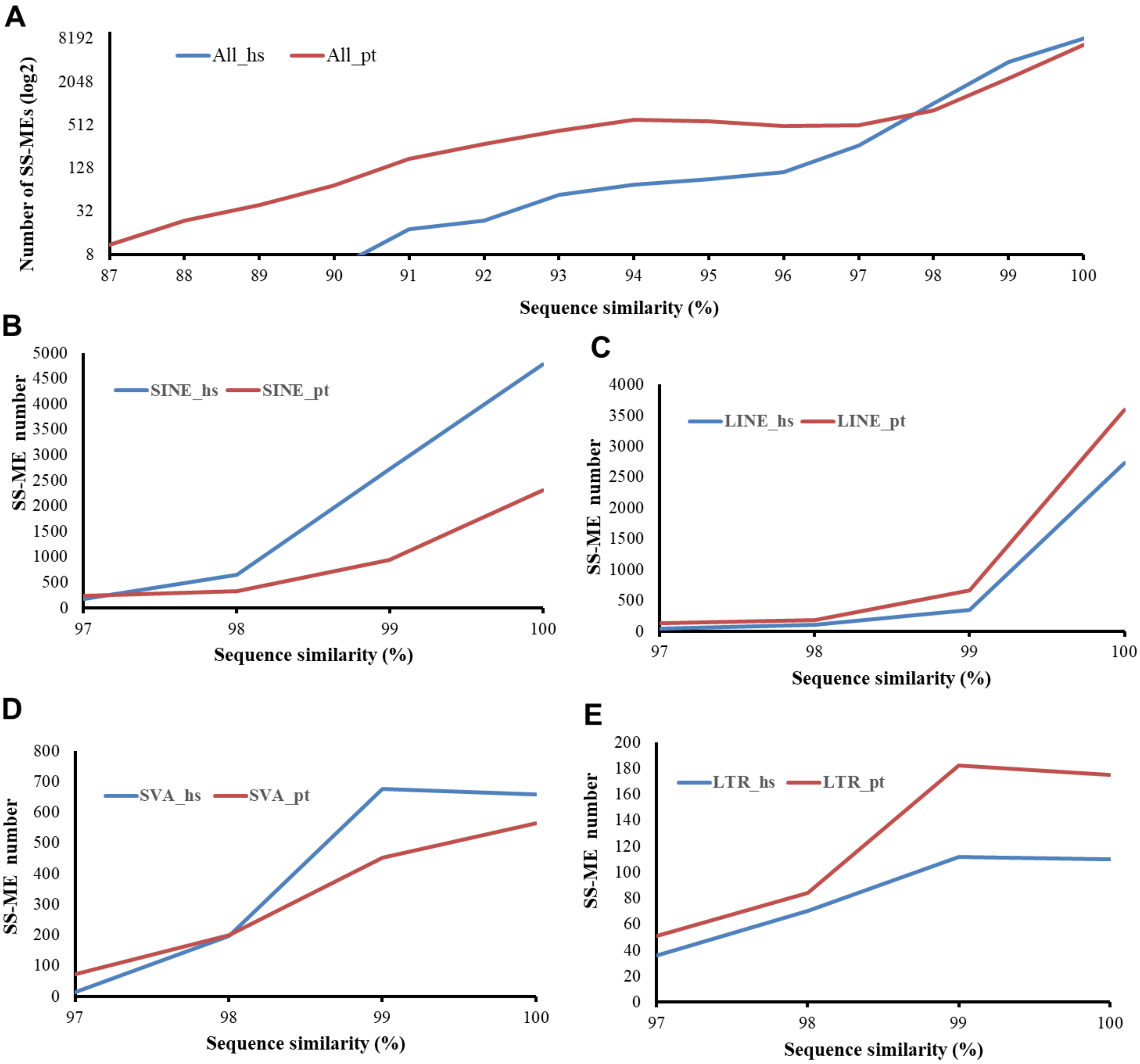


Figure 2.4 The comparison of activity profiles of species-specific mobile elements (SS-MEs) in the human and chimpanzee genomes

A. The numbers of SS-MEs with sequence similarity at 87% or more to another copy of SS-MEs in the human and chimpanzee genomes with Y-axis shown in log2 scale. B-E: The number of SS-MEs with sequencing similarity at 97% or more with another copy of SS-ME in the same genome for SINE (B), LINE C), SVA (D), and LTR (E) in the human and chimpanzee genomes. "_hs" and "_pt" in the data labels indicates for human and chimpanzee genome, respectively.

The higher ratio of the most recent SS-SINEs in human than in chimpanzee is consistent with the higher number of most recent SS-SINEs in human despite chimpanzee having more SS-SINEs. This indicates that the human genome has a higher most recent SINE activity than in the chimpanzee genome, while the latter had a higher earlier SINE activity. To verify this, we analyzed the activity profiles of SS-MEs in associate with the ME age by ME class in these two genomes based on sequence divergence level of SS-MEs by performing an all-against-all sequence similarity search among all MEs in each genome. In this case, the analysis was based on raw SS-MEs since the two genomes were mutually the closest among the eight genomes, therefore the raw SS-MEs are directly comparable. As shown in Fig. 2.4, the age profiles of SS-ME classes are quite different between different ME classes in the same genome and between the two genomes for the same ME classes. The human genome showed a lower level of overall activity earlier, but a much more rapid increase of activity towards the more recent period as reflected by the higher ratios of SS-MEs at high sequence similarity levels (Fig. 2.4A). The higher most recent ME transposition activity in the human genome seems to be contributed by SINEs and SVAs with SINEs showing the largest differences in activity with the chimpanzee and contributing most to the higher number of most recent SS-MEs in the human genome compared to the chimpanzee genome (Fig. 2.4B & D). The chimpanzee genome showed a higher most recent activity for LINEs and LTRs (Fig. 2.4C & E). Interestingly, SVAs in the human genome showed a lower activity early on, but a quicker acceleration, followed by a trend of plateauing or even a slightly lower towards the most recent period, while SVAs in chimpanzee genome showed lower but steady increase of activity all the way to the most recent period (Fig. 2.4D). This seems to correlate well with the observation that human genome has the younger SVA-F and SVA-E subfamilies being more active than the older SVA-D, while the chimpanzee genome has only SVA-D active (Fig. 2.5), supporting SVA-E and SVA-F being human-specific (Wang et al. 2005).

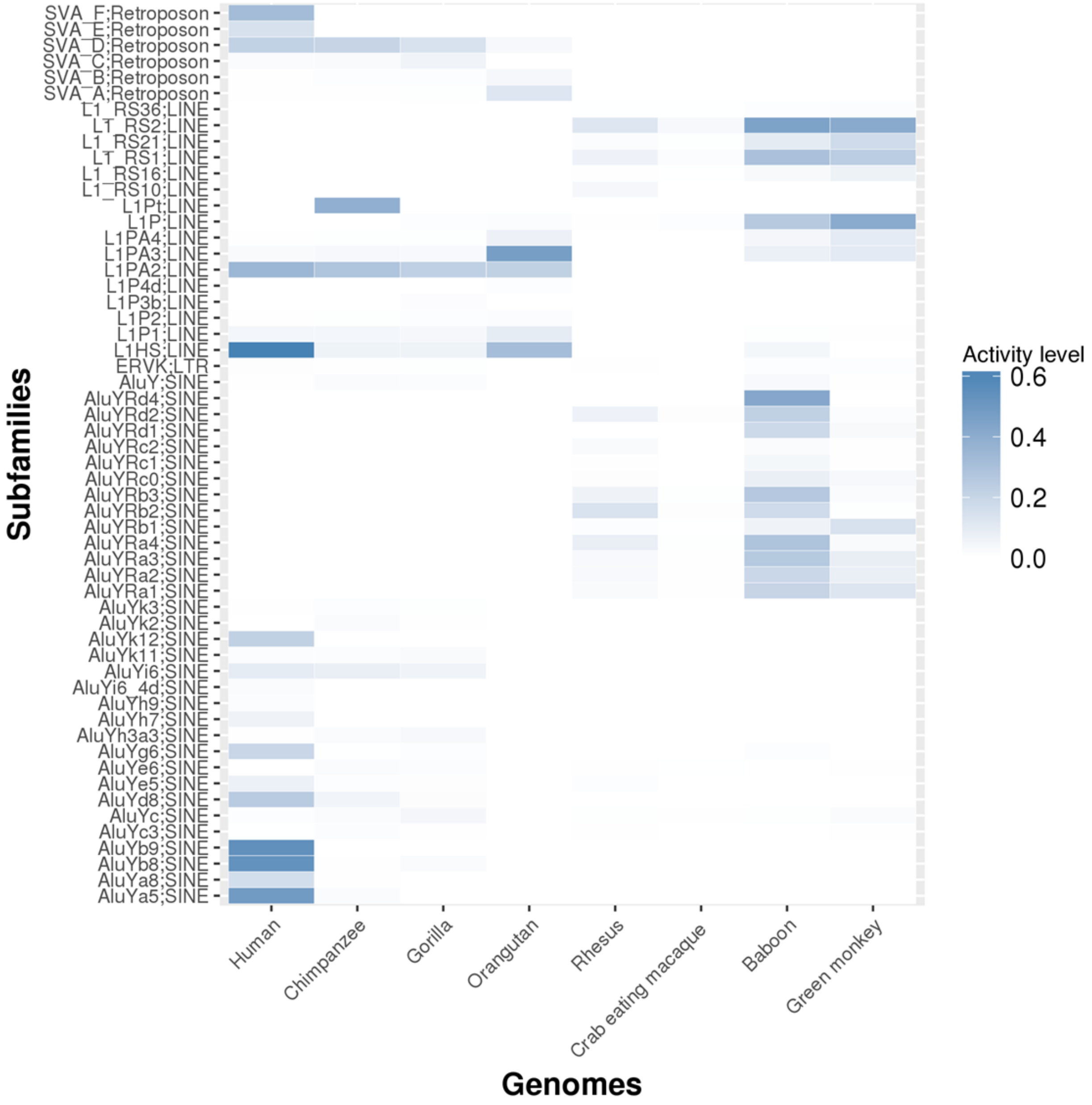


Figure 2.5 A heat map of mobile element (ME) subfamily activity in primate genomes based on most recent species-specific MEs.

A total of 56 different ME subfamilies, which have activity ≥1% in at least one genome, are selected and represented in the heat map. The activity level was calculated as the percentage of the most recent SS-MEs among the total number of MEs in the same subfamily in the genome. The detailed numeric values used to generate this heat map can be found in Table 2.5.

2.4.4 The most active ME subfamilies in the eight primate genomes based on the most recent SS_MEs

The lists of most recent SS-MEs provide an unbiased measure for the relative level of ME accumulation during the most recent/current period across the genomes, as well as among different ME classes and subfamilies. Table 2.5 shows the most recent transposition activity by ME class in each genome calculated as the percentage of the most recent SS-MEs in all MEs in a class. Only the ME subfamilies showing a minimum of 1% in activity in at least one of the genomes were kept. A total of 56 non-redundant subfamilies were collected across the eight genomes, among which 32, 16, 6, and 1 belong to SINE, LINE, SVA, and LTR, respectively (Table 2.5). A visual representation for active ME subfamilies and their relative activity levels in the eight genomes is shown as a heatmap (Fig. 2.5), while the top 5 active ME subfamilies in each genome were also shown in Fig. 2.6. As shown in Fig. 2.5 and Fig. 2.6, each genome has a unique profile of active MEs that differ not only by ME subfamilies but also by their relative activity levels.

Table 2.5 The most active mobile element (ME) subfamilies in eight primate genomes*

ME Subfamily	Class	Human	Chimpanzee	Gorilla	Oran-gutan	Green monkey	Crab-eating macaque	Rhesus	Baboon
AluY	SINE	0.5%	1.8%	1.5%	0.0%	0.3%	0.0%	0.1%	2.9%
AluYa5	SINE	49.7%	1.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
AluYa8	SINE	15.7%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
AluYb8	SINE	54.5%	0.5%	2.1%	0.0%	0.0%	0.0%	0.0%	0.0%
AluYb9	SINE	55.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
AluYc	SINE	0.9%	1.9%	4.1%	0.0%	2.0%	0.4%	0.5%	0.7%
AluYc3	SINE	0.2%	1.3%	0.4%	0.0%	0.5%	0.0%	0.3%	0.0%
AluYd8	SINE	25.3%	4.9%	1.3%	0.2%	0.0%	0.0%	0.0%	0.0%
AluYe5	SINE	6.8%	1.0%	0.8%	0.0%	0.0%	0.0%	0.9%	0.0%
AluYe6	SINE	0.0%	1.8%	1.5%	0.0%	0.4%	0.2%	0.4%	0.0%
AluYg6	SINE	19.6%	0.2%	0.9%	0.0%	0.0%	0.0%	0.0%	1.0%

AluYh3a3	SINE	0.3%	1.7%	3.4%	0.0%	0.0%	0.0%	0.0%	0.0%
AluYh7	SINE	5.9%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
AluYh9	SINE	1.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
AluYi6	SINE	9.2%	7.6%	5.4%	0.0%	0.0%	0.0%	0.0%	0.0%
AluYi6_4d	SINE	2.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
AluYk11	SINE	1.5%	1.6%	2.2%	0.0%	0.0%	0.0%	0.0%	0.0%
AluYk12	SINE	22.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
AluYk2	SINE	0.1%	1.9%	0.4%	0.0%	0.0%	0.0%	0.0%	0.0%
AluYk3	SINE	0.3%	1.0%	0.6%	0.0%	0.0%	0.0%	0.0%	0.0%
AluYRa1	SINE	0.0%	0.0%	0.0%	0.0%	12.2%	0.3%	2.5%	20.6%
AluYRa2	SINE	0.0%	0.0%	0.0%	0.0%	7.4%	0.3%	2.8%	19.5%
AluYRa3	SINE	0.0%	0.0%	0.0%	0.0%	8.1%	0.3%	2.9%	26.4%
AluYRa4	SINE	0.0%	0.0%	0.0%	0.0%	2.2%	0.6%	7.5%	29.1%
AluYRb1	SINE	0.0%	0.0%	0.0%	0.0%	14.0%	0.2%	1.1%	5.6%
AluYRb2	SINE	0.0%	0.0%	0.0%	0.0%	0.6%	0.9%	13.7%	17.3%
AluYRb3	SINE	0.0%	0.0%	0.0%	0.0%	2.0%	0.6%	5.7%	26.7%
AluYRc0	SINE	0.0%	0.0%	0.0%	0.0%	3.1%	0.2%	0.8%	7.4%
AluYRc1	SINE	0.0%	0.0%	0.0%	0.0%	0.4%	0.2%	0.5%	4.0%
AluYRc2	SINE	0.0%	0.0%	0.0%	0.0%	0.3%	0.0%	2.5%	1.9%
AluYRd1	SINE	0.0%	0.0%	0.0%	0.0%	3.0%	0.0%	0.0%	18.5%
AluYRd2	SINE	0.0%	0.0%	0.0%	0.0%	1.5%	0.8%	6.1%	22.2%
AluYRd4	SINE	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	43.2%
ERVK	LTR	0.7%	0.5%	0.7%	0.0%	1.2%	0.1%	0.3%	1.1%
L1_RS1	LINE	0.0%	0.0%	0.0%	0.0%	24.3%	1.6%	6.2%	30.2%
L1_RS10	LINE	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	3.7%	0.0%
L1_RS16	LINE	0.0%	0.0%	0.0%	0.0%	6.4%	0.2%	0.5%	3.0%
L1_RS2	LINE	0.0%	0.0%	0.0%	0.0%	42.1%	3.1%	11.7%	45.7%
L1_RS21	LINE	0.0%	0.0%	0.0%	0.0%	17.3%	0.6%	1.3%	9.7%
L1_RS36	LINE	0.0%	0.0%	0.0%	0.0%	1.3%	0.2%	0.2%	0.9%
L1HS	LINE	63.0%	6.0%	5.9%	32.4%	0.0%	0.0%	0.0%	4.0%
L1P	LINE	0.0%	0.0%	1.2%	1.6%	41.2%	0.9%	0.5%	26.4%
L1P1	LINE	4.3%	4.5%	3.6%	9.1%	0.3%	0.0%	0.1%	0.7%
L1P2	LINE	0.2%	0.6%	1.1%	1.3%	0.1%	0.1%	0.0%	0.0%
L1P3b	LINE	0.0%	0.0%	1.2%	0.0%	0.0%	0.0%	0.0%	0.0%
L1P4d	LINE	0.0%	0.0%	0.0%	1.0%	0.0%	0.0%	0.0%	0.0%
L1PA2	LINE	35.4%	28.6%	22.6%	22.4%	0.0%	0.0%	0.0%	0.0%
L1PA3	LINE	2.3%	3.2%	2.8%	48.3%	9.9%	0.0%	0.0%	7.3%
L1PA4	LINE	0.3%	0.7%	0.6%	6.5%	9.4%	0.0%	0.0%	4.1%
L1Pt	LINE	0.0%	39.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
SVA_A	Retroposon	0.4%	0.3%	0.7%	12.1%	0.0%	0.0%	0.0%	0.0%
SVA_B	Retroposon	0.3%	1.1%	1.0%	3.7%	0.0%	0.0%	0.0%	0.0%
SVA_C	Retroposon	1.7%	2.3%	5.4%	0.0%	0.0%	0.0%	0.0%	0.0%
SVA_D	Retroposon	22.0%	20.6%	14.1%	3.2%	0.0%	0.0%	0.0%	0.0%

SVA_E	Retroposon	14.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
SVA_F	Retroposon	32.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

*The activity is calculated as the percentage of most recent SS-MEs in all MEs in the same subfamily in a genome. Only subfamilies with activity >=1% in at least one primate genome are included.

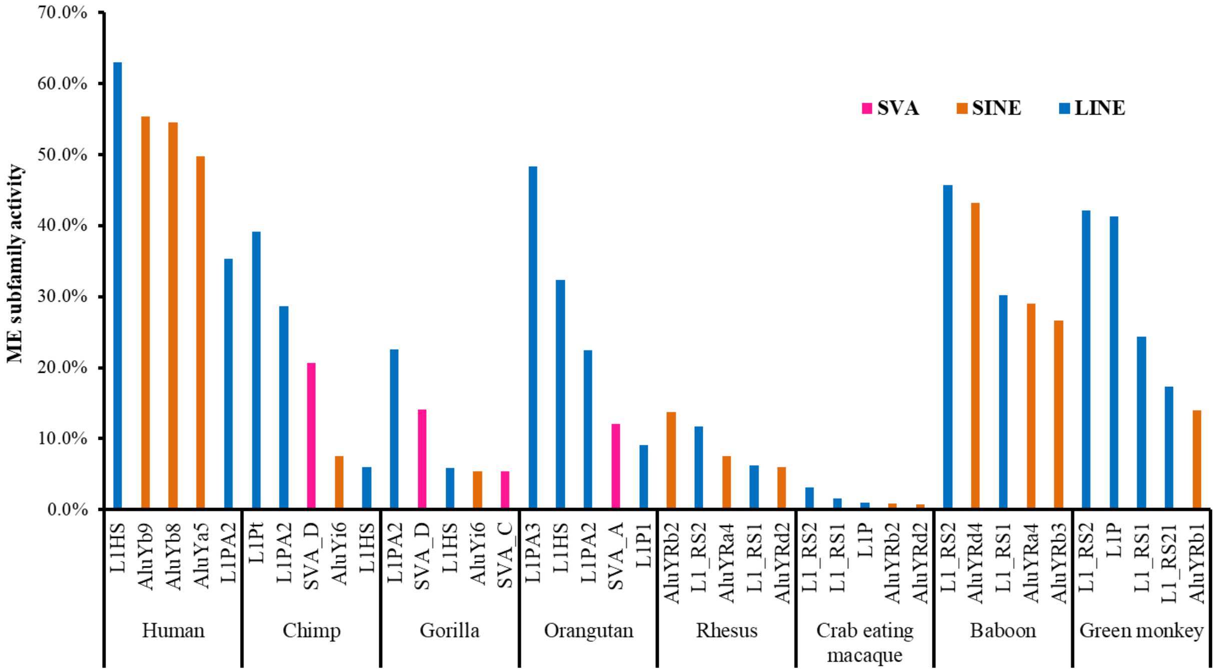


Figure 2.6 Most active subfamilies of mobile elements (MEs) in the eight primate genomes.

The top 5 active ME subfamilies in each primate genome are listed. The activity level of each ME subfamily was calculated by dividing the numbers of most recent SS-MEs with the total numbers of MEs in the subfamily.

Consistent with having the largest number of SS-SINEs, the baboon genome has the largest number of Alu subfamilies at high activities (10 subfamilies at 10% or more) despite none being the highest among all genomes (Table 2.5 and Fig. 2.6). Similarly, the orangutan genome has the largest number of recently active L1 subfamilies (4 at 17% or more) and with 4 of its top 5 active ME subfamilies being from LINEs, all at relatively high activity, explaining its largest number of SS-LINEs. Next to the orangutan genome, the green monkey genome also seems to have a high level of recent L1 activity by having 4 of the 5 top active ME subfamilies from L1, all with relatively high levels of activity, supporting its high number of most recent SS-LINEs (Tables 2.4 & 2.5, Fig. 2.1B & 2.6). In the human genome, AluYa5, AluYb8, and AluYb9 are the most active SINE subfamilies, while L1HS and L1PA2 are the most active LINE subfamilies. Four of these five subfamilies (L1HS, AluYb9, AluYb8, and AluYa5) have the highest activity among all ME subfamilies from all genomes, with the 5th ME subfamily (L1PA2) and 3 SVA subfamilies also have the highest among the same subfamilies from all *Hominidae* genomes (Table 2.5, Fig. 2.5). These data indicate that the human genome has the highest most recent ME transposition activity among the eight genomes. For SVA as the youngest ME class uniquely found in the *Hominidae* group, all of its 6 subfamilies got onto the list of active ME subfamilies with activity more than 1% (Table 2.5 and Fig. 2.5). The highest activity seen among the SVA subfamilies is with the youngest SVA-F in the human genome (32.6%). There seems to a high positive correlation between the age of the species and the age of active SVA subfamilies with the orangutan as the oldest and having the oldest active SVA subfamily and the human genome being the youngest having the youngest active SVA subfamilies and at the highest activities (Table 2.5 and Fig. 2.5). For LTRs, only the ERVK subfamily barely got onto the list of active ME subfamilies with the green monkey and baboon genomes have higher activity (~1.1%), indicating the overall low activity of LTRs in all these genomes compared to the non-LTR retrotransposons (Table 2.5).

It is worth noting that, in contrast with all other genomes, the crab-eating macaque genome lacks a single highly active ME subfamily (Table 2.5, Fig. 2.5 & 2.6) with the highest being 1.6% for L1RS1. It explains the extremely small number of SS-MEs, and it once again reinforces the possibility for the existence of a universal mechanism in suppressing all ME transposition.

2.4.5 Differential impact of ME transposition on primate genome sizes

We compared across the eight genomes the impact of SS-MEs on genome size via insertion of MEs and generation of TSDs and transductions, as well as possible genome size reduction through insertion-mediated deletions (IMD) of flanking sequences. In this case, we used the raw SS-MEs for the initial size calculation followed by normalizing the total size change based on the evolutionary distance for comparison.

Table 2.6 Impact of species-specific mobile elements (SS-MEs) on genome size (Kb)

Genome	ME insertion	TSD	Transduction	IMD	raw total	Normalized total
Human	14,259	171	687	-977	14,141	11,473
Chimpanzee	16,274	118	1,033	-11,403	6,021	4,885
Gorilla	5,895	89	1,086	-4,073	2,996	1,611
Orangutan	33,924	243	3,741	-12,381	25,527	6,776
Rhesus	11,074	139	2,616	-10,700	3,128	3,128
Crab-eating macaque	1,797	15	646	-1,184	1,274	1,274
Baboon	29,342	581	6,063	-12,448	23,537	23,003
Green monkey	17,330	353	4,435	-16,377	5,742	2,599
Total	129,894	1,709	20,307	-69,543	82,368	NA
Average	16,237	214	2,538	-8,693	10,296	6,844

*: TSD, target site duplications; IMD, insertion-mediated deletions

As shown in Table 2.6, in all eight genomes, SS-MEs have led to a net genome size increase. Collectively, SS-MEs have contributed to a combined ~82.3 Mbp increase in the eight genomes or on average ~10 Mbp per genome or ~7 Mbp with normalization. However, the degree of size increase varies significantly among the genomes with the baboon genome gaining the largest increase (~23.5 Mb) and the crab-eating macaque genome gaining the least (~1.2 Mb), which is directly correlated with the overall levels of SS-MEs. Among the different types of size impact, the insertion of ME sequences is responsible for the majority of the size increase as expected, followed by transductions, and TSDs and with insertion mediated deletion (IMD) contributing to a significant amount of size loss offsetting the size increases from the insertions (Table 2.6).

2.4.6 SS-MEs impact genes in the primate genomes

To predict the functional impact of SS-MEs, we analyzed the gene context of their insertion sites based the gene annotation data in human from the GENCODE project (Release 23, July 2015) (Harrow et al. 2012) combined with the NCBI RefGene annotation set (Pruitt et al. 2007) and ENSEMBL gene annotation data for the non-human primates (Zerbino et al. 2018). For this purpose, we used the raw list of SS-MEs as these represent the accumulated differences among the species examined.

Table 2.7 The distribution of species-specific mobile elements (SS-MEs) in the genic regions in eight primate genomes

Genic region*	CDS	NR	Promoter	5' UTR	3' UTR	Intron	Total
Human	40	205	242	11	55	7,033	7,586
Chimpanzee	32	11	100	1	11	6,909	7,064
Gorilla	52	16	103	0	3	4,028	4,202
Orangutan	27	26	240	1	43	8,798	9,135
Green monkey	27	28	397	4	192	14,071	14,719
Crab-eating macaque	12	3	24	0	7	1,156	1,202
Rhesus	48	18	261	2	107	9,617	10,053
Baboon	13	51	579	9	260	21,773	22,685
Total	**251**	**358**	**1,946**	**28**	**678**	**73,385**	**76,646**

*, CDS: coding sequence; NR: non-coding RNA; UTR: untranslated region

As shown in Table 2.7, a total of 76,646 SS-MEs, representing ~33.5% of all SS-MEs, are located in genic regions, which include protein-coding genes, non-coding RNAs and transcribed pseudogenes. Similar to our observation for the human-specific MEs (Tang et al. 2018), most of these genic SS-MEs (95.7%) are located in intron regions, while 609 SS-MEs contribute to exon regions as part of transcripts. Furthermore, these SS-MEs potentially impact the CDS regions of more than 251 unique genes, which cover all eight genomes (Tables 2.8 & Appendix I).

Table 2.8 SS-MEs in potential protein coding genes in eight primate genomes

ME Class	SINE	LINE	LTR	SVA	Total
Human	1	5	2	32	40
Chimpanzee	13	16	0	3	32
Gorilla	23	26	3	0	52
Orangutan	5	11	2	9	27
Green monkey	20	7	0	0	27
Crab-eating macaque	9	3	0	0	12
Rhesus	39	7	2	0	48
Baboon	9	0	4	0	13
Total	**119**	**75**	**13**	**44**	**251**

2.5 Discussions

In this study, we deployed a comparative computational genomic approach recently developed for the analysis of human-specific MEs (Tang et al. 2018) for a larger scale comparative genomic analysis involving a total of eight primate genomes with four representing each of the top two families of primates, the *Hominoidea* and *Cercopithecoidea*. Our analysis provided the first set of comprehensive lists of MEs that are uniquely owned by each of these primate genomes based on the most updated reference sequences. Collectively, we identified a total of 228,450 SS-MEs from these eight primate genomes, among which 84,274 were considered to have occurred very recently in these genomes (Table 2.3). These lists of SS-MEs and most recent SS-MEs allowed us to observe the differential ME transposition and its impact in primate evolution. We discussed below the relevance of our results in several aspects.

2.5.1 The challenges in the identification of SS-MEs

The reason for the lack of large-scale comparative studies for ME transposition in primates is partly due to many challenges in this task as previously discussed in our recent work on human-specific MEs (Tang et al. 2018). These challenges include, but are not limited to 1) the high content of MEs in the primate genomes, 2) the reference genome sequences are still incomplete, especially for the non-human primate genomes, 3) genome assembly errors, especially for regions rich of repeat elements, which can mislead the results, 4) variable quality of ME annotation from different genomes from the use of different versions of repeat reference sequences (i.e., Repbase) and RepeatMasker (Jurka et al. 2005; Smit et al. 2013; Tarailo-Graovac and Chen 2009), and 5) variable mutation rate across species (Scally and Durbin 2012), which could have an impact on the analysis of the most-recent SS-MEs based on a sequence similarity cutoff. For non-human primate genomes, the 2nd and 3rd issues are larger than for human genome due to the generally lower quality of the reference genome assemblies (Table 2.1). The gap regions are usually biased towards the repeat sequence regions, and therefore, the different quality level of the reference genomes might have contributed to an unknown but likely small portion of the SS-ME differences reported in our study. For the 4th issue, in our tests with different versions of Repbase and RepeatMasker, different numbers of annotated MEs in the same version of the genome were seen, but the difference in the total numbers of MEs are all below 1%, while the discrepancies in ME subfamily assignment can be higher in some cases, especially for some small and new subfamilies, but are no more than 10%, mostly below 5% (data not shown). Therefore, the variation in annotation quality may affect the subfamily activity calculation, but it should have a very small impact on the total number of SS-MEs by ME class. In addition to these 4 issues, we also faced the lack of certain resources, for example, data linking the orthologous regions across closely related genomes (e.g. liftOver overchain files on the UCSC genome browser) and functional annotation data are mostly missing for comparative analysis among non-human primates. For these reasons, we believe that our lists of SS-MEs still suffer a certain level of both false negatives and false positives. We can expect the situation to improve with continuing improvement of the genome assemblies, for example, benefiting from the use of newer generations of sequencing platforms that can provide much longer reads, such as the Nanopore and PacBio platforms (Roberts et al. 2017; Schneider and Dekker 2012). The numbers of SS-MEs can be expected to have a certain level of increase from regions with sequencing gaps, especially regions highly rich of repeats, such as the centromere and telomere regions, which may be hot spots for certain types of MEs, such as LTRs (Tang et al. 2018).

2.5.2 The differential ME transposition among primate genomes

Despite more and more non-human primate genomes having been sequenced and assembled in the recent years, prior studies on ME transposition have mostly focused on the analysis of ME profiles for individual genomes separately (Battilana et al. 2006; Ewing and Kazazian 2011; Jha et al. 2009; Jordan et al. 2018; Mills et al. 2006; Ray et al. 2005; Steely et al. 2018; Stewart et al. 2011; Tang et al. 2018; Wang et al. 2006). So far, only very limited comparative analyses involving a small number of genomes have been reported. Among these, the work by Mills et al (Mills et al. 2006) compared the ME profile between human and chimpanzee, and a recent study has focused on lineage-specific *Alu* subfamilies in the baboon genome (Steely et al. 2018). Due to the challenges described above, a large scale systematic comparative analysis of mobile elements in primate genomes still represents a gap in the field. In this study, we focused on the SS-MEs that represent the results of ME transposition events uniquely occurred in each of the eight primate genomes since divergence from their perspective closely related genomes in this group.

Our SS-ME data demonstrate that each primate genome displays a remarkably different ME accumulation profile as measured both by the total number of SS-MEs (both raw and normalized), the most recent SS-MEs, and the specific ME composition by ME class and subfamilies for each of these sets of SS-MEs. Among the eight primate genomes examined, the raw number of SS-MEs in a genome varies from the highest at 66,578 copies in the baboon genome to the lowest at 3,281 copies in the crab-eating macaque genome, and with the remaining six genomes ranked from high to low as green monkey, rhesus, orangutan, chimpanzee, human, and the gorilla genomes (Table 2.4 and Fig. 2.1A). While these raw numbers of SS-MEs did provide us a quick snap shot of the SS-ME transposition among these genomes, they are not appropriate for accurate measurement of the differential ME transposition in these genomes. This is because the raw number of SS-MEs in each primate genome represents the total number of new MEs accumulated from past ME transposition since the divergence from the relative last common ancestor (LCA) among the species included in this analysis. Therefore, the number of SS-MEs is directly impacted by both the level of ME transposition and the relative distance from their LCA, with the latter being variable among the eight primates. To avoid this bias caused by the variable evolutionary distance, we obtained the normalized numbers of SS-MEs and the numbers of most recent SS-MEs. The normalized numbers of SS-MEs based on the relative evolutionary distance permits comparison of the relative total ME accumulation in a genome since its relative LCA, while the numbers of most recent SS-MEs are independent of the evolutionary distance and reflect the most recent or current ME transposition level in a genome.

Among the eight genomes, the baboon genome stands out with the largest numbers of SS-MEs and the most recent SS-MEs, mainly due to its most successful Alu transposition from a large number of highly active Alu subfamilies (Table 2.5, Fig. 2.5). This is supported by the findings from two recent studies, showing that the baboon genome has a dramatically elevated recent Alu insertions contributed to the presence of a larger number of baboon-specific Alu subfamilies (Rogers et al. 2019; Steely et al. 2018). The fact that, despite the great success of Alu transposition in the baboon genome, none of the active Alu subfamilies have the top level of most recent activity among the eight genomes suggests that Alu transposition might have been kept at a more constant and high rate during the evolution of the baboon genome, unlike the human genome, which seems to have a more recent accelerating for SINEs/Alus (Fig. 2.4B).

The crab-eating macaque genome has a strikingly low number of SS-MEs, being less than 1/12 of that for averages across all eight primate genomes, ~1/16 of *Cercopithecidae* family average, and ~1/35 of that for the baboon genome (Table 2.3 & 2.4). Along with the fact that its most recent number of SS-MEs is also in the same situation, the observation that the crab-eating macaque genome lacks a single highly active ME subfamily from any ME class (Table 2.4, Fig. 2.1 & 2.3) strongly suggest the existence of a molecular mechanism, which imposes a strong genome-wide suppression of ME transposition in this genome. One possible such mechanism may be related to epigenetic regulation, such as a genome-wide DNA hyper-methylation during gametogenesis, as DNA methylation has been known to suppress ME transposition (Law and Jacobsen 2010).

The normalization of SS-MEs based on evolutionary distance is not without caveats. First, the normalization is based on the assumption that the ME transposition rate was constant over the time for all ME types, which turned out be untrue by data from this study. Second, having an accurate estimation of the evolutionary for species seems to be an unreachable target due to lack of the ground truth. This is because the evolution distance estimate can vary significantly from gene to gene and from study to study and getting a consensus from multiple studies, which is what offered by the TimeTree database (http://timetree.org) (Hedges et al. 2006), still does not provide an ultimate answer. For example, the reported distance between baboon and rhesus ranges drastically from 6.6 to 49.1 million years (My) among the 36 studies collected in the TimeTree database, and TimeTree provides 12.4 My as the estimate for the distance between the two species. This is larger than the distance between gorilla and human (9.06 My from TimeTree). While our *ACTB* CDS sequence-based phylogeny shows a similar tree topology with the tree from TimeTree (data not shown), it shows a closer distance among the four monkey genomes than the four ape genomes (Fig. 2.2A), with the distance between baboon and the two macaque species being much closer. While our data seem to be better supported by a closer distance between baboon with the macaques than a very large distance, like the one from TimeTree, some of the details in pattern of normalized SS-MEs we observed among the eight primate genomes (Fig. 2.1A) might not be very accurate due to the uncertainty in the accuracy of the distance estimates.

To overcome the above issues, we used the numbers of the most recent SS-MEs, which is independent of the evolutionary distance of the genomes, to provide an alternative approach in measuring the differential ME transposition among these genomes. In identifying the most recent SS-MEs, we applied the highest stringency (100% sequence similarity), such that it allows us to focus on the shortest period of the speciation towards the current genomes. It also helps to reduce the problem with the normalized SS-MEs being smaller than the most recent SS-ME mentioned earlier. Certainly, this approach is still not perfect, because the mutation rate may be variable across the genomes (Smith and Donoghue 2008), meaning that a genome with a higher rate of the mutation will show an underestimated number of the most recent SS-MEs by this method. However, we can expect that the degree of the mutation rate variation to be small and focusing on the most recent and shortest period of the genome evolution may help minimize its effect on our result.

The comparison of between the profile of SS-MEs and most recent SS-MEs across closely-related genomes provides us more details about the differences of ME transposition among genomes. For example, between human and chimpanzee genomes, even though the latter has a higher number of SS-MEs for all ME classes examined, the human genome has a higher number and higher ratio of most recent SS-MEs (Table 2.4, Fig. 2.3C). The largest difference is seen for SS-SINEs and in this case the normalization would not have an impact on the comparison, as the two genomes are the mutually closest; while the human genome has significantly fewer SS-SINEs than the chimpanzee genome (8,844 vs. 10,612 for raw SS-MEs), it has more than double of the most recent SS-SINEs than in the chimpanzee genome (8,131 vs. 3,587) (Table 2.4). The ratio of the most recent SS-SINEs is 67% for human genome compared to ~27% for chimpanzee genome (Fig. 2.3C). Higher ratios of most recent SS-MEs in the human genome are also seen for LINEs and LTRs (Fig. 2.3C), despite their numbers being lower than the count parts in the chimpanzee genome (Table 2.4). These data may suggest that, relatively speaking, between the two genomes as shown in Fig. 2.4A, the overall ME transposition was relatively lower in the human genome during earlier stage but accelerated more due to the emergence of a few very young and active SINE subfamilies, such as AluYa5, AluYb8, and AluYb9, along with L1HS, and SVA-F (Fig. 2.6, Table 2.5). For SVAs, human has two species-specific subfamilies which are highly active, in addition to the older and active SVA-D subfamily that is shared with chimpanzee. These young and highly active ME subfamilies contributed to the higher numbers of most recent SS-SINEs and SS-SVAs, as well as for the larger total number of most recent SS-MEs in the human genome than in the chimpanzee genome (Fig. 2.4A, B & D). It is worth noting that our lists of SS-MEs for human and chimpanzee (14,947 and 22,087, respectively) are significantly larger than the number of species-specific MEs reported in an earlier comparative study involving just a pairwise comparison between the same two genomes with earlier versions of the genome sequences (Mills, et al. 2006) (7,786 and 2,933, respectively). Further, our data reveal a different trend with more detailed picture, showing chimpanzee with a larger number of SS-MEs (vs a smaller number of SS-MEs reported by Mills, et al, 2006), while human having a larger number of most recent SS-MEs. This is likely attributed mainly to the much improved genome assembly quality and our more robust methodology involving multi-way genome comparison.

Similar to human genome, the baboon genome also has a very high recent activity of SINEs due to highly active subfamilies, such as AluRd4 and AluRd2 (Fig. 2.5 & 2.6, Table 2.5), and this is in good agreement with results of two recent studies (Rogers et al. 2019; Steely et al. 2018). Interestingly, in the human genome, the activities for four of the five most active ME subfamilies (L1HS, AluYb9, AluYb8, and AluYa5) are higher than any active ME subfamilies in the other genomes (Fig. 2.6), revealing the human genome as the most active among the eight primate genomes by its most active recent ME transposition. It is also worth noting that all of the most active ME subfamilies from all genomes belong to the non-LTR retrotransposons, which are all driven by the L1-based TPRT mechanism (Goodier 2016). This agrees with the data in L1base and our recent observation, which show that the human genome has the largest number of functional L1s among primates (Penzkofer et al. 2017) and with most of these L1s being human-specific and even polymorphic (Nanayakkara, et al, manuscript in preparation). We would like to believe that the presence of a large number of human-specific functional L1s might have provided the unique opportunities for the emergence of many young and active non-LTR ME subfamilies during human evolution, a trend which may extend to future evolution. In a similar way, it is interesting to observe that in all eight genomes the ratio of most recent SS-LINEs (L1s) is the highest among all ME classes (Fig. 2.3C) regardless of the overall level of SS-MEs. This is expected, as the functional L1 retrotransposition machinery is necessary for the activity of all non-LTR retrotransposons, including LINE, SINE, SVA, and processed pseudogenes (Goodier 2016).

In summary, our data indicate that the overall differential ME transposition among the eight primate genomes came as a result of their different composition of young ME by class and subclass, as well as differential temporal profile of ME transposition during the evolution of these genomes since the divergence from their perspective LCA.

2.5.3 The impact of differential ME transposition on primate genomes.

ME transition is known to be one of the dominant contributors for genome size variation among species with a positive linear relationship between the percentage of MEs and genome size (Kidwell 2002; Lee and Kim 2014). As an example, maize has one of the largest genomes among plants with 85% of its genome contributed by repetitive sequences, among which 63% are recognizable MEs, being the genome with the highest percentage of MEs reported so far (Baucom et al. 2009; Jiao et al. 2017). The 230,855 SS-MEs from the eight primate genomes have collectively contributed to ~82 Mbp net increase in these genomes (Table 2.6) with a net increase in each genome, ranging from ~1.2 Mbp in the crab-eating macaque genome to ~25.5 Mbp in the orangutan genome, showing ME transposition as a very important, likely the most significant molecular mechanism contributing to genome size increases in primate genomes as previously discussed (Tang et al. 2018).

In addtion to impact on genome size, MEs are also known to have direct impacts on gene function by participating in or interrupting protein coding or by participating in gene regulation (see reent review by Bourque et al (Bourque et al. 2018)). By using the latest annotation data for these genomes, we were able to provide a preliminary assessment of SS-MEs' potential impact on genes. Our results showed that a total of 76,646 SS-MEs, representing 33.5% of all SS-MEs, are located in genic regions, which include protein coding genes, non-coding RNAs, and transcribed pseudogenes (Table 2.7). This ratio is lower than the 50.7% previously reported for the human-specific MEs (Tang et al. 2018), likely due to the fact that the human genome is much better annotated than the non-human primate genomes. Among these genic SS-MEs, 609 can potentially be part of the primate transcriptomes by contributing to exons. Interestingly, in 251 of these cases, an SS-ME contributes to the protein coding sequence (CDS) in a transcript. As shown in Appendix I, most of these CDS SS-MEs are SS-SINEs (119/251), even more so in the the *Cercopithecidae* family (77/100). In this study, we did not cover the analysis of SS-MEs' contribution to regulatory elements in consideration that less mature related data resources are available for non-human primates, especially for the species-specific portion. Such analysis can be certainly be part of the future studies on these SS-MEs. For these and other reasons, our assessment of the functional impact of SS-MEs certainly represents an underestimation of what could exist in these genomes.

In summary, our data suggest that, similar to human-specific MEs (Tang et al. 2018), SS-MEs in non-human primate genomes have the potential to participate in gene function by their presence in the gene vicinity in a species-specific fashion and along with other genetic variations are likely responsible for lineage-specific traits as illustrated in literature (Oliver and Greene 2011).

2.6 Conclusions and future perspectives

In summary, our comparative genomic analysis of eight primate genomes involving representatives from the top two primate families, *Hominidae* and *Cercopithecidae*, revealed remarkable differential levels of ME transposition among primate genomes. Each of these genomes was shown to have a unique profile of SS-MEs in terms of their composition by ME class and activity level, and there are also common trends characteristic of lineages. Notably, the ME transposition seems to be lowered to a ground level for all ME classes in the crab-eating macaque genome, likely due to a genome-wide suppression of ME transposition, while it is highly active in the baboon and human genome, each due to the existence of several unique highly active ME subfamilies. Overall, *Hominidae* has relatively more successful LINEs, while *Cercopithecidae* has SINEs as more successful. Remarkable differences in ME transposition are also seen among closely related genomes, as seen between human and chimpanzee genomes, with ME transposition showing a later and quicker acceleration in the human genome compared to the chimpanzee genome. Furthermore, differential ME transposition has made a significant differential impact on the genome size and with the potential also impacting gene function in these genomes, responsible for unique genomic and phenotypic characteristics of each species along with other mechanisms. Future studies may focus on the elucidation of the specific mechanisms underlying such differential ME transpositions in each species and the specific functional impacts on gene functions in the context of species-specific phenotypes. Follow-up studies on the specific mechanism responsible for the extremely low level of ME transposition in the crab-eating macaque genome and its impact on the organism would also be very interesting.

Chapter 3 The identification and characterization of retro-DNA, a new type of retrotransposons originated from DNA transposons, in primate genomes

(The content of this chapter is mostly copied from a manuscript prepared for publication: "The identification and characterization of retro-DNA, a new type of retrotransposons originated from DNA transposons, in primate genomes" Wanxiangfu Tang and Ping Liang with some minor changes for table formats and figure reorganization (renumbered after combining with supplementary figures).

The candidate is the main author of this manuscript and was responsible for generating most of the data included in the manuscript. The manuscript was drafted by the candidate and edited by the corresponding author, Dr. Liang, to its current form.)

3.1 Abstract

Mobile elements (MEs) can be divided into two major classes based on their transposition mechanisms as retrotransposons and DNA transposons. Retrotransposons utilize an RNA-intermediate to transpose in a "copy-and-paste" fashion, and DNA transposons move in the genomes directly in the form of DNA in a cut-and-paste style. In addition to the target site duplications (TSDs), a hallmark of transposition shared by both classes, the DNA transposons also carry terminal inverted repeats (TIRs). DNA transposons constitute ~3% of primate genomes, and they are thought to be inactive in the recent primate genomes since ~37 My ago despite their success during early primate evolution. Retrotransposons can be further divided into Long Terminal Repeat retrotransposons (LTRs), which are characterized by the presence of LTRs at the two ends, and non-LTR which lack LTRs. In the primate genomes, LTRs constitute ~9% of genomes and have a low level of ongoing activity, while non-LTR retrotransposons represent the major types of MEs contributing to ~37% of the genomes with some members being very young and currently active in retrotransposition. The four known types of non-LTR retrotransposons include LINEs, SINEs, SVAs, and processed pseudogenes, all characterized by the presence of a polyA tail and TSDs, which mostly range from 8 to 15 bp in length. All non-LTR retrotransposons are known to utilize the L1-based target-primed reverse transcription (TPRT) machineries for retrotransposition. In this study, we report a new type of non-LTR retrotransposon, which we named as retro-DNA, to represent DNA transposons by sequence but non-LTR retrotransposons by the transposition mechanism in the recent primate genomes. By using a bioinformatics comparative genomics approach, we identified a total of 1,750 retro-DNA elements, which represent 748 unique insertion events in the human genome and nine non-human primate genomes from the ape and monkey groups. These retro-DNA elements, mostly as

fragments of full-length DNA transposons, carry no TIRs but longer TSDs with ~23.5% also carrying a polyA tail and with their insertion site motifs and TSD length pattern characteristic of non-LTR retrotransposons. These features suggest that these retro-DNA elements are DNA transposon sequences likely mobilized by the TPRT mechanism. Further, at least 40% of these retro-DNA elements locate to genic regions, presenting significant potentials for impacting gene function. More interestingly, some retro-DNA elements, as well as their parent sites, show certain levels of current transcriptional expression, suggesting that they have the potential to create more retro-DNA elements in the current primate genomes. The identification of retro-DNA, despite small in number, reveals a new mechanism in propagating the DNA transposon sequences in the primate genomes with the absence of the canonical DNA transposon activity. It also suggests that the L1 TPRT machinery may have the ability to retrotranspose a wider variety of DNA sequences than what we currently know.

3.2 Introduction

Mobile elements (MEs), also known as transposable elements, as a whole constitute significant proportions of the genomes for most higher organisms, being around 50% in primate genomes (Carbone et al. 2014; Chimpanzee Sequencing and Analysis 2005; Cordaux and Batzer 2009; Deininger et al. 2003; Higashino et al. 2012; Lander et al. 2001; Locke et al. 2011; Rhesus Macaque Genome Sequencing and Analysis et al. 2007; Scally et al. 2012; Tang and Liang 2019; Warren et al. 2015; Yan et al. 2011). MEs are defined as genomic sequences capable of changing locations or making copies into other locations within genomes. Despite being initially considered "junk DNA", research from the last few decades has demonstrated that MEs have made significant contributions to genome evolution, and they can impact gene function via a variety of mechanisms: they are known to generate insertional mutations and genomic instability, create new genes and splicing isoforms, exon shuffling, alternate gene expression and epigenetic regulation (Callinan et al. 2005; Chuong et al. 2016; Feschotte and Pritham 2007; Han et al. 2004; Han et al. 2005; Han et al. 2007; Konkel and Batzer 2010; Mita and Boeke 2016; Quinn and Bubb 2014; Sen et al. 2006; Symer et al. 2002; Szak et al. 2003; Wheelan et al. 2005). Additionally, retrotransposons, via germline or somatic insertions, can contribute to genetic diseases in humans (see reviews by Anwar et al. 2017; Goodier 2016).

Based on the types of their transposition intermediates, MEs can be divided into two major classes: the Class I MEs or retrotransposons, which utilize an RNA-intermediate to transpose in a "copy-and-paste" fashion, and the Class II MEs or DNA transposons, which utilize a DNA-intermediate to transpose in a "cut-and-paste" style. Despite both having target site duplications (TSDs), the two ME classes differ in their sequence characteristics, not only in their actual sequences but also in TSD length and whether there are terminal inverted repeats (TIRs) and polyA tail sequence, etc. (Feschotte and Pritham 2007; Pace Ii and Feschotte 2007; Smit and Riggs 1996).

Retrotransposons represent the majority of MEs in primate genomes, owing to their “copy-and-paste” style transposition, which results in direct copy number increase. In this process, a retrotransposon is first transcribed into RNA and then reverse transcribed into DNA as a new copy inserting into a new location in the genome (Kazazian and Goodier 2002). Retrotransposons can be divided into two major subtypes: the LTR and non-LTR retrotransposons, with the former carrying long terminal repeats (LTRs) that are absent from the latter, while the latter mostly have a polyA tail at the 3’-end (Cordaux and Batzer 2009; Deininger et al. 2003). In primates, LTR retrotransposons, mainly as endogenous retrovirus (ERVs) originated from retrovirus affecting and integrating into the germline genomes at various times during primate evolution, constitute ~9.0% of the genomes. In comparison, non-LTR retrotransposons, as the most successful MEs in primate genomes, contribute to more than 35% of the genomes and more than 80% of all MEs (Tang and Liang 2019). By sequence features, currently known non-LTR MEs belong to four subclasses, including Short-INterspersed Elements (SINEs), Long-INterspersed Elements (LINEs), SINE-R/VNTR/Alu (SVA), and processed pseudogenes (i.e. retro-copies of mRNAs). All subclasses of non-LTR retrotransposons, despite having many differences such as size, sequencing feature, and coding capacities, share the common property of having a 3’ polyA tail and the use of target-prime reverse transcription (TPRT) mechanism for retrotransposition (Goodier 2016; Ostertag and Kazazian 2001).

LINE-1s (L1s), being the only subfamily of autonomous non-LTR retrotransposons in the primate genomes, provide the TPRT machinery for all other non-LTR retrotransposons, which are considered non-autonomous for transposition (Cost and Boeke 1998; Goodier 2016; Jurka 1997; Mita and Boeke 2016; Tang et al. 2018; Xing et al. 2006). A functional L1, which is ~6,000 bp long, consists of an internal RNA polymerase II promoter, two open reading frames (ORF1 and ORF2) and a polyadenylation signal followed by a polyA tail (Kazazian and Goodier 2002). The ORF1 gene encodes an RNA-binding protein and ORF2 encodes a protein with endonuclease and reverse transcriptase activity (Goodier 2016; Kazazian and Goodier 2002). Several studies have shown that Alus, L1s, and SVAs have an identical core sequence motif of "TT/AAAA" for the insertion sites, confirming that all non-LTR retrotransposition use the same TPRP mechanism (Cost and Boeke 1998; Jurka 1997; Tang et al. 2018; Wang et al. 2006).

In contrast to retrotransposons, DNA transposons, initially known as the "jumping genes," move in genomes using a transposase encoded by autonomous copies (Deininger et al. 2003). Ten out of the twelve DNA transposon superfamilies are known to excise themselves out from their original locations as double-stranded DNA and move to new sites in the genome, which leads to no direct change in copy numbers (Feschotte and Pritham 2007; Pace Ii and Feschotte 2007). Two of the superfamilies, *Helitrons* and *Mavericks*, transpose through non-canonical mechanisms by utilizing a single-stranded DNA as intermediate, which leads to a "copy-and-paste" style (Feschotte and Pritham 2007; Kapitonov and Jurka 2001; Pritham et al. 2007). The ten "cut-and-paste" DNA transposon superfamilies, as well as *Mavericks*, have the presence of TIRs and TSDs, while *Helitrons* is the only superfamily with neither TIRs nor TSDs, owing to its rolling-circle mechanism (Feschotte and Pritham 2007; Kapitonov and Jurka 2001). In addition to these aforementioned DNA transposons, there is another group of DNA transposons named miniature inverted-repeat transposable element (MITEs) characterized by the presence of both TSDs and TIRs yet lacking the coding capacity for the transposase (Zhang et al.

2000). By using DNA transpose encoded by other autonomous DNA transposons, these non-autonomous, short (50-600 bp) MITE entries can transpose in the host genome (Feschotte et al. 2003; Feschotte and Pritham 2007).

Past studies on MEs in the primate genomes had been mainly focused on the retrotransposons due to their significant contribution to the genome and their active contribution to inter- and intra-species genetic variations as lineage-specific or species-specific MEs driven by young and active members (Ahmed et al. 2013; Battilana et al. 2006; Ewing and Kazazian 2011; Liang and Tang 2012; Stewart et al. 2011). On the contrary, DNA transposons have been considered inactive in the current primate genomes and have received very little research attention. While initial analysis shown evidence in the draft human genome that there was no DNA transposon activity since ~50 My ago (Lander et al. 2001), Pace Ii and Feschotte later suggested that DNA transposon had been highly active till ~37 My ago (Pace Ii and Feschotte 2007). There have been very few, if any, published reports for lineage-specific or species-specific DNA transposons in primate genomes. However, in our recent comparative analysis of species-specific MEs in eight primates from the *Hominidae* and the *Cercopithecidae* families, in addition to the identification of 228,450 species-specific retrotransposons (Tang and Liang 2019), we also identified a total of 2,405 DNA transposons which are also species-specific that were not included in our report. As part of efforts to understand the mechanisms underlying these species-specific DNA transposons, we report in this study a new type of non-LTR retrotransposons derived from DNA transposons. These DNA transposons share sequence features characteristic of L1-based retrotransposons, and we therefore name them as retro-DNA, adding them as the fifth subclass of non-LTR retrotransposons after LINEs, SINEs, SVAs, and processed pseudogenes.

3.3 Materials and Methods

3.3.1 Sources of primate genome sequences

In this study, we chose to use ten primate genomes including human, among which eight genomes were included in our previous study for identifying species-specific MEs in primates (Tang and Liang 2019). These 10 primate species include human (GRCh38/UCSC hg38), chimpanzee (May 2016, CSAC Pan_troglodytes-3.0/panTro5), gorilla (Dec 2014, NCBI project 31265/gorGor4.1), orangutan (Jul. 2007, WUSTL version Pongo_albelii-2.0.2/ponAbe2), gibbon (Oct. 2012 GGSC Nleu3.0/nomLeu3.0), green monkey (Mar. 2014 VGC Chlorocebus_sabeus-1.1/chlSab2), crab-eating macaque (Jun. 2013 WashU Macaca_fascicularis_5.0/macFas5), rhesus monkey (Nov. 2015 BCM Mmul_8.0.1/rheMac8), baboon (Anubis) (Mar. 2012 Baylor Panu_2.0/papAnu2), and marmoset (Mar. 2009 WUGSC 3.2/calJac3). All genome sequences in fasta format and the RepeatMasker annotation files were downloaded from the UCSC genomic website (http://genome.ucsc.edu) onto our local servers for in-house analyses. We have used the most recent genome versions available on the UCSC genome browser site in all cases except for gorilla. For the gorilla genome, there is a newer version (Mar. 2016, GSMRT3/gorGor5) available, but it was not scaffolded into chromosomes, making it difficult to be used for our purpose.

3.3.2 LiftOver overchain file generation

A total of 90 liftOver chain files were needed for pair-wise comparisons of the ten genomes used in this study. These files contain the information linking the orthologous positions in a pair of genomes based on lastZ alignment (Harris 2007). Twenty-two of these were available and downloaded from the UCSC genome browser site, and another 34 liftOver chain files were generated using a modified version of UCSC pipeline RunLastzChain (http://genome.ucsc.edu) from a previous study (Tang and Liang 2019). The remaining 36 liftOver chain files were newly generated using the same pipeline.

3.3.3 Identification of DNA transposons with both insertion and pre-integration allele

3.3.3.1 Pre-processing of DNA transposon

The starting list of DNA transposons in each primate genome was obtained based on the RepeatMasker ME annotation data from the UCSC website (https://genome.ucsc.edu). As previously described, we performed a pre-processing to integrate the ME fragments annotated by RepeatMasker back to ME sequences representing the original transposition events (Tang et al. 2018).

3.3.3.2 Identification of DNA transposons with both insertion and pre-integration allele

We modified a previously reported bioinformatics comparative genomics approach (Tang et al. 2018) to identify diallelic DNA transposons (da-DNAs) that have the presence of both the insertion and pre-integration alleles in the ten primate genomes. Briefly, this pipeline uses a robust multi-way computational comparative genomics approach to determine the presence or absence status of a ME among a group of genomes by using both the whole genome alignment-based liftOver tool and the local sequence alignment-based BLAT tool (Hinrichs et al. 2006; Kent 2002). The sequences of a DNA transposon at the insertion site and its two flanking regions in a genome were compared to the sequences of the orthologous regions available in all other genomes. If a DNA transposon is absent from the orthologous regions of any of the other nine genomes not due to the existence of a sequence gap, it is selected as a potential candidate for a da-DNA subject to further analyses.

3.3.4 Identification of retro-DNA elements

3.3.4.1 Identification of target site duplications (TSDs) and terminal inverted repeats (TIRs)

For the candidate entries from the previous step, using in-house PERL scripts as described previously (Tang et al. 2018), we performed identification of the TSDs for the da-DNA entries from the above step. Additionally, we have modified our scripts to identify the TIRs, which are the hallmarks of all cut-and-paste transposons except for *Helitrons* (Feschotte and Pritham 2007). da-DNA entries without identifiable TSDs or TSD length < 8 bp, as well as entries with identifiable TIRs, were excluded from further analysis. The 8 bp TSD length cutoff was chosen based on our observation for human-specific retrotransposons that 95% of identified TSDs are at least 8 bp long (Tang et al. 2018). Additionally, we have used *MiteFinderII*, a tool designed to identify MIMEs (Hu et al. 2018) to verify that none of our candidate entries contain TIRs.

3.3.4.2 Filtering against retrotransposon transducitons

To ensure the presence of a DNA transposon is a result of active transposition, not a passive result of other events, such as retrotransposition-mediated transductions, we mapped the candidate entries against the known retrotransposons in the ten primate genomes based on their genomic positions. Specifically, the sequences of candidates from the last step were mapped back onto the host genome using BLAT, followed by removing all entries located within 50 bp to a retrotransposon (excluding entries inserted into a retrotransposon), as such entries could be a result of retrotransposition-based transduction. All entries left at this point were considered candidates of "retro-DNA" for being retrotransposons derived from DNA transposons but apparently using a retrotransposition mechanism.

3.3.4.3 Identification of poly(A) tails

For each candidate retro-DNA element, we retrieved the 10 bp sequence from the 3' end of the positive-strand (by the DNA transposon consensus sequence). If the sequence contains 6 or more "A"s, the entry is considered to have a polyA tail.

3.3.5 Clustering retro-DNA elements to eliminate redundancy

The retro-DNA candidates identified from the last step in the ten primate genomes were subject to a round of "all-against-all" sequence similarity search using BLAT with the sequences of the retro-DNA plus 100 bp of the flanking region on each side. Entries with 95% or higher sequence similarity across the entirety of the sequences were identified as one orthologous cluster, representing one retro-DNA insertion event during the evolution of these primates.

3.3.6 Estimating the timeline for retro-DNA insertions

A phylogenetic tree of the ten primate genomes plus the marmoset genome as the outgroup was obtained from the TimeTree database (http://www.timetree.org)(Hedges et al. 2006). The treeview program (Page 1996) was used to display the organismal phylogenetic tree.

We then added the numbers of non-redundant retro-DNA elements onto the nodes and branches of this tree based on their presence or absence in the specific genomes or lineages.

3.3.7 Multiple sequence alignment of retro-DNA elements and parent sites

We performed multiple sequence alignment for a few selected retro-DNA entries, including their parent sites. We first collected retro-DNA sequences including 100 bp on both flankings, as well as the orthologous sequences of the parent sites from the rest of primate genomes using the online version of MUSCLE (MUltiple Sequence Comparison by Log-Expectation)(Madeira et al. 2019) from European Bioinformatics Institute website (https://www.ebi.ac.uk/Tools/msa/muscle/).

3.3.8 Expression analysis of retro-DNA elements and their parent sites

RNA-Seq data for the blood and the generic (mixed) samples from chimpanzee, gorilla, crab-eating macaque, rhesus and baboon were retrieved from the Non-Human Primate Reference Transcriptome Resource (NHPRTR)(Pipes et al. 2013) for expression analysis of the retro-DNA elements and their parent copies. We also collected data for six human transcriptomes (Shin et al. 2014) and two green monkey transcriptomes (Jasinska et al. 2013; Jasinska et al. 2017). Tophat2 (version 2.1.1) was used to align the RNA-seq reads to the reference primate genomes (Kim et al. 2013). Reads mapped to the retro-DNA or parent copies regions were retrieved in fasta format and aligned back to the reference genome using the NCBI BLASTn to ensure that each RNA-seq read was only assigned to only one genomic location based on the perfect match, and they were used to calculate the Fragments Per Kilobase of transcript per Million reads (fpkm) values for each DNA transposon entry using an in-house Perl script.

3.3.9 Data analysis

The data analysis and figure plotting were performed using a combination of Linux shell scripting, R, and Microsoft Excel. The computational analysis was mostly performed on Compute Canada high-performance computing facilities (http://computecanada.ca).

3.4 Results

3.4.1 Overall profiles of DNA transposons and lineage-specific retro-DNAs in the ten primate genomes

To identify the retro-DNA events in the primate genomes, we first identified in the ten primate genomes the da-DNAs that represent DNA transposons with both the insertion allele and pre-integration allele identifiable in the set of the primate genomes. These DNA transposons were likely to be the results of relatively recent transposition events with a low level of sequence divergence to permit accurate identification of TSDs and TIRs. The starting lists of DNA transposons were based on the RepeatMasker annotation subject to a consolidation process to ensure the accuracy in identifying DNA transposons with both insertion and pre-integration alleles as well as their TSDs. As shown in Table 3.1, the raw number of DNA transposons in the primate genomes ranged from 392,937 in the marmoset as the lowest to 510,250 in the chimpanzee genome as the highest, averaging at 459,521/genome. After integration, the counts dropped ~18%, leading to an average of 376,720 DNA transposons per genome, and they contributed to a total of ~98 Mbp or ~3.6% of these primate genomes on averages (Table 3.1). While the numbers and percentages of DNA transposons in these genomes were similar overall with the variation falling within 10% of the average (data not shown), visible differences were also observed based on the integrated DNA entries with marmoset showing the least counts at 324,248 (3.2%) and chimpanzee showing the highest at 421,580 (3.8%). Various factors could have impacted the DNA transposons numbers in these genomes, which include, but are not limited to, the differences for versions of RepeatMasker and the ME reference sequences used for ME annotation, the quality of genome assembly, and evolution history of the primate genomes.

Table 3.1 Summary of DNA transposons in the 10 primate genomes

Genomes	Raw counts	integrated counts	% count reduction	total size (bp)	% genome	full-length count	% full-length	diallelic DNA counts
Human	483,994	399,590	17	102,664,356	3.5	119,368	29.9	25,933
Chimpanzee	510,250	421,580	17	107,832,154	3.8	119,265	28.3	28,273
Gorilla	503,480	418,454	17	106,573,049	3.8	117,263	28.0	27,386
Orangutan	429,467	347,471	19	93,420,030	3.4	113,425	32.6	23,923
Gibbon	438,800	363,738	17	93,531,426	3.6	108,334	29.8	24,206
Crab-eating macaque	443,909	359,802	19	94,910,440	3.5	109,444	30.4	26,218
Rhesus	486,991	401,546	18	102,546,356	3.7	111,558	27.8	28,149
Baboon	459,662	369,684	20	97,943,467	3.7	109,523	29.6	25,844
Green monkey	445,724	361,048	19	95,097,218	3.5	108,139	30.0	26,252
Marmoset	392,937	324,288	17	83,220,943	3.2	91,946	28.4	34,901
Average	459,521	376,720	18	97,773,944	3.6	110,827	29.5	27,109

*: full-length is defined as >=90% of consensus

Using a multi-way comparative genomics approach modified from our previous analysis of human-specific MEs (Tang et al. 2018), we identified a total of 271,085 da-DNAs in the ten primate genomes (Table 3.1). Specifically, for each da-DNA, we require the presence of a pre-integration allele in at least one of the nine remaining genomes.

As shown in Table 3.1, the number of da-DNAs varies from 23,923 in the orangutan genome as the lowest to 34,901 in the marmoset as the highest, averaging at 27,109 for the ten genomes. The largest number of da-DNAs in the marmoset is expected as marmoset has the largest evolutionary distance from the remaining primate species. Notable differences were also seen between genomes with a mutually closest evolutionary relationship in the group, making the numbers directly comparable for the paired genomes. For example, between the human and chimpanzee pair, the chimpanzee genome has more than 10% of da-DNAs than the human genome (28,273 vs. 25,933), while between the two macaques, the rhesus genome has ~10% more than the crab-eating macaque genome (28,149 vs. 26,218) (Table 3.1). Interestingly, this difference is much less than that for the species-specific non-LTRs, which shows crab-eating macaque genome having a much lower retrotransposition activity than the rhesus genome (Tang and Liang 2019). This may indicate that the majority of these da-DNAs were generated by a mechanism different from retrotransposition.

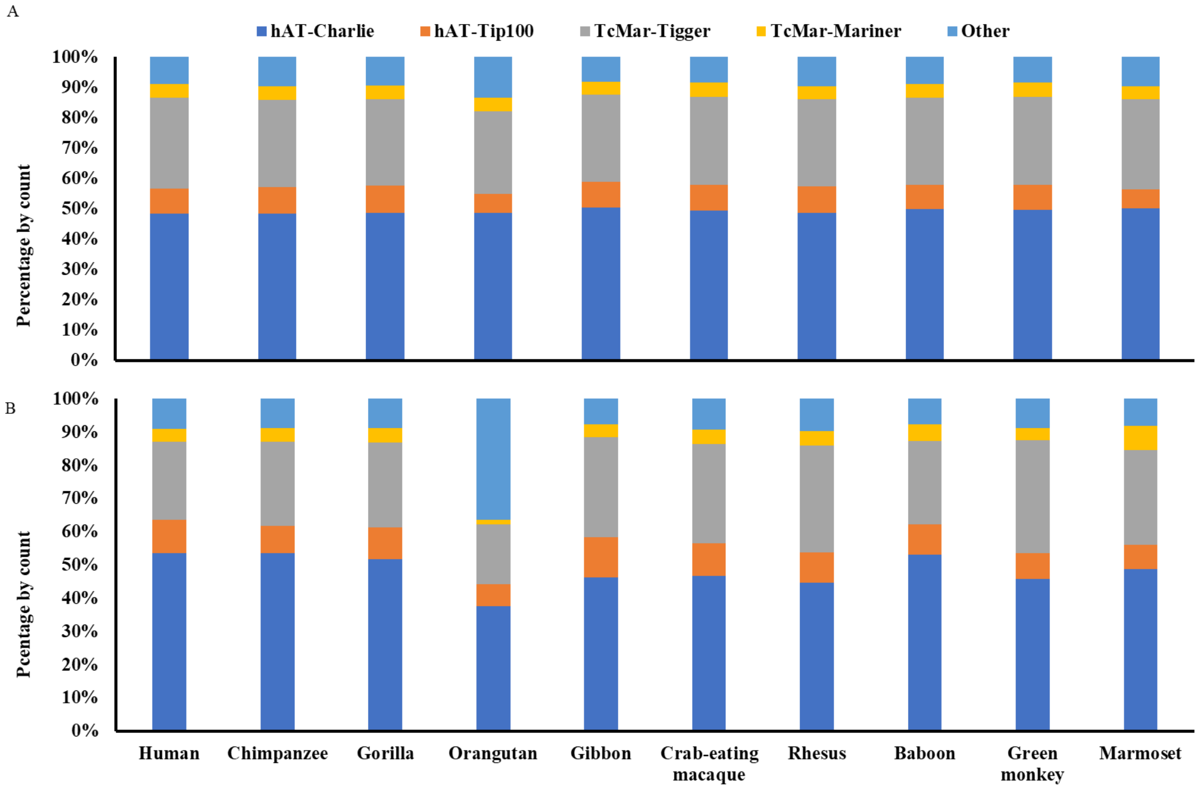


Figure 3.1 The distribution of diallelic DNA transposons and retro-DNA elements by family in the ten primate genomes

Stacked bar plots showing the family of the composition of diallelic DNA transposons (A) and retro-DNA elements (B) in each of the 10 primate genomes. The colour scheme in panel B is the same as in panel A.

By the composition in DNA transposon type, the majority of the da-DNAs belong to the hAT and TcMar superfamilies (Table 3.2, Fig. 3.1). The two hAT families, hAT-Charlie and hAT-Tip100, contributed to ~57% of da-DNAs in all genomes with the hAT-Charlie family alone contributing to ~50% of all da-DNAs. The two TcMar families, TcMar-Tigger and TcMar-Mariner, contribute ~30% of da-DNAs, while the remaining families contributed to ~10% of da-DNAs. This composition pattern was quite similar among all genomes, except for the orangutan genome, which has fewer da-DNAs from the TcMar-Trigger and the hAT-Tip100 families but more from families other than the hAT and TcMar superfamilies (Fig. 3.1).

Table 3.2 Composition of diallelic DNA transposons (da-DNAs) by family in the ten primate genomes

Genome	hAT-Charlie	hAT-Tip100	TcMar-Tigger	TcMar-Mariner	Other	Total
Human	12585	2128	7753	1189	2278	**25933**
Chimpanze e	13670	2504	8120	1288	2691	**28273**
Gorilla	13310	2466	7777	1268	2565	**27386**
Orangutan	11659	1501	6512	1074	3177	**23923**
Gibbon	12206	2065	6959	983	1993	**24206**
Crab-eating macaque	12966	2181	7604	1248	2219	**26218**
Rhesus	13690	2449	8069	1262	2679	**28149**
Baboon	12905	2058	7410	1203	2268	**25844**
Green monkey	13037	2140	7657	1211	2207	**26252**
Marmoset	17466	2239	10352	1459	3385	**34901**

3.4.2 Retro-DNA elements in the primate genomes show non-LTR retrotransposon sequence characteristics

While analyzing these da-DNAs in detail for understanding the possible mechanisms involved, we came across an unusual case of DNA transposon located at *chr4:146335052-146335253* of the human genome (GRCh38) as being a human-specific ME (Fig. 3.2A).

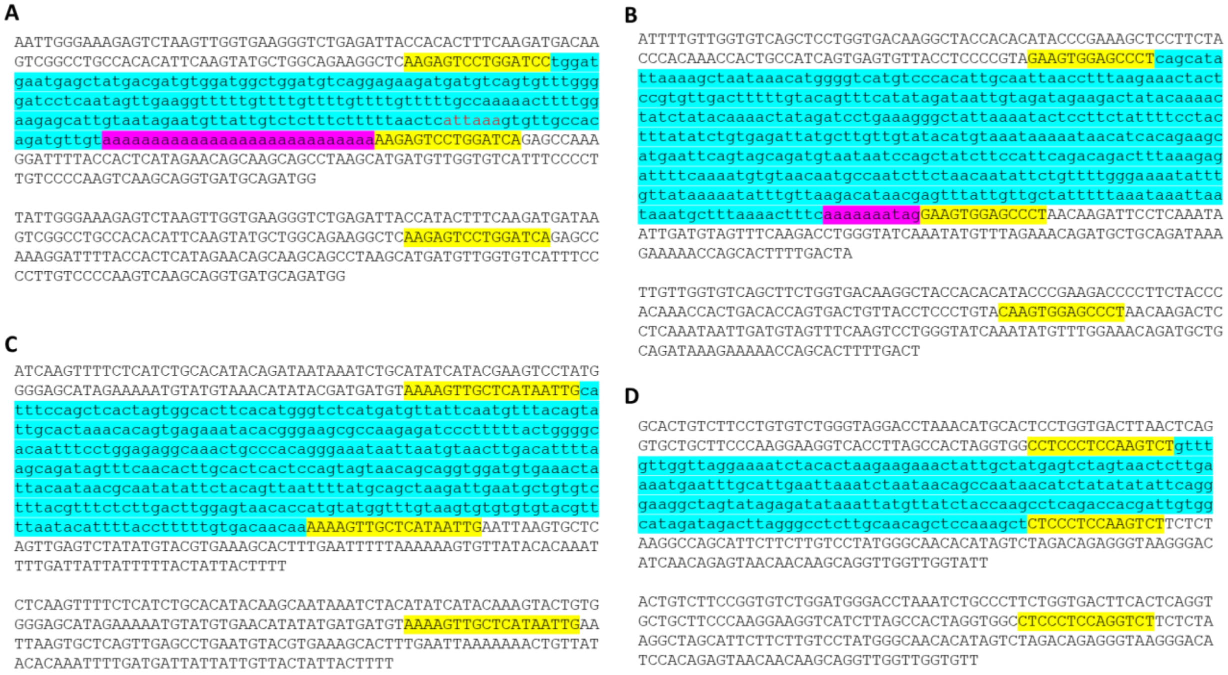


Figure 3.2 Examples of retro-DNA elements in different primate genomes

A. A retro-DNA element from the human genome (hg38_chr4:146335052-146335253) with the pre-integration allele from the chimpanzee genome (panTro5_chr4:38758218-38758438). B. A retro-DNA element from the green monkey genome (chlSab2_chr8:30005081-30005527) with the pre-integration allele from the gibbon genome (nomLeu3_chr8:37535028-37535236); C. A retro-DNA element located from the green monkey genome (chlSab2_chrX:73456937-73457324) with the pre-integration allele from the orangutan genome (ponAbe2_chrX:82896142-82896360). D. A retro-DNA element found from the human genome (hg38_chr4:38758216-38758442) with the pre-integration allele from the green monkey genome (chlSab2_chr27:11529606-11529817). In each panel, the sequence at the top is the insertion allele containing the retro-DNA element and the sequence at the bottom is the pre-integration allele without the retro-DNA element. The yellow boxes indicate TSDs, the blue boxes indicate the DNA transposon sequences, while the purple boxes indicate possible polyA tail sequences.

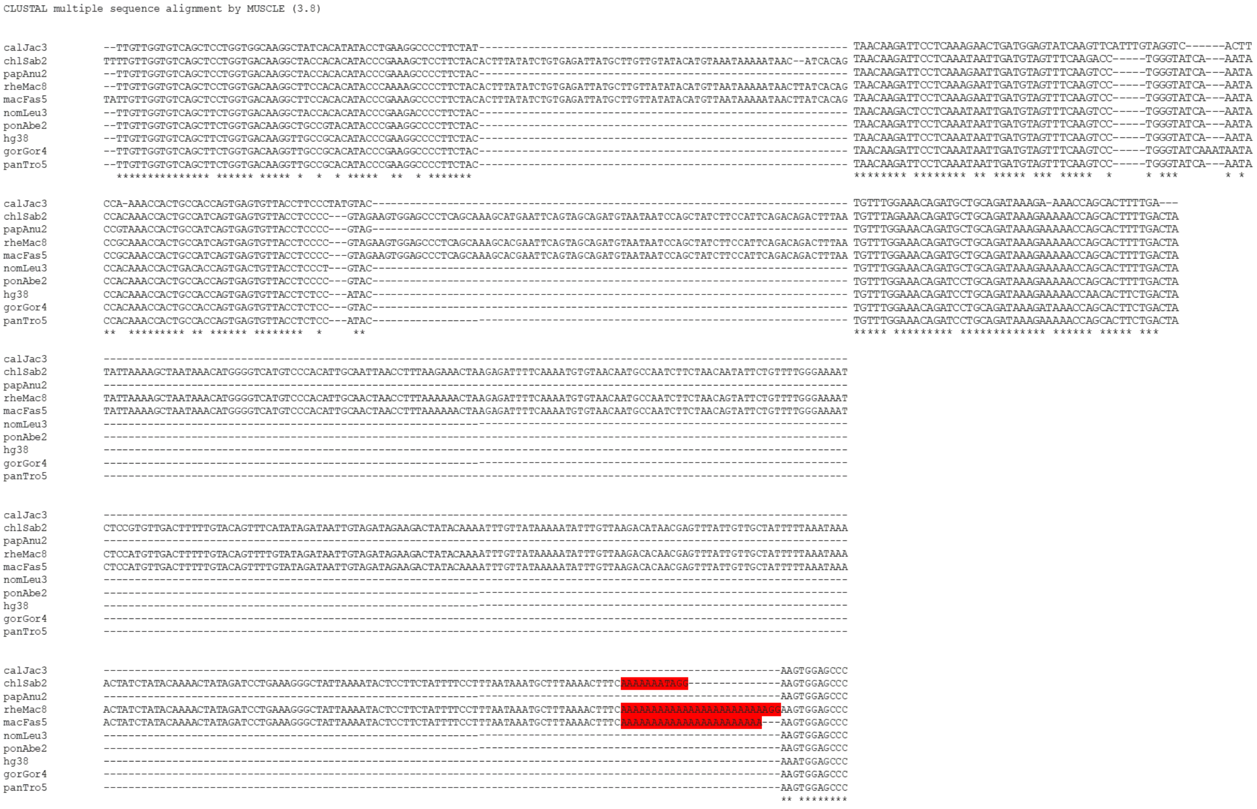
The 201 bp DNA transposon fragment was annotated by RepeatMasker as a *Tigger7* element from the *TcMar-Tigger* family. As shown in the multiple sequence alignments (Fig. 3.3) with its orthologous sequences from the other eight primate genomes except for marmoset (not identifiable), the *Tigger7* element was absent from the orthologous sites of all non-human primate genomes, making it as an authentic case of human-specific ME. More interestingly, this insertion has a 15 bp TSD “AAGAGTCCTGGATCC/AAGAGTCCTGGATCA” that was much longer than TSDs for DNA transposons, and it has no identifiable TIR typical of a DNA transposon (Fig. 3.2A). Furthermore, it has a 27 bp polyA tail at the 3’-end of the insertion sequence and a predicted polyadenylation signal “ATTAAA” before the polyA tail. Despite being part of a *Tigger7* DNA transposon sequence, all these features point to it being a non-LTR retrotransposon rather than a canonical *Tigger7* DNA transposon, which is expected to have TIRs and 2 bp (TA) TSDs. Because it is a DNA transposon sequence with the characteristics of a non-LTR retrotransposon, something not reported before, we named it as a retro-DNA element for being a retrotransposon-like element derived from a DNA transposon sequence.

```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)

nomLeu3         TATTGGGAAAGAGTCTAAGATGGTGAAGGGTCTGAGATTACCACACTTTCAAGATGACAA
panTro5         TATTGGGAAAGAGTCTAAGTTGGTGAAGGGTCTGAGATTACCATACTTTCAAGATGATAA
hg38            AATTGGGAAAGAGTCTAAGTTGGTGAAGGGTCTGAGATTACCACACTTTCAAGATGACAA
ponAbe2         TATTGGGAAAGAGTCTAAGATGGTGAAGGGTCTGAGATTACCACACTTTCAAGATGACAA
gorGor4         TATTGGGAAAGAGTCTAAGATGGTGAAGGGTCTGAGATTACCACACTTTCAAGATGACAA
chlSab2         TATTGGGAGAGAGTCTAAGATGGTGAAGGTTCTGAAATTACCACACTTTCAAGACGACAA
macFas5         TATTGGGAGAGAGTCTAAGATGGTGAAGGGTCTGAAATTACCACACTTTCAAGACGACAA
papAnu2         TATTGGGAGAGAGTCTAAGATGGTGAAGGGTCTGAAATTACCACACTTTCAAGACGACAA
rheMac8         TATTGGGAGAGAGTCTAAGATGGTGAAGGGTCTGAAATTACCACACTTTCAAGACGACAA
                 ******* ********** ********* ***** ******* ********** ** **

nomLeu3         GTCGGCCTGCCACACTTTCAAGTATGCTGGCAGAAGGCTC--------------------
panTro5         GTCGGCCTGCCACACATTCAAGTATGCTGGCAGAAGGCTC--------------------
hg38            GTCGGCCTGCCACACATTCAAGTATGCTGGCAGAAGGCTCAAGAGTCCTGGATCCTGGAT
ponAbe2         GTCTGCCTGCCACACATTCAAGTGTGCTGGCAGAAGGCTC--------------------
gorGor4         GTCGGCCTGCCACACATTCAAGTATGCTGGCAGAAGGCTC--------------------
chlSab2         GTTGGTCTGCCACACATTCAAGTATACTGGGAGAAGACTC--------------------
macFas5         GTTGGTCTGCCA----TTCAAGTATACTGGCAGAAGACTC--------------------
papAnu2         GTTGGTCTGCCA----TTCAAGTATACTGGCAGAAGACTC--------------------
rheMac8         GTTGGTCTGCCA----TTCAAGTATACTGGCAGAAGACTC--------------------
                **  * ******    ******* * **** ***** ***

nomLeu3         ------------------------------------------------------------
panTro5         ------------------------------------------------------------
hg38            GAATGAGCTATGACGATGTGGATGGCTGGATGTCAGGAGAAGATGATGTCAGTGTTTGGG
ponAbe2         ------------------------------------------------------------
gorGor4         ------------------------------------------------------------
chlSab2         ------------------------------------------------------------
macFas5         ------------------------------------------------------------
papAnu2         ------------------------------------------------------------
rheMac8         ------------------------------------------------------------

nomLeu3         ------------------------------------------------------------
panTro5         ------------------------------------------------------------
hg38            GATCCTCAATAGTTGAAGGTTTTTGTTTTGTTTTGTTTTGTTTTTGCCAAAAACTTTTGG
ponAbe2         ------------------------------------------------------------
gorGor4         ------------------------------------------------------------
chlSab2         ------------------------------------------------------------
macFas5         ------------------------------------------------------------
papAnu2         ------------------------------------------------------------
rheMac8         ------------------------------------------------------------

nomLeu3         ------------------------------------------------------------
panTro5         ------------------------------------------------------------
hg38            AAGAGCATTGTAATAGAATGTTATTGTCTCTTTCTTTTTAACTCATTAAAGTGTTGCCAC
ponAbe2         ------------------------------------------------------------
gorGor4         ------------------------------------------------------------
chlSab2         ------------------------------------------------------------
macFas5         ------------------------------------------------------------
papAnu2         ------------------------------------------------------------
rheMac8         ------------------------------------------------------------

nomLeu3         -------------------------------------AAAAGTCCTAGATCAGCGCCAAA
panTro5         -------------------------------------AAGAGTCCTGGATCAGAGCCAAA
hg38            AGATGTTGTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGAGTCCTGGATCAGAGCCAAA
ponAbe2         -------------------------------------AAGAGTCCTGGATCAGAGCCAAA
gorGor4         -------------------------------------AAGAGTCCTGGATCAGAGCCAAA
chlSab2         -------------------------------------AAGAGTCCTGGATCAGAGCC-AA
macFas5         -------------------------------------AAGAGTCCTGGATCACAGCCAAA
papAnu2         -------------------------------------AAGAGTCCTGGATCACAGCCAAA
rheMac8         -------------------------------------AAGAGTCCTGGATCACAGCCAAA
                                                     ** ****** *****  *** **

nomLeu3         GGATTTTACCACTCACAGAACAGCAAGCAGCCTAAGCATGATGTTGGTGTCATTTCCCCT
panTro5         GGATTTTACCACTCATAGAACAGCAAGCAGCCTAAGCATGATGTTGGTGTCATTTCCCCT
hg38            GGATTTTACCACTCATAGAACAGCAAGCAGCCTAAGCATGATGTTGGTGTCATTTCCCCT
ponAbe2         GGATTTTACCACTCATAGAACAGCAAGCAGCCTAAGCATGATGTTGGTGTCATTTCCCCT
gorGor4         GGATTTTACCACTCATAGAACAGCAAGCAGCCTAAGCATGATGTTGGTGTCATTTCCCCT
chlSab2         GGATTTTACCACTCACAGAACAGCAAGTAGCCTAAGCATGATGTTGGTGACACTTCCCCT
macFas5         GGATTTTACCACTCACAGAACAGCAAGCAGCCTAAGCATGATGTTGGTAACATTTCCCCT
papAnu2         GGATTTTACCACTCACAGAACAGCAAGCAGCCTAAGCATGATGTTGGTGACATTTCCTCT
rheMac8         GGATTTTACCACTCACAGAACAGCAAGCAGCCTAAGCATGATGTTGGTGACATTTCCCCT
                *************** *********** ********************  ** **** **

nomLeu3         TGTCCCCAAGTCAAGCAGGTGATGCAGATGG
panTro5         TGTCCCCAAGTCAAGCAGGTGATGCAGATGG
hg38            TGTCCCCAAGTCAAGCAGGTGATGCAGATGG
ponAbe2         TGTCCCCAAGTCAAGCAGGTGATGCAGATGG
gorGor4         TGTCCCCAAGTCAAGCAGGTGATGCAGATGG
chlSab2         TGTCCCCAAGTCAAGCAGGTGATGCAGATGG
macFas5         TGTCCCCAAGTCAAGCAGGTGATGCAGATGG
papAnu2         TGTCCCCAAGTCAAGCAGGTGATGCAGATGG
rheMac8         TGTCCCCAAGTCAAGCAGGTGATGCAGATGG
                *******************************
```

Figure 3.3 Multiple sequence alignment for a species-specific retro-DNA element

Multiple sequence alignment for the retro-DNA element located in the human genome (hg38_chr4:146335052-146335253) and the corresponding pre-integration sequences from eight other primate genomes. The pre-integration sequences from the marmoset genome are unavailable likely due to the high level of sequence divergence.

Following the discovery of this retro-DNA case, we searched the human genome and other genomes to identify more similar circumstances, as shown in Fig. 3.2. For instance, a 446 bp *Charlie1a* fragment from the *hAT-Charlie* family was identified as a retro-DNA element in three primate genomes (the green monkey, rhesus, and crab-eating macaque genomes with the locations being chlSab2_chr8:30005081-30005527, rheMac8_chr8:31992158-31992606, and macFas5_chr8:32527581-32528029, respectively). This entry has a 13 bp TSD "GAAGTGGAGCCCT" and has no TIRs (Fig. 3.2B).



```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)


calJac3         --TTGTTGGTGTCAGCTCCTGGTGGCAAGGCTATCACATATACCTGAAGGCCCCTTCTAT------------------------------------------------------------ TAACAAGATTCCTCAAAGAACTGATGGAGTATCAAGTTCATTTGTAGGTC------ACTT
chlSab2         TTTTGTTGGTGTCAGCTCCTGGTGACAAGGCTACCACACATACCCGAAAGCTCCTTCTACACTTTATATCTGTGAGATTATGCTTGTTGTATACATGTAAATAAAAATAAC--ATCACAG TAACAAGATTCCTCAAATAATTGATGTAGTTTCAAGACC-----TGGGTATCA---AATA
papAnu2         --TTGTTGGTGTCAGCTCCTGGTGACAAGGCTACCACACATACCCGAAAGCCCCTTCTAC------------------------------------------------------------ TAACAAGATTCCTCAAAGAATTGATGTAGTTTCAAGTCC-----TGGGTATCA---AATA
rheMac8         --TTGTTGGTGTCAGCTCCTGGTGACAAGGCTTCCACACATACCCAAAAGCCCCTTCTACACTTTATATCTGTGAGATTATGCTTGTTATATACATGTTAATAAAAATAACTTATCACAG TAACAAGATTCCTCAAAGAATTGATGTAGTTTCAAGTCC-----TGGGTATCA---AATA
macFas5         TATTGTTGGTGTCAGCTCCTGGTGACAAGGCTTCCACACATACCCGAAAGCCCCTTCTACACTTTATATCTGTGAGATTATGCTTGTTATATACATGTTAATAAAAATAACTTATCACAG TAACAAGATTCCTCAAAGAATTGATGTAGTTTCAAGTCC-----TGGGTATCA---AATA
nomLeu3         --TTGTTGGTGTCAGCTTCTGGTGACAAGGCTACCACACATACCCGAAGACCCCTTCTAC------------------------------------------------------------ TAACAAGACTCCTCAAATAATTGATGTAGTTTCAAGTCC-----TGGGTATCA---AATA
ponAbe2         --TTGTTGGTGTCAGCTTCTGGTGACAAGGCTGCCGTACATACCCGAAGGCCCCTTCTAC------------------------------------------------------------ TAACAAGATTCCTCAAATAATTGATGTAGTTTCAAGTCC-----TGGGTATCA---AATA
hg38            --TTGTTGGTGTCAGCTTCTGGTGACAAGGTTGCCGCACATACCCGAAGGCCCCTTCTAC------------------------------------------------------------ TAACAAGATTCCTCAAATAATTGATGTAGTTTCAAGTCC-----TGGGTATCA---AATA
gorGor4         --TTGTTGGTGTCAGCTTCTGGTGACAAGGTTGCCGCACATACCCGAAGGCCCCTTCTAC------------------------------------------------------------ TAACAAGATTCCTCAAATAATTGATGTAGTTTCAAGTCC-----TGGGTATCAAATAATA
panTro5         --TTGTTGGTGTCAGCTTCTGGTGACAAGGTTGCCGCACATACCCGAAGGCCCCTTCTAC------------------------------------------------------------ TAACAAGATTCCTCAAATAATTGATGTAGTTTCAAGTCC-----TGGGTATCA---AATA
                  *************** ****** ***** *  *  * *****  **  * *******                                                              ******** ******** ** ***** *** *****  *     * ***       * *

calJac3         CCA-AAACCACTGCCACCAGTGAGTGTTACCTTCCCTATGTAC----------------------------------------------------------------------------- TGTTTGGAAACAGATGCTGCAGATAAAGA-AAACCAGCACTTTTGA---
chlSab2         CCACAAACCACTGCCATCAGTGAGTGTTACCTCCCC---GTAGAAGTGGAGCCCTCAGCAAAGCATGAATTCAGTAGCAGATGTAATAATCCAGCTATCTTCCATTCAGACAGACTTTAA TGTTTAGAAACAGATGCTGCAGATAAAGAAAAACCAGCACTTTTGACTA
papAnu2         CCGTAAACCACTGCCATCAGTGAGTGTTACCTCCCC---GTAG----------------------------------------------------------------------------- TGTTTGGAAACAGATGCTGCAGATAAAGAAAAACCAGCACTTTTGACTA
rheMac8         CCGCAAACCACTGCCATCAGTGAGTGTTACCTCCCC---GTAGAAGTGGAGCCCTCAGCAAAGCACGAATTCAGTAGCAGATGTAATAATCCAGCTATCTTCCATTCAGACAGACTTTAA TGTTTGGAAACAGATGCTGCAGATAAAGAAAAACCAGCACTTTTGACTA
macFas5         CCGCAAACCACTGCCATCAGTGAGTGTTACCTCCCC---GTAGAAGTGGAGCCCTCAGCAAAGCACGAATTCAGTAGCAGATGTAATAATCCAGCTATCTTCCATTCAGACAGACTTTAA TGTTTGGAAACAGATGCTGCAGATAAAGAAAAACCAGCACTTTTGACTA
nomLeu3         CCACAAACCACTGACACCAGTGACTGTTACCTCCCT---GTAC----------------------------------------------------------------------------- TGTTTGGAAACAGATGCTGCAGATAAAGAAAAACCAGCACTTTTGACTA
ponAbe2         CCACAAACCACTGCCACCAGTGAGTGTTACCTCCCC---GTAC----------------------------------------------------------------------------- TGTTTGGAAACAGATCCTGCAGATAAAGAAAAACCAGCACTTTTGACTA
hg38            CCACAAACCACTGCCACCAGTGAGTGTTACCTCTCC---ATAC----------------------------------------------------------------------------- TGTTTGGAAACAGATCCTGCAGATAAAGAAAAACCAACACTTCTGACTA
gorGor4         CCACAAACCACTGCCACCAGTGAGTGTTACCTCTCC---GTAC----------------------------------------------------------------------------- TGTTTGGAAACAGATCCTGCAGATAAAGATAAACCAGCACTTCTGACTA
panTro5         CCACAAACCACTGCCACCAGTGAGTGTTACCTCTCC---ATAC----------------------------------------------------------------------------- TGTTTGGAAACAGATCCTGCAGATAAAGAAAAACCAGCACTTCTGACTA
                **  ********* ** ****** ********  *     **                                                                               ***** ********* ************* ****** ***** ***

calJac3         ------------------------------------------------------------------------------------------------------------------------
chlSab2         TATTAAAAGCTAATAAACATGGGGTCATGTCCCACATTGCAATTAACCTTTAAGAAACTAAGAGATTTTCAAAATGTGTAACAATGCCAATCTTCTAACAATATTCTGTTTTGGGAAAAT
papAnu2         ------------------------------------------------------------------------------------------------------------------------
rheMac8         TATTAAAAGCTAATAAACATGGGGTCATGTCCCACATTGCAACTAACCTTTAAAAAACTAAGAGATTTTCAAAATGTGTAACAATGCCAATCTTCTAACAGTATTCTGTTTTGGGAAAAT
macFas5         TATTAAAAGCTAATAAACATGGGGTCATGTCCCACATTGCAACTAACCTTTAAAAAACTAAGAGATTTTCAAAATGTGTAACAATGCCAATCTTCTAACAGTATTCTGTTTTGGGAAAAT
nomLeu3         ------------------------------------------------------------------------------------------------------------------------
ponAbe2         ------------------------------------------------------------------------------------------------------------------------
hg38            ------------------------------------------------------------------------------------------------------------------------
gorGor4         ------------------------------------------------------------------------------------------------------------------------
panTro5         ------------------------------------------------------------------------------------------------------------------------


calJac3         ------------------------------------------------------------------------------------------------------------------------
chlSab2         CTCCGTGTTGACTTTTTGTACAGTTTCATATAGATAATTGTAGATAGAAGACTATACAAAATTTGTTATAAAAATATTTGTTAAGACATAACGAGTTTATTGTTGCTATTTTTAAATAAA
papAnu2         ------------------------------------------------------------------------------------------------------------------------
rheMac8         CTCCATGTTGACTTTTTGTACAGTTTTGTATAGATAATTGTAGATAGAAGACTATACAAAATTTGTTATAAAAATATTTGTTAAGACACAACGAGTTTATTGTTGCTATTTTTAAATAAA
macFas5         CTCCATGTTGACTTTTTGTACAGTTTTGTATAGATAATTGTAGATAGAAGACTATACAAAATTTGTTATAAAAATATTTGTTAAGACACAACGAGTTTATTGTTGCTATTTTTAAATAAA
nomLeu3         ------------------------------------------------------------------------------------------------------------------------
ponAbe2         ------------------------------------------------------------------------------------------------------------------------
hg38            ------------------------------------------------------------------------------------------------------------------------
gorGor4         ------------------------------------------------------------------------------------------------------------------------
panTro5         ------------------------------------------------------------------------------------------------------------------------


calJac3         -------------------------------------------------------------------------------------------------------------AAGTGGAGCCC
chlSab2         ACTATCTATACAAAACTATAGATCCTGAAAGGGCTATTAAAATACTCCTTCTATTTTCCTTTAATAAATGCTTTAAAACTTTCAAAAAAATAGG---------------AAGTGGAGCCC
papAnu2         -------------------------------------------------------------------------------------------------------------AAGTGGAGCCC
rheMac8         ACTATCTATACAAAACTATAGATCCTGAAAGGGCTATTAAAATACTCCTTCTATTTTCCTTTAATAAATGCTTTAAAACTTTCAAAAAAAAAAAAAAAAAAAAAAAAAGGAAGTGGAGCCC
macFas5         ACTATCTATACAAAACTATAGATCCTGAAAGGGCTATTAAAATACTCCTTCTATTTTCCTTTAATAAATGCTTTAAAACTTTCAAAAAAAAAAAAAAAAAAAAAAA---AAGTGGAGCCC
nomLeu3         -------------------------------------------------------------------------------------------------------------AAGTGGAGCCC
ponAbe2         -------------------------------------------------------------------------------------------------------------AAGTGGAGCCC
hg38            -------------------------------------------------------------------------------------------------------------AAATGGAGCCC
gorGor4         -------------------------------------------------------------------------------------------------------------AAGTGGAGCCC
panTro5         -------------------------------------------------------------------------------------------------------------AAGTGGAGCCC
                                                                                                                             ** ********
```

Figure 3.4 Multiple sequence alignment for a lineage-specific retro-DNA element

The multiple sequence alignment for a retro-DNA element located in the green monkey, crab-eating macaque and rhesus genomes (chlSab2_chr8:30005081-30005527/rheMac8_chr8:31992158-31992606/macFas5_chr8:32527581-32528029) and their pre-integration sites from 7 other primate genomes. The red boxes represent possible polyA tail with various lengths in different genomes.

As shown in Fig. 3.4, the 446 bp *Charlie1a* fragment was absent in the orthologous regions of the remaining seven primate genomes, which could be explained as a lineage-specific insertion event that happened in the last common ancestor of green monkey, rhesus, and crab-eating macaque. Also, it appears that this retro-DNA sequence in these three genomes had been subject to variations as the polyA tails have different lengths, indicating its relatively older age as a lineage-specific da-DNA in comparison to the species-specific element as an example in Fig. 3.3.

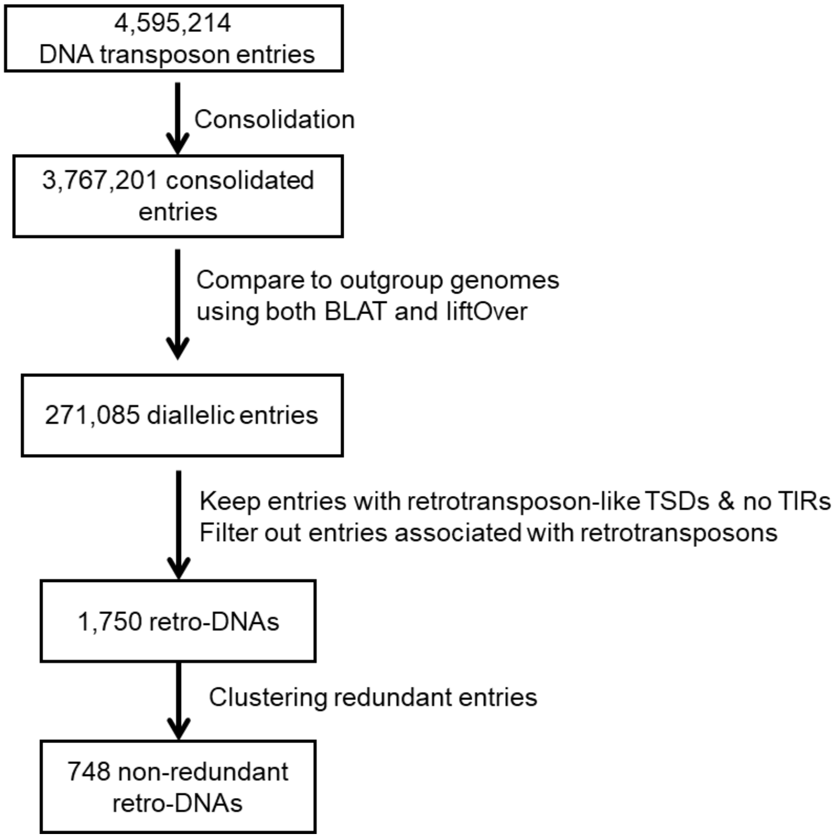

4,595,214
DNA transposon entries
Consolidation
3,767,201 consolidated entries
Compare to outgroup genomes using both BLAT and liftOver
271,085 diallelic entries
Keep entries with retrotransposon-like TSDs & no TIRs
Filter out entries associated with retrotransposons
1,750 retro-DNAs
Clustering redundant entries
748 non-redundant retro-DNAs

Figure 3.5 Flow chart for identification of retro-DNA elements.

By requiring the presence of longer TSDs (≥8 bp) and the absence of TIRs**,** we identified a total of 1,750 retro-DNA elements from the da-DNAs using a workflow shown in Fig. 3.5. By classification, these retro-DNAs consist of 847, 478, 156, 74, and 195 entries from the hAT-Charlie, TcMar-Tigger, hAT-Tip100, TcMar-Mariner, and other families, respectively. As seen in Table 3.3, these 1,750 retro-DNA elements cover all ten genomes and can be clustered into 748 unique retro-DNA insertion events. It is worth noting that our list of retro-DNA elements may suffer a certain level of false-negatives and false-positives due to the uses of a set of criteria which may not be very optimal and challenges associated with the analysis of transposable elements as well as the deficiencies of the resources, including the quality of the reference genomes and the RepeatMasker annotation, especially for the non-human primates as discussed in our recent study(Tang and Liang 2019).

Table 3.3 Retro-DNA elements in the 10 primate genomes

Genome	hAT-Charlie	hAT-Tip100	TcMar-Tigger	TcMar-Mariner	Other	All Retro-DNA elements
Human	100	19	44	7	17	**187**
Chimpanzee	108	17	51	8	18	**202**
Gorilla	99	18	49	8	17	**191**
Orangutan	58	10	28	2	56	**154**
Gibbon	72	19	47	6	12	**156**
Crab-eating macaque	76	16	49	7	15	**163**
Rhesus	79	16	57	8	17	**177**
Baboon	76	13	36	7	11	**143**
Green monkey	78	13	58	6	15	**170**
Marmoset	101	15	59	15	17	**207**
Total	847	156	478	74	195	**1,750**
NR total	317	63	221	34	113	**748**

By sequence length, these 748 retro-DNA elements averaged at 209 bp (±190 bp) and represented only part of DNA transposons in all cases, covering ~21% of their consensus sequences (Table 3.4). While the consensus sequences for DNA transposon families differ in length significantly from 380 bp for TcMar-Mariner to 1506 bp for hAT-Tip100, the average length of retro-DNA elements seems to be relatively consistent among the families, ranging from 122 bp for TcMar-Mariner to 251 bp for TcMar-Tigger. Nevertheless, in general, the retro-DNA elements from the longer families do have a longer average length than those from the shorter families, despite those from the longer families have a lower proportion of their consensus sequences (Table 3.4).

Table 3.4 The composition of retro-DNA elements by family and the size information

DNA transposon Family	copy number	% of all retro-DNA elements	Average size (bp)	Std (bp)	Average consensus length (bp)	% of consensus
hAT-Charlie	317	42.4	190	110	515	37
TcMar-Tigger	221	29.5	251	256	1,162	22
hAT-Tip100	63	8.4	200	209	1,506	13
TcMar-Mariner	34	4.5	122	115	380	32
Other	113	15.1	210	200	1,053	20
Total	**748**	**100**	**209**	**190**	**923**	**21**

Additionally, we examined whether there are any hotspots on the consensus sequences that are used as the sources of these retro-DNA elements. By using the retro-DNA entries from the Tigger1 DNA transposon subfamily, the largest subfamily of retro-DNA elements containing 41 non-redundant entries, we generated a frequency plot showing the usage of the consensus sequences by the retro-DNAs. As illustrated in Fig. 3.6, while all regions of the consensus sequence were used by the 41 retro-DNA elements, the frequency ranges from 2.4% to 29.3%, showing a few regions including those from ~1310-1440 bp and from ~1840-2240 bp of consensus sequence had been used more frequently than the rest of the regions.

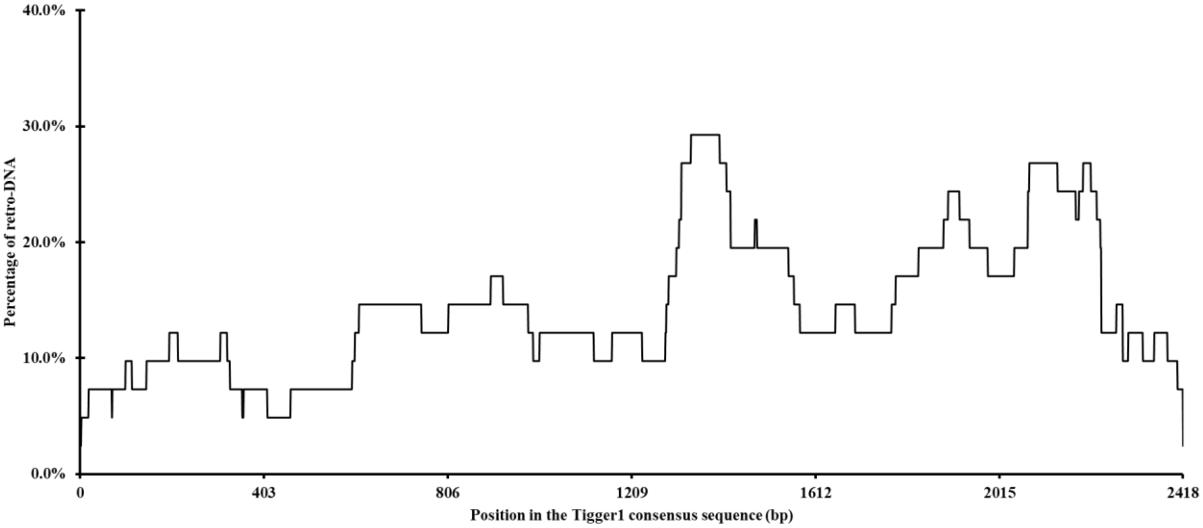


Figure 3.6 The frequency of the Tigger1 subfamily DNA transposon consensus sequence used for retro-DNA sequences

The plot is based on the data for a total of 41 non-redundant retro-DNA entries from the Tigger1 subfamily.

As shown in Table 3.5, we have identified a total of 176 non-redundant retro-DNA entries carrying a potential polyA tail. We speculate that the relatively low percentage (23.5%) of entries with a polyA tail might be partially due to post-insertion mutations in the polyA sequences, which are more prone to random mutations than other regions. For these retro-DNA insertion events, we further examined the sequence motifs at the insertion sites and the TSD length distribution pattern.

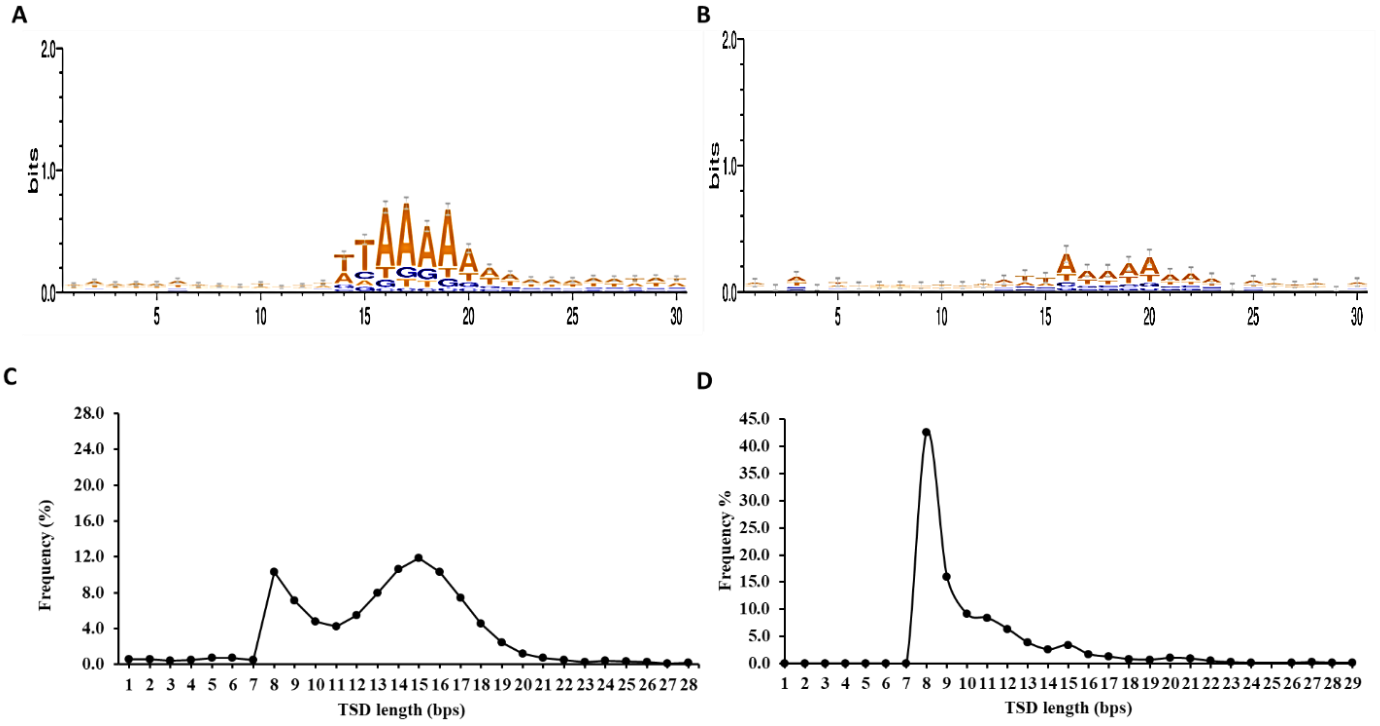


Figure 3.7 Sequence motifs of pre-integration sites and TSD length distribution pattern for retro-DNA elements

A. Sequence motif logos for human-specific L1s at the integration sites, adopted from authors' publication (Tang et al., 2018). B. Sequence motif logos for retro-DNA elements at the integration sites. C. A line plot showing the distribution of TSD length for human-specific L1s, adopted from the authors' publication (Tang et al., 2018). D. A line plot showing the distribution of TSD length for retro-DNA elements.

As shown in Fig. 3.7A, a sequence motif of 'TT/AAAA', which was the same as the motif for *Alu*s, L1s, and SVAs (Goodier 2016; Tang et al. 2018; Wang et al. 2006), was clearly seen despite the signal being much weaker. This, nevertheless, is strongly suggesting the use of the L1-based non-LTR retrotransposition TPRT mechanism (Cost and Boeke 1998; Jurka 1997). As further support, the TSD length distribution peaked at 8 bp, which is similar to a secondary peak for the TSD lengths of human-specific L1s, despite being shorter than the peak around 15 bp, which is the major peak seen for non-LTR retrotransposons (Tang et al. 2018).

Table 3.5 The number of retro-DNA elements with identifiable polyA tails in the 10 primate genomes

Genome	# of retro-DNA elements	# of retro-DNA elements with polyA tail*	polyA entry ratio (%)
Human	187	32	17.1
Chimpanzee	202	36	17.8
Gorilla	191	32	16.8
Orangutan	154	32	20.8
Gibbon	156	23	14.7
Green monkey	170	29	17.1
Crab-eating macaque	163	22	13.5
Rhesus	177	27	15.3
Baboon	143	22	15.4
Marmoset	207	49	23.7
Total (non-redundant)	**748**	**176**	**23.5**
Average	**175**	**30.4**	**17.2**

*: polyA tail is defined as >=6 "A"s within the 10 bp of 3' end sequence

3.4.3 The patterns of retro-DNA elements and their parent sites in genome distribution and expression

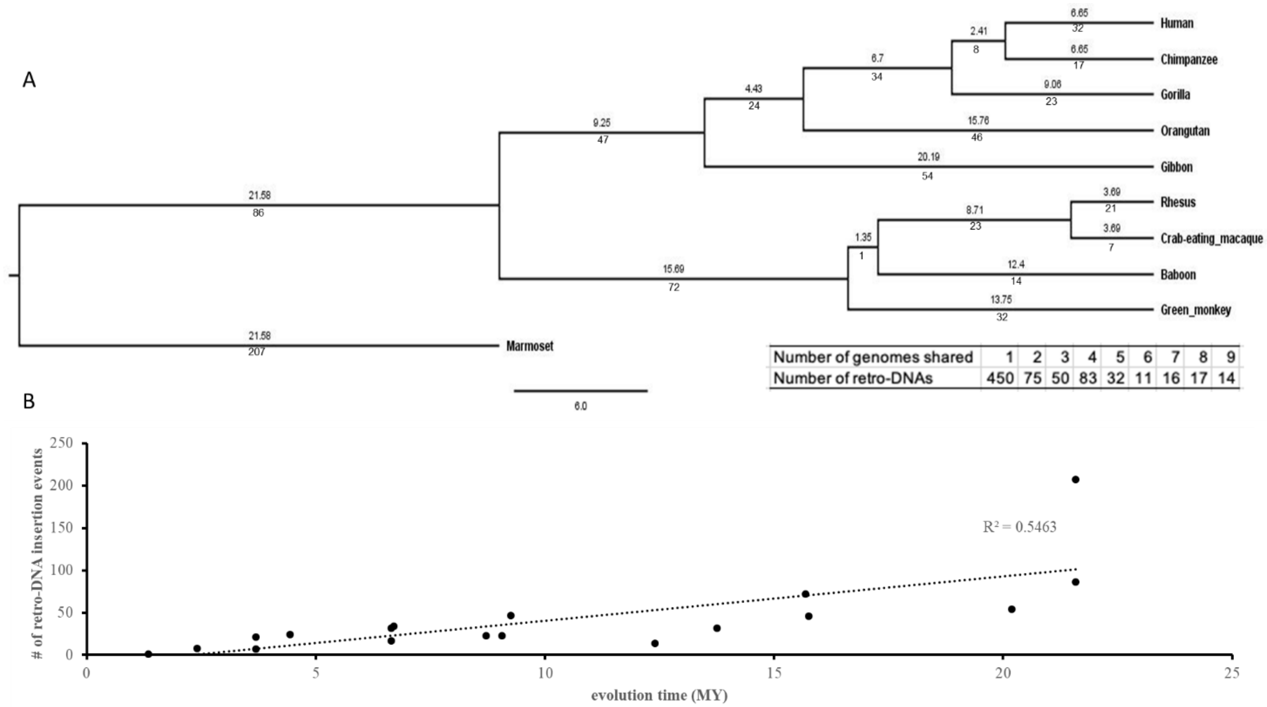
To assess the potential functional impact of these retro-DNA elements, we examined their gene context based on the Ensemble gene annotation for these genomes (Release 95 for all genomes except Release 90 and 91 were for baboon and marmoset, respectively)(Zerbino et al. 2018). A total of 698 retro-DNA elements, representing ~40% of the 1,750 retro-DNA elements, are located within different genic regions, including non-coding RNAs, intron regions, untranslated regions, and promoter regions for 734 transcripts representing 414 unique genes (Table 3.6 & Appendix II). The majority of these retro-DNA elements were located within the intron regions (699/734), while 27 entries were inserted into promoter regions and the untranslated regions. The presence of these retro-DNA elements in the genic regions provides the potential for them to impact gene regulation or splicing.

Table 3.6 The numbers of retro-DNA elements located in the genic regions in the 10 primate genomes

Genic region*	NR	Promoter	5' UTR	3' UTR	Intron	Total
Human	4	9	0	1	114	128
Chimpanzee	1	5	1	0	78	85
Gorilla	1	2	0	0	70	73
Orangutan	1	1	0	0	60	62
Gibbon	0	0	0	0	61	61
Crab-eating macaque	0	1	1	0	62	64
Rhesus	1	1	0	0	67	69
Baboon	0	0	0	0	42	42
Green monkey	0	0	0	0	53	53
Marmoset	0	3	0	2	92	97
Total	**8**	**22**	**2**	**3**	**699**	**734**

*, NR: non-coding RNA; UTR: untranslated region

We also examined the timeline of these retro-DNA insertion events by mapping them onto a phylogenetic tree of these primates based on the data in the TimeTree database (http://www.timetree.org)(Hedges et al. 2006).



Number of genomes shared	1	2	3	4	5	6	7	8	9
Number of retro-DNAs	450	75	50	83	32	11	16	17	14

Figure 3.8 Timeline of the retro-DNA elements generation during the evolution of the ten primate genomes

A. A rooted phylogenetic tree of the ten primate genomes from the TimeTree database(http://www.timetree.org/). The numeric values below each branch represent the number of retro-DNA insertion events that happened during the corresponding period of primate evolution. The numeric value above each branch represents the millions of years (My) for that branch. The insert table below the tree shows the distribution of the retro-DNA elements by the degree of conservation among the genomes as measured by the number of genomes owning a retro-DNA element. B. A scatter plot between the number of retro-DNA insertion events and the evolutionary time based on the data in panel A. The trend line shows that the number of retro-DNA insertion events is positively correlated with the relative evolutionary distance with R^2 = 0.5463.

As shown in Fig. 3.8A (insert), 450 or 60.2% of these retro-DNA elements appeared to be species-specific by being uniquely present in only one genome, while another 295 were found in multiple genomes as being lineage-specific. On average, a retro-DNA element is shared among two genomes, suggesting an average age older than the species-specific MEs reported in our earlier studies (Tang and Liang 2019). Some manual corrections were made for placing the lineage-specific retro-DNA elements on the phylogenetic tree. For example, 7 retro-DNA elements are found to be shared between human and gorilla but not in chimpanzee, and we decided to include these entries on the branch common for these three genomes. We argue that this manual correction is necessary as retro-DNA identification can suffer false negatives from the insufficient sequence assembly quality for the non-human primate genomes. As shown in Fig. 3.8B, the number of retro-DNA insertion events seems to show a positive linear correlation with the relative evolutionary spans of the species and lineages ($R^2 = 0.5463$), suggesting that these retro-DNA insertion events have occurred at a relatively consistent rate during primate evolution.

Further, we identified the potential parent sites for these retro-DNA entries by performing a sequence similarity search using these retro-DNA elements as query sequences against each primate genome. For each retro-DNA element, the best non-self-match was selected as its potential parent site. As shown in Appendix III, we have identified a total of 715 potential parent sites for the 1,750 retro-DNA entries. This converts to 325 non-redundant entries of the 748 unique retro-DNAs. The failure in finding the parent copies for the remaining entries could be due to the loss of the parent copy as a result of genomic rearrangements or due to incomplete coverage of the genome sequences. As for retro-DNAs, we have examined the gene context for these potential parent sites. As shown in Table 3.7, 351 (or 49.1%) of these redundant potential retro-DNA parent sites were located within 410 different genic regions with 13 coding sequences, 19 non-coding RNAs, 112 promoter regions, one 5' UTR, four 3'UTRs and 274 intron regions, which collectively represent 371 unique genes or transcripts. In these cases, the transcripts of these potential parent sites, likely as part of the transcripts of their host genes or their splicing products, might have had the chance to hijacked L1's TPRT machinery as in the case of processed pseudogenes to generate the retro-DNA. The ratio of genic entries (49.1%) is higher for parent sites than that for retro-DNA (~40%), and the implication is discussed in later sections.

Table 3.7 Parent sites associated with genic regions in the ten primate genomes

Genic region*	CDS	NR	Promoter	5' UTR	3' UTR	Intron	Total
Human	0	10	34	0	1	60	**105**
Chimpanzee	0	2	17	0	0	37	**56**
Gorilla	0	2	12	0	0	36	**50**
Orangutan	0	2	6	0	0	19	**27**
Gibbon	0	1	9	0	0	21	**31**
Green monkey	12	0	12	0	1	25	**38**
Crab-eating macaque	0	0	5	0	1	21	**27**
Rhesus	0	2	6	0	0	25	**33**
Baboon	0	0	7	0	0	16	**23**
Marmoset	1	0	4	1	1	14	**20**
Total	**13**	**19**	**112**	**1**	**4**	**274**	**410**

*, NR: non-coding RNA; UTR: untranslated region

We have also examined the expression level of retro-DNA elements and their potential parent sites using RNA-seq data from Non-Human Primate Reference TRanscriptome (NHPRTR) dataset (Pipes et al. 2013) and two other studies (Jasinska et al. 2017; Shin et al. 2014) to see if any of these entries have any transcriptional activities in the current primate genomes. For this, we collected a total of 21 transcriptomes for seven primates, excluding orangutan, gibbon, and marmoset, for which no transcriptome data is available. To minimize the false positives due to the high sequence similarity among members in the same family, we included only the reads with the perfect match to the retro-DNA elements or their parent site regions in the primate genomes and with each read to be used only once for calculating the expression level. However, since the specific transcriptome sequences can diverge from the corresponding reference genomes due to intra-species variations, we believe this process has inevitably introduced false negatives in the results and therefore lead to an underestimation of the retro-DNA and parent site expression level. As seen in Table 3.8 and Appendix IV, 966 loci from the 1,750 retro-DNA elements and 715 parent sites in these seven primate genomes were shown to have a certain level of expression ranging in fragments per kilobase exon per million reads (fpkm) value from 0.0003 to 27.30.

Table 3.8 The numbers of expressed retro-DNA elements and parent sites in 21 primate transcriptomes

Species	**# of RNA-seq sets**	**retro-DNA elements**			**parent sites**		
		entry #	**expressed**	**%**	**entry #**	**expressed**	**%**
Human	6	187	93	49.7	98	57	58.2
Chimpanzee	2	202	99	49.0	101	67	66.3
Gorilla	1	191	55	28.8	99	42	42.4
Rhesus	4	177	97	54.8	64	46	71.9
Crab-eating macaque	4	163	115	70.6	63	55	87.3
Baboon	2	143	68	47.6	53	34	64.2
Green monkey	2	170	90	52.9	62	48	77.4
Total	**19**	**1063**	**527**	**49.6**	**478**	**301**	**63.0**

We further investigated the relationship between retro-DNA elements and their parent sites based on their expression levels. Specifically, three human testis transcriptome samples (SRR2040581, SRR2040582, SRR2040583) retrieved from the NCBI SRA (Sequence Read Archive; https://www.ncbi.nlm.nih.gov/sra) were used for the analysis of expression level of the retro-DNA/parent pair sites.

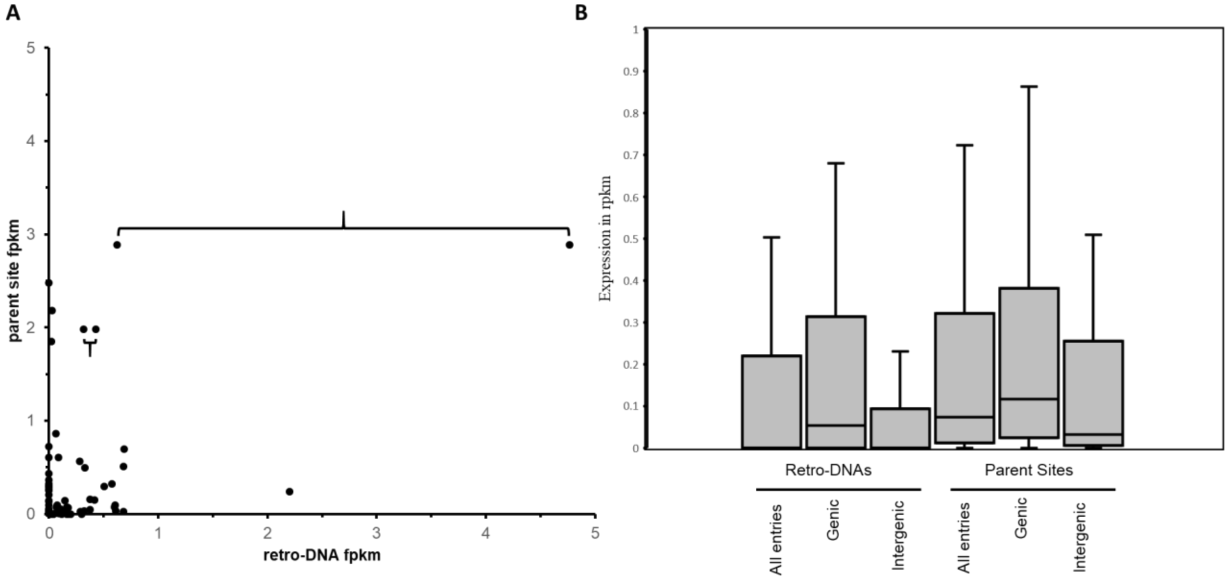


Figure 3.9 The expression (fpkm) of retro-DNA elements and their parent sites in three human testis transcriptomes

A. A scatter plot based on 66 retro-DNA/parent site pairs which show a certain level of expression (fpkm > 0) for the retro-DNA element and/or parent site. The retro-DNA elements connected by brackets indicate entries with possible same parent copy. B. Box plots showing the expression levels of the 66 retro-DNA elements and parent sites divided into genic and intergenic groups. Expression data was based on the average fpkm value in the three human testis transcriptomes.

As shown in Fig. 3.9A, a total of 66 retro-DNA/parent site pairs were shown to have a certain level of expression (fpkm > 0) for either the retro-DNA element or the parent site among the three human testis samples. Notably, within these 66 pairs, 57 parent sites were being expressed (fpkm > 0) compared to only 42 entries for retro-DNA elements (Appendix III & IV, Fig. 3.9A). This might be due to the fact that the generation of a retro-DNA element requires the expression of its parent site, while a retro-DNA element itself may not be expressive depending on its landing location, which is random. Therefore, a higher ratio of transcriptionally active sites can be expected for the parent sites than the progeny (retro-DNA) sites. More interestingly, we noticed that the two parent sites responsible for multiple retro-DNA entries were shown to have the highest levels of expression among the parent sites (Fig. 3.9A, labeled in brackets). This may suggest that the expression level of the parent sites is positively correlated to their potential in generating retro-DNA and that there seems to be no relationship between the expression levels of the parent sites and the progeny retro-DNA sites. Furthermore, the ongoing expression of the parent sites suggests that they can potentially generate more retro-DNA elements in the future.

We have also examined the expression levels of retro-DNA elements and parent sites in the three human testis transcriptomes based on positions in gene contest. As shown in Fig. 3.9B, the average fpkm values of the parent sites are always higher than the average fpkm values of the retro-DNA entries as a whole group or divided into genic and intergenic regions. In addition, the entries located within genic regions showed higher expression levels than the ones located outside the genic regions for both retro-DNA elements and the parent sites (Fig. 3.9B), suggesting that entries located in the genic region may have more opportunities to be expressed passively as part of the host gene expression. This difference is larger for retro-DNA elements than for the parent sites, likely because parent sites have to be expressed in order to be able to generate new copies, and their expression might be driven by other factors other than the host gene expression if located outside genes. None of these differences are statistically significant, likely due to the small sample size.

3.5 Discussions

3.5.1 Retro-DNA as a new type of retrotransposons derived from DNA transposons

In this study, we focused our attention on a small number of species-specific DNA transposons identified in primate genomes using our computational comparative genomics pipelines, which revealed unprecedented numbers of species-specific retrotransposons in the human genome and seven other genomes (Tang et al. 2018; Tang and Liang 2019). Unlike for the retrotransposons, for which the ongoing activity during primate evolution and in the current genomes have been well established (Goodier 2016; Jordan et al. 2018; Tang and Liang 2019), the presence of species-specific DNA transposons in these primate genomes present a puzzle, which cannot be answered by existing literature. This is because DNA transposons are thought to have become inactive about 37 My ago (Feschotte and Pritham 2007; Pace Ii and Feschotte 2007), meaning that no canonical DNA transposition activity could have existed during the evolution of these primate genomes. In trying to understand the mechanism underlying these mystery species-specific DNA transposon insertions, we started examining the sequence features by manual analysis and spotted a few interesting entries as exemplified by the case shown in Fig. 3.2A, which shows evident characteristics of non-LTR retrotransposons by having longer TSDs and presence of a polyA tail, while lacking TIRs that are the hallmark of new DNA transposon insertions. The remaining cases, shown in Fig. 3.2, have the same non-LTR features, but not necessarily have the typical polyA tail. For their non-LTR retrotransposon characteristics, we name them as “retro-DNA” as retrotransposons derived from DNA transposons. We then performed a systematic analysis to look for more of such “retro-DNA” cases.

For this, we expanded from the strict species-specific DNA transposons, which are defined as those present in only one of the primate genomes (Tang et al. 2018; Tang and Liang 2019), to diallelic DNA transposons or da-DNAs, which are defined as those with a pre-integration site (i.e., the orthologous region without the DNA transposon) present in at least one of the ten genomes. We obtained a total of 271,085 da-DNAs, and from these, we then specifically searched for retro-DNA cases that have long TSDs (>=8 bp) and the absences of the TIRs (Fig. 3.5). This led to the identification of 1,750 of retro-DNA cases, which represent 748 unique events, covering all ten primate genomes with the over half being species-specific and the other half being lineage-specific at different levels on the evolution tree (Fig. 3.8A). Our results indicate that the presence of retro-DNA elements is not limited to the human genome but can be found in all ten primate genomes included in this study and along different stages of primate evolution. Furthermore, these retro-DNA elements are not limited to one DNA transposon family but cover all major DNA transposon families, suggesting that the existence of such “retro-DNA”, a novel type of retrotransposons, is not just for rare incidental cases, but is rather the product of a consistent mechanism shared by all these primate genomes.

3.5.2 The likely mechanism underlying the generation of retro-DNA elements

Several lines of evidence from our results guided us to propose that these retro-DNA elements were the products of the L1-based TPRT machinery, similar for the known non-autonomous non-LTR retrotransposons, i.e., SINEs, SVAs and processed pseudogenes (Cost and Boeke 1998; Jurka 1997; Mita and Boeke 2016; Tang et al. 2018; Xing et al. 2006).

The major pieces of evidence include the presence of the “TT/AAAA” sequence motif at the insertion sites and the long TSDs. As seen in Fig. 3.7A, the integration sites for the 748 retro-DNA elements display a core sequence motif of “TT/AAAA”, which is identical to the insertion site sequence motif observed for non-LTR retrotransposons in the human genome (Jurka 1997; Tang et al. 2018; Wang et al. 2006). The length distribution of the TSDs for these retro-DNA elements, as shown in Fig. 3.7B, shows a dominant peak at 8 bp, which is much longer than that of TSDs typically found for DNA transposons (2 bp) and is similar to the secondary peak of TSD length observed for the human-specific L1s (Tang et al. 2018).

As additional pieces of evidence supporting our proposal, the presence of parent sites in the same genome for the majority of the retro-DNA elements (325/748 or 43.5%) indicates their use of a “copy-and-paste” mechanism rather than the “cut-and-paste” mechanism used by canonical DNA transposons. Furthermore, the presence of a polyA tail in many (176/748 or 23.5%) of these retro-DNA elements provides further support for the use of L1-based TPRT mechanism.

It is worth noting that, as described above, while there is sufficient similarity in sequence features between these retro-DNA elements and the known non-LTR retrotransposons to consider these retro-DNA elements as a new type of non-LTR retrotransposons, unique aspects of these retro-DNA elements are also visible. These include the missing of the major TSD length peak at 15 bp for non-LTR retrotransposons, the low percentage of entries with a polyA tail, and the weaker signal of the sequence motif, “TT/AAAA”, at the integration sites. All of these characteristics might be contributed by the relative older average age of these retro-DNA elements as shown by the relatively high percentage (298/748 or ~40%) of being lineage-specific (Fig. 3.2A) than the non-LTR retrotransposons used in most prior studies for analysis of the non-

LTR integration site sequence motifs (Cost and Boeke 1998; Jurka 1997; Mita and Boeke 2016; Tang et al. 2018; Xing et al. 2006). In other words, the older age of the retro-DNA elements leads to higher sequence divergence, which in turn lowers the sensitivity for detecting all of these sequence features. An additional reason for the weaker signal in the integration sequence motif for the retro-DNA elements could be due to the small sample size. In the meantime, it is also possible that these differences might suggest that some differences exist in the detailed retrotransposition process of these DNA transposons, likely the interaction between the retro-DNA transcripts and the ORF1 and ORF2 proteins.

3.5.3 The relative retro-DNA activity during primate evolution

In comparison with the other types of non-autonomous non-LTR retrotransposons, including Alus, SVAs, and processed pseudogenes, in the primates (Bennett et al. 2008; Goodier 2016; Lander et al. 2001; Tang and Liang 2019), the number of retro-DNA element per genome is much lower, averaging below 200 per genome (Table 3.5). This number is even much lower than that of processed pseudogenes, which repesent the smallest class of non-LTR retrotransposons with 10,190 in the human genome (Tutar 2012). We reason that the small copy numbers of retro-DNA elements may be mainly attributed to one factor, which is the lack of internal promoters to drive their own transcription, leading to an overall low level of their transcripts available for retrotransposition. This is in agreement with the fact that there is no clear hotspot in the DNA transposon consensus sequences used in generating retro-DNA elements, as shown in Fig. 3.6 for Tigger 1. Should there be internal promoters driving the transcription, we would expect to observe one or more clear dominant peaks in the frequency of the regions used for retro-DNA elements.

Without the ability in driving their own transcription, the only way for DNA transposons to get transcribed is to be become part of genes and get transcribed as a part of the host gene transcripts. If this is how retro-DNA elements were generated, then we would expect to see a high percentage of retro-DNA elements having their parent sites located in the genic regions, more specifically in the transcribed regions, i.e. exons and introns. By examining the gene context, 351 of the 715 parent sites (49.0%) for the retro-DNA elements locate in 371 unique genes or transcripts in the ten primate genomes. This ratio is higher than the ratio of all DNA transposons in the genic regions (39%, detailed data not shown).

Following the same rationale, we would expect that on average the parent sites should have a higher expression level than that of retro-DNA elements since the parent sites were selected to be biased for those in the genic regions, while the location of the retro-DNA is more or less random, leading to the latter having a relatively higher proportion in non-genic and non-transcriptive regions. As shown in Fig. 3.9A, among the 66 retro-DNA/parent site pairs, 57 pairs have parent sites with fpkm > 0 compared to only 42 expressed entries for retro-DNA elements. Additionally, we identified two parent sites, which are potentially responsible for generating multiple retro-DNA entries, and they showed the highest levels of expression among the parent sites (Fig. 3.9A). By comparing the expression levels of all parent sites with that of retro-DNA in the human genome, we can see an overall higher expression for the parent sites (Fig. 3.9B), and this is also true when comparing the two groups of sites in the genic and intergenic regions (Fig. 3.9B). Furthermore, the expression level of parent sites in the genic regions is much higher than their counterparts in the intergenic regions as expected (Fig. 3.9B).

The use of ten primate genomes, representing several lineages in the group spanning a certain time span in primate evolution, allowed us to examine whether there is any positive correlation between the length of the evolutionary span and the number of retro-DNA insertion events. As shown in Fig. 3.8B, a positive correlation between the two ($R^2 = 0.5463$) is observed, suggesting that the generation of retro-DNA is relatively steady during the evolution of this group of primates. Furthermore, the fact that many of the retro-DNA parent sites, as well as 966 of the 1773 (~54.5%) retro-DNA elements show certain levels of expression in the seven primate transcriptomes (Table 3.8 & Appendix IV), suggests the possible ongoing activity of retro-DNA generation from the parent sites and retro-DNAs.

3.5.4 The functional potential of retro-DNA elements

As shown in Table 3.6 & Appendix II, 698 retro-DNA entries (redundant across genomes), representing ~40% of the 1,750 retro-DNA elements, are located within 734 different genic regions, including non-coding RNAs, introns, untranslated regions, and promoter regions for 414 unique genes in the ten primate genomes. Furthermore, 8 entries of these retro-DNA elements contribute to part of transcripts, despite none found to be in CDS regions. Therefore, we could speculate that these retro-DNA elements may have a potential impact on gene function via the regulation of transcription and splicing, similar to what has been shown for retrotransposons (Ward et al. 2013).

3.6 Conclusions and future perspectives

In this study, through a comparative analysis of ten primate genomes including the human genome, we identified a new type of non-autonomous non-LTR retrotransposons, which derived from DNA transposon sequences. Named as “retro-DNA”, these elements represent the 5th type of non-LTR retrotransposons after LINE, SINE, SVA, and processed pseudogene, very likely using the same L1-based TPRT mechanism. The generation of these retro-DNAs serve to propagate DNA transposon sequences in the absence of the canonical DNA transposon activity. Despite their relatively small number, they do contribute to the genetic diversity among primate species along with other non-LTR retrotransposons. Furthermore, the discovery of these retro-DNA elements suggests that the L1 TRPT machinery may have been used by more diverse types of RNA transcripts than we currently know. Future work may include the verification of the retrotransposition activity of these retro-DNA elements using *in vitro* and in *vivo* assays and expanding of similar analysis to other types of expressive DNA sequences, such as non-coding RNA genes. In addition, analysis to identify the mechanisms underlying the remaining majority of the diallelic DNA transposons would also be very interesting.

Chapter 4 Alu master copies serve as the drivers of differential SINE transposition in recent primate genomes

(The content of this chapter is mostly copied from a manuscript in preparation for publication: “Alu master copies serve as the drivers of differential SINE transposition in recent primate genomes” Wanxiangfu Tang and Ping Liang with some minor changes for table formats and figure reorganization (renumbered after combining with supplementary figures).

The candidate is the main author of this manuscript and was responsible for generating most of the data included in the manuscript. The manuscript was drafted by the candidate and edited by the corresponding author, Dr. Liang, to its current form.)

4.1 Abstract

Alu elements, averaging ~300 bp in length, are a family of primate-specific short intersperse nucleotide elements (SINEs) with more than one million copies contributing to ~11% of primate genomes. Despite mostly being shared among primates, our recent study revealed highly differential recent Alu transposition among the genomes of primates from *Hominidae* and *Cercopithecidae* families. To understand the underlying mechanism, we analyzed six primate genomes and revealed species- and lineage-specific Alu profile exclusively defined by AluY composition. Among all Alus from the 6 genomes, we identified 5,401 Alu master copies with 99% being from the AluY subfamily. The numbers of Alu master copies are positively correlated to the number of AluY elements in the genomes with the baboon genome having the largest number of most recent Alu master copies at high activity, while the crab-eating macaque genome having a low number of Alu master copies with low activity. Furthermore, the expression level of Alu master copies is positively correlated with their transposition activity. Our results support the concept that Alu transposition in the primate genome is driven by a small number of master copies, the number and relative activity of which contribute to the differential Alu transposition in recent primate genomes.

4.2 Introduction

Mobile elements (MEs) constitute more than 50% of the current primate genomes (Carbone et al. 2014; Chimpanzee Sequencing and Analysis 2005; Cordaux and Batzer 2009; Deininger et al. 2003; Lander et al. 2001; Locke et al. 2011; Rhesus Macaque Genome Sequencing and Analysis et al. 2007; Tang and Liang 2019), and they are known to play important roles in genome evolution and gene function through a variety of mechanisms (Batzer and Deininger 2002; Cordaux and Batzer 2009; Goodier 2016; Kazazian 2000; Kazazian and Goodier 2002; Kazazian 2004). Continuous ME transposition has served as an important mechanism in generating inter- and intra-species genomic diversity (Callinan et al. 2005; Steely et al. 2018; Tang et al. 2018; Tang and Liang 2019; Wang et al. 2006).

In the primate genomes, the retrotransposons, Long INterspersed Elements (LINEs) and Short INterspersed Elements (SINEs), represent the major types of MEs, both using LINE-1's Target Prime Reverse Transcription (TPRT) machinery for retrotransposition (Goodier 2016; Jurka 1997; Kazazian and Goodier 2002). Alu elements are a family of primate-specific SINEs averaging ~300 bp in length. Despite being one of the shortest ME families, Alus have shown great success by contributing to ~11% primate genomes, second only to the ~18% contributed by the LINE-1s (L1s). The higher percentage of Alu elements is primarily due to their extremely high copy numbers in the primate genomes, averaging ~1.2 million copies per genome (Tang and Liang 2019). A typical Alu element consists of two diverged dimers, which are believed to have derived from the 7SL RNA gene during the very early stage of primate evolution (Quentin 1992). The 3' end of Alu usually has a long consecutive "A"s, which is referred to as the poly-A tail. Alu elements carry an internal RNA polymerase II promoter and have the ability to express them as RNAs, which can hijack the L1's TPRT machinery for retrotransposition (Batzer and

Deininger 2002; Deininger 2011; Wang and Huang 2014). However, despite both using the same mechanism, there seems to be a difference between L1 retrotransposition and Alu transposition; while L1s depend on both ORF1p and ORF2p to retrotranspose, Alus only rely on the presence of ORF2p protein to retrotranspose (Dewannieux et al. 2003; Goodier 2016; Wallace et al. 2008). According to the data in L1base (Penzkofer et al. 2017) and our recent observation (Nanayakkara et al., manuscript in preparation), the primate genomes usually have only a handful of function L1s with the ability to code intact ORF1p and ORF2p proteins. Meanwhile, there are more L1s with intact ORF2p coding capacity but have lost the capability to encode intact ORF1p protein, as ORF1 seems to subject to a higher level of mutations (Penzkofer et al. 2017). This may explain the fact that Alus have been able to amplify in most of the primate genomes more efficiently than L1s by having much larger copy numbers (Tang and Liang 2019). In particular, the baboon genome showed an extremely high level of Alu expansion in its recent evolution through a large number of highly active baboon-specific Alu subfamilies (Jordan et al. 2018; Steely et al. 2018; Tang and Liang 2019).

There are three major Alu subfamilies in the current primate genomes: the AluJ, AluS and AluY (Bennett et al. 2008; Jurka and Smith 1988). The AluJ is the oldest subfamily, averaging roughly a quarter-million entries per genome among the primates which contribute to ~2.4% of the primate genomes based on our recent analysis (Tang and Liang 2019). The AluS subfamily, which has been highly active during the early stages of the primate evolution, has contributed to ~6.4% of the primate genomes (Batzer et al. 1996; Tang and Liang 2019). The 3rd major Alu subfamily is the Y subfamily, which is responsible for the more recent Alu amplification in the primate genomes (Bennett et al. 2008). While the numbers of AluS and AluJ elements are relatively constant in individual primate genomes as the old and shared Alu elements, the number of AluY elements varies substantially, especially between the ape and monkey groups. Additionally, only a small number of AluY copies can be found in lower primate genomes such as marmoset, indicating that AluY amplification did not start until the late stage of the primate evolution (Tang and Liang 2019).

It has been reported that despite many Alu elements are capable of making Alu transcripts, there are only a few “master copies” which can generate new Alu copies in the primate genomes (Cordaux et al. 2004; Han et al. 2005; Shen et al. 1991). Shen et al. first surveyed Alu elements and proposed a model, in which all Alu elements were made either from a single master gene or from a series of sequential master genes (Shen et al. 1991). A later study by Johnson and Brookfield suggests that it is highly unlikely, if not at all impossible, for any Alu subfamilies to have accumulated from the activity of one single master copy (Johnson and Brookfield 2006). Additionally, some new Alu subfamilies in the human genome are shown to have originated from a small number of young active Alu elements characterized by accumulated new mutations and serving as master copies (Ahmed et al. 2013). Therefore, it is currently believed that the Alu elements in the primate genomes have been contributed by multiple master copies (Brookfield and Johnson 2006; Johnson and Brookfield 2006; Tachida 1996).

Recently, studies on Alu elements have focused on the young and active members, which can contribute to genetic variation among individuals in the primate genomes, particularly the human genome (Ahmed et al. 2013; Battilana et al. 2006; Bennett et al. 2008; Jordan et al. 2018; Mills et al. 2006; Wang et al. 2006). For example, some studies have shown that Alu elements are still active in the human genome, namely the young AluYa5, Yb8 and Yb9, and they are responsible for generating population-specific or polymorphic Alu elements (Ahmed et al. 2013; Battilana et al. 2006; Liang and Tang 2012; Mamedov et al. 2010; Stewart et al. 2011). Similar research has been extended to the non-human primates. For example, Steely et al. recently ascertained 28,114 baboon-specific Alu elements by comparing the genomic sequences from baboon to both the rhesus macaque and the human genome (Steely et al. 2018). More recently, our group generated a comprehensive compilation of MEs that are uniquely present in the human genome by making use of the most recent genome sequences for human and many other closely related primates and a robust multi-way comparative genomic approach. It led to the identification of 14,870 human-specific MEs, among which 8,817 are Alu elements (Tang et al. 2018).

In our recent comparative analysis of species-specific MEs (SS-MEs) in eight primates from the Hominidae and the Cercopithecidae families, we identified a total of 148,753 species-specific Alus (SS-Alus), showing striking differences in recent Alu transposition among these primate genomes (Tang and Liang 2019). For example, the baboon genome has the largest number of SS-Alus, and the human genome has the largest number of Alu subfamilies with the highest activity among all recent primate genomes, while the crab-eating macaque genome has a sustained low transposition for all MEs, including Alu elements (Tang and Liang 2019). To better understand the differential Alu transposition in recent primate genomes and the driving factors behind them, we performed a focused analysis of Alu master copies by examining them both at the DNA sequence and transcriptome levels.

4.3 Materials and Methods

4.3.1 Sources of primate genome annotation files

In this study, we chose to use six primate genomes, which cover both the *Hominidae* and *Cercopithecidae* families based on the availability of genome and transcriptome data. These six primate genomes include human genome (GRCh38/UCSC hg38), chimpanzee genome (May 2016, CSAC Pan_troglodytes-3.0/panTro5), gorilla genome (Dec 2014, NCBI project 31265/gorGor4.1), rhesus monkey genome (Nov. 2015 BCM Mmul_8.0.1/rheMac8), crab-eating macaque genome (Jun. 2013 WashU Macaca_fascicularis_5.0/macFas5), and the baboon (Anubis) genome (Mar. 2012 Baylor Panu_2.0/papAnu2). The most updated version of gene annotation data for each primate genome was downloaded from Ensemble Release 95 except for baboon (only Release 90 available). Besides, the RepeatMasker annotation files were downloaded from the UCSC genomic website (http://genome.ucsc.edu) onto our local servers for in-house analyses.

4.3.2 Analysis of age and composition profiles of Alu elements in the primate genomes

To estimate the age of Alu elements, we compared their sequences against the corresponding subfamily consensus sequences and determined the divergence level. For each Alu element, the average number of substitutions per 100 bp (K) was calculated using the mismatch level p, which was provided in the RepeatMasker annotation file (Smit et al. 2004), according to the one-parameter Jukes-Cantor model ($K = -3/4\ln(1 - 4/3p)$)(Jukes and Cantor 1969). These Alus, divided into the three major Alu subfamilies (AluJ, AluS, and AluY) were then grouped into bins with incremental 1% of the substitution and the percentage of genomes for the Alus in each bin was calculated to show the Alu age and composition profile in the genomes. For the age profiling of more recent periods with higher time resolution for the three *Cercopithecidae*, we used the sequence similarity of Alu elements based on their non-self best match from an all-against-all search using the NCBI BLASTn (Altschul et al. 1990). A minimum of 100 bp in length and 80 in BLASTn alignment score was applied for analyzing the Alu matches.

4.3.3 Identification of Alu master copies in the primate genomes

To identify the recent Alu master copies in the primate genomes, we first performed a pre-processing to integrate the Alu fragments annotated by RepeatMasker back to Alu sequences representing the original transposition events as previously described (Tang et al. 2018). The sequences of all Alu elements in the six primate genomes were then retrieved from the reference primate genomes, followed by an all-against-all BLAST search (Altschul et al. 1990) for Alus in a genome. Using a set of in-house Linux shell and Perl scripts, the best non-self-match based on BLAST score was identified for each Alu element based on the BLASTn output. To exclude random sequence match, a minimum sequence similarity at 95% over a minimum length of 100 bp were applied for a qualified match. We reason that an Alu has the best match to its parent copy or to its own daughter copies, and we define a master copy as an Alu being the 2nd best non-self-match for 10 or more other Alu copies. We further require an Alu master copy to have a full sequence by being 280 bp or more in length.

4.3.4 Profiling Alu element expression using RNA-seq

4.3.4.1 Sources of RNA-Seq data

For the five non-human primates, RNA-Seq data generated using the generic (mixed) samples were downloaded from the Non-Human Primate Reference Transcriptome Resource (NHPRTR) (Pipes et al. 2013). These five primates include chimpanzee, gorilla, rhesus, crab-eating macaque and baboon.

Table 4.1 RNA-seq data used for generating simulated human mixed (generic) samples

Tissue	**SRA experiment ID**	**Number of clusters**
Adipose	ERR030880	77300072
Blood	SRR1060757	33353039
Brain	ERR030882	73513047
Brain	SRR2040575	73730964
Brain	SRR2040576	62770724
Colon	SRR2012208	68628998
Colon	SRR2012209	71105369
Heart	ERR030886	82918784
Heart	SRR2040577	70846205
Heart	SRR2040578	82628997
Kidney	ERR030885	80397337
Kidney	SRR1536710	33422761
Kidney	SRR1536711	32198935
Liver	ERR030887	80048623
Liver	SRR2040579	67904633
Liver	SRR2040580	77680887
Lung	ERR030879	79296905
Lymph Node	ERR030878	82078157
Skeletal muscle	ERR030876	82111139
Testis	SRR2040581	68228322
Testis	SRR2040582	85271817
Testis	SRR2040583	16202310
Thymus	SRR1299440	34930991
Thymus	SRR1299441	31238820
Total		**1547807836**

Additionally, as shown in Table 4.1, we collected 24 RNA-seq samples across 12 different human tissues from the NCBI Short Read Archive database (Leinonen et al. 2011; Shin et al. 2014). For a better comparison between the human and non-human primate RNA-seq samples, we generated a simulation human generic (mixed) sample by merging the RNA-seq data from different tissue samples in a composition similar to the NHPRTR dataset.

4.3.4.2 Processing of RNA-Seq data

The RNA-seq data were first mapped to the primary genome assemblies (sequences assigned to specific chromosomes) of the aforementioned primate reference genomes using TopHat2 (Kim et al. 2013). To identify RNA-seq reads representing individual Alus, only a single primary alignment from the RNA-seq alignment file was used for collecting reads mapped to Alus using Sambamba (Tarasov et al. 2015).

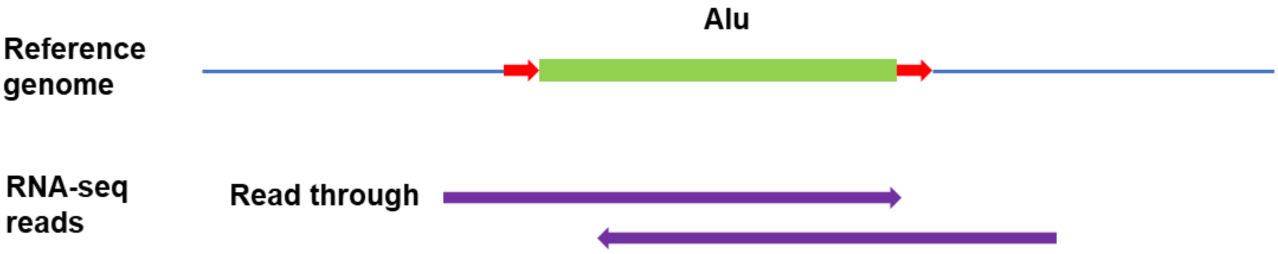


Figure 4.1 A schematic diagram for filtering out passively expressed Alu elements

The green box indicates an Alu element flanked by a pair of TSDs (red arrows) in the reference genome. The purple arrows represent reads from passive Alu transcripts not driven by Alu internal promoters but by host genes in the region.

To ensure that only authentic Alu expression driven by the internal promoter was included, reads representing passive expression as shown in Fig. 4.1 were discarded. The remaining reads mapped to Alus were used to calculate the normalized Alu expression in fragments per kilobase of transcript per million reads (fpkm) using in-house PERL scripts.

4.3.5 Data analysis

Most sequence analyses were performed on the high performance computing facilities at Compute Canada (www.computecanada.ca) running CentOS Linux. Data analysis and figure generation were performed using a combination of Linux shell scripting, R and Microsoft Excel.

4.4 Results

4.4.1 Alu age and composition profiles in the primate genomes

For this study, we chose to use six primate genomes including human, chimpanzee, gorilla, rhesus monkey, crab-eating macaque, and baboon, based on the availability of transcriptome data. Utilizing an integration process described previously (Tang et al. 2018), we pre-processed the Alu elements annotated by Repeatmasker to integrate the fragmented Alu elements to represent original insertion events.

Table 4.2 The copy numbers and total sequence length of Alu elements in the six primate genomes

Genome	AluJ			AluS			AluY		
	copy number	total size (bp)	genome %	copy number	total size (bp)	genome %	copy number	total size (bp)	genome %
Human	286,597	73,139,136	2.5	656,214	188,566,363	6.4	136,483	39,032,011	1.3
Chimpanzee	264,591	68,452,900	2.4	663,055	190,339,136	6.6	129,811	37,276,336	1.3
Gorilla	256,223	65,872,448	2.4	632,028	178,626,066	6.4	121,315	33,634,842	1.2
Rhesus	246,746	64,059,981	2.3	624,699	177,385,009	6.4	231,221	65,980,015	2.4
Crab-eating macaque	263,384	67,258,541	2.5	604,788	171,895,354	6.3	208,673	58,879,705	2.2
Baboon	260,221	66,288,450	2.5	589,181	166,999,630	6.2	229,490	62,906,049	2.3

As shown in Table 4.2, there are ~6.4 million Alu elements in the six primate genomes, contributing to a total of ~1.8 billion base pair in genomic sequences. The AluS subfamily, averaging ~628,000 copies per primate genome and being ~2.4 and ~3.6 times larger than the other two major subfamilies, AluJ and AluY, is the most successful Alus in these primate genomes. The numbers of AluJ elements and AluS elements have shown no significant differences between the *Hominidae* and *Cercopithecidae* primate families (P-value > 0.1, data not shown), indicating they were generated in the common ancestor of these lineages during the early phases of primate evolution. In contrast, the numbers of AluY elements in the *Cercopithecidae* genomes are about ~1.7 times larger than these for the *Hominidae* genomes and the difference is statistically significant (P-value = 0.0003, data not shown), suggesting that the emergence of AluY occurred around the divergence of these two primate lineages and continued into the evolution of each lineage.

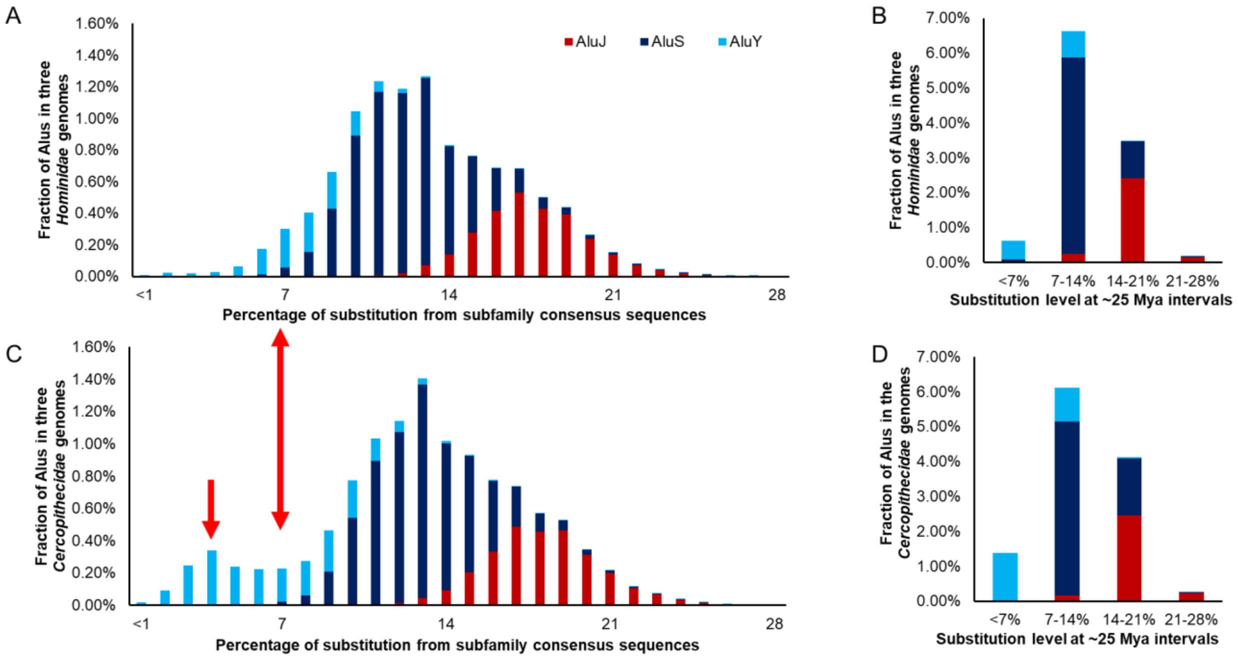


Figure 4.2 The age and composition profiles of Alu elements in the six primate genomes

The average number of substitutions per 100 bp (K) was calculated using the mismatch level p, which was provided in the RepeatMasker annotations, according to the one-parameter Jukes-Cantor model (K = -3/4ln(1 - 4/3p)). A, bar plots showing the percentage of Alu elements in the genome at different ages (K bins). The data is based on all Alus in the three *Hominidae* genomes including human, chimpanzee and gorilla; B, the data in A plotted in large substitution bins at 25 million years (My) intervals; C, bar plots showing the percentage of Alu elements at different ages (K bins). The data is based on all Alu elements in the three *Cercopithecidae* genomes including rhesus, crab-eating macaque, and baboon. D, the data in C plotted in large substitution bins at 25 My intervals. The color legend for B, C, and D is the same as in A. The double arrow between A and C indicates the starting point of differential Alu transposition between the two primate families, while the single arrow in C indicates the additional AluY transposition peak in the *Cercopithecidae* genomes.

The differential ages among the three Alu major subfamilies and the AluY profile differences between the two primate families were more clearly seen in the Alu age profiles based on the average substitution level of Alu sequences from their perspective consensus sequences. As shown in Fig. 4.2A & C, the distribution of Alus by their age and genome composition is very similar between the *Hominidae* and *Cercopithecidae* families for the AluJ and AluS subfamilies. In contrast, the profile of AluY elements being the youngest Alu subfamily showed clear differences between the two primate families. As shown in Fig. 4.2A & C, the AluY subfamily kept a relatively high activity in the genomes of the *Hominidae* and the *Cercopithecidae* families till their divergence ~25-30 Myr ago (from 7% substitution rightwards in Fig. 4.2A & C) (the evolution time was based on data in TimeTree database; http://timetree.org)(Hedges et al. 2006). However, while the activity of the AluY subfamily has quickly slowed down in the ape genomes, it retained at a high level in the monkey genomes for a much longer period with even a recent activity peak visible (the left sections in Fig. 4.2A &C), leading to ~2.5 times more AluY elements in their genomes compared to the ape genomes (Fig. 4.2B & D). The profile similarity for AluJ and AluS subfamilies and the differences for AluY subfamily between the two primate groups are also clearly seen when Alu elements were grouped into three larger evolutionary periods at ~25 My intervals (Fig. 4.2B & D).

We also performed more detailed comparison between the three monkey genomes, since they were shown in our previous study to have highly variable species-specific SINE (SS-SINE) transposition. Specifically, the baboon genome has an extremely high number of SS-SINEs, while the crab-eating macaque genome has an extremely low number of SS-SINEs, and the rhesus genome lies in the middle (54,969, 2,257, and 22,069, respectively, after normalization for the relative evolutionary distance)(Tang and Liang 2019).

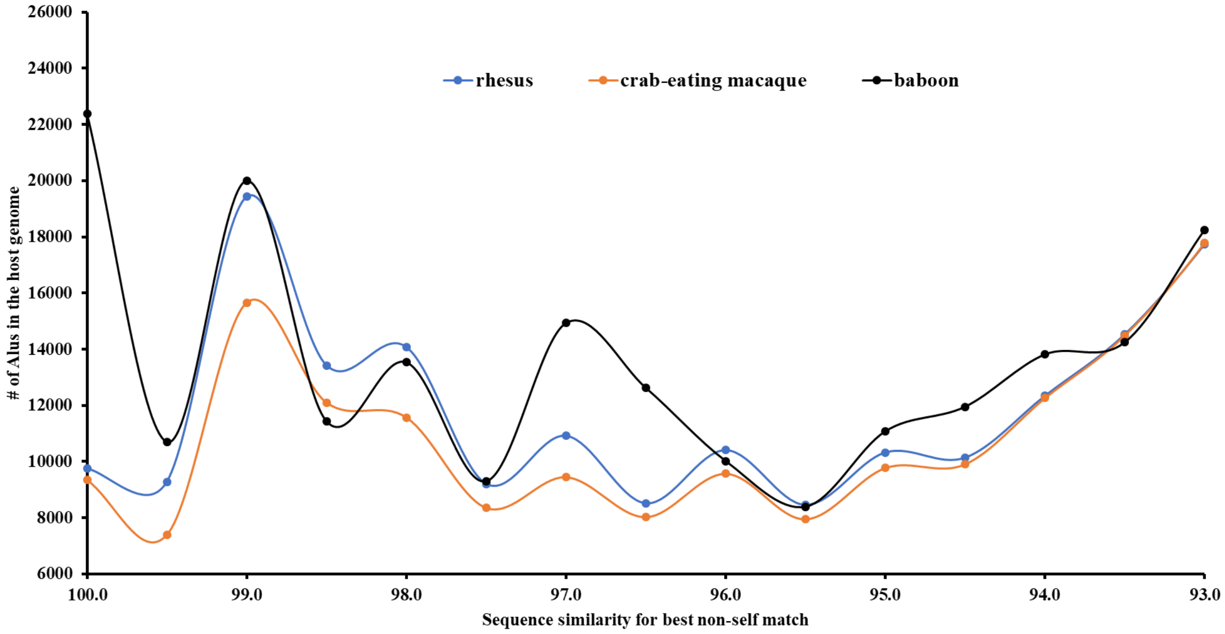


Figure 4.3 The age and composition profiles of Alu elements in the three *Cercopithecidae* genomes

Line plots (trendlines) for the distribution of Alu elements at different sequence similarity levels based on their non-self best matches in the perspective genome. The data is based on all Alu elements in the three *Cercopithecidae* genomes including rhesus, crab-eating macaque, and baboon.

As shown in Fig. 4.3, the recent Alu profile is quite different among the three genomes for the more recent period represented by sequence similarity up from 93.3%. As expected, the rhesus and crab-eating macaque genomes shared a similar Alu profile over a longer period (up to 94.5% sequence similarity) due to their closer evolutionary relationship. Interestingly, the baboon genome seemed to have experienced a few major hikes (at least 3) of Alu transposition after its divergence from the ancestor of the other two monkey species, with the highest occurred towards the most recent or current period. These activity peaks are all much higher than all activity peaks in the rhesus and crab-eating macaque genomes (Fig. 4.3), and this is consistent with the highest SINE transposition reported in the baboon genome (Jordan et al. 2018; Steely et al. 2018; Tang and Liang 2019). Between the rhesus and the crab-eating macaque genomes, the activity profiles seem to be similar, both showing similar peaks that are much lower relative to those seen in the baboon genome. However, the level in the crab-eating macaque genome was consistently lower than that in the rhesus genome over the entire period since their divergence, accumulatively leading to the large difference of SS-Alus between the two genomes as previously reported (Tang and Liang 2019).

4.4.2 The number of Alu master copies is positively correlated to the total number of Alu elements in the primate genomes

To evaluate the contribution of Alu master copies towards differential Alu activity in the most recent primate genomes, we identified the potential Alu master copies by performing an all-against-all BLAST search (Altschul et al. 1990) for all Alu sequences in each of the six primate genomes. We used the best non-self-match (>=95% similarity; >=100 bp) to identify the parent-daughter copy relationship for each Alu element in a genome, and we defined an Alu master copy using two criteria: 1) the master copy is the 2nd best non-self-match for ten or more Alu copies; 2) the master copy has a full sequence by being 280 bp or more in length.

As shown in Table 4.3, we identified a total of 5,401 Alu master copies in the six primate genomes, which represent only a very small proportion (~0.08%) of the total Alu population. As expected, based on the Alu age profile, the majority (~99.4%) of these Alu master copies belong to the AluY subfamily, with only 32 and 1 Alu master copies from the AluS and AluJ subfamilies, respectively (Table 4.3). The number of Alu master copies from the AluY subfamily differs substantially among the six primate genomes, ranging from only 235 copies in the gorilla genome to 1,710 copies in the rhesus genome (Table 4.3). Notably, the average number of AluY master copies in the three *Cercopithecidae* genomes (~1,468 copies per genome) is ~4.5 times higher than the number for the three *Hominidae* genomes (~321 copies per genome) (Table 4.3).

Table 4.3 Number of Alu master copies by major subfamilies in the six primate genomes

Genome	**AluJ**			**AluS**			**AluY**		
	Total number	**Master copy #**	**Master copy (%)**	**copy number**	**Master copy #**	**Maser copy (%)**	**copy number**	**Master copy #**	**Master copy (%)**
Human	286,597	0	0	656,214	3	0	136,483	410	0.3
Chimpanzee	264,591	0	0	663,055	5	0	129,811	319	0.2
Gorilla	256,223	1	0	632,028	4	0	121,315	235	0.2
Rhesus	246,746	0	0	624,699	6	0	231,221	1,710	0.7
Crab-eating macaque	263,384	0	0	604,788	7	0	208,673	1,330	0.6
Baboon	260,221	0	0	589,181	7	0	229,490	1,364	0.6
Total	**1,577,762**	**1**	**0**	**3,769,965**	**32**	**0**	**1,056,993**	**5,368**	**0.5**

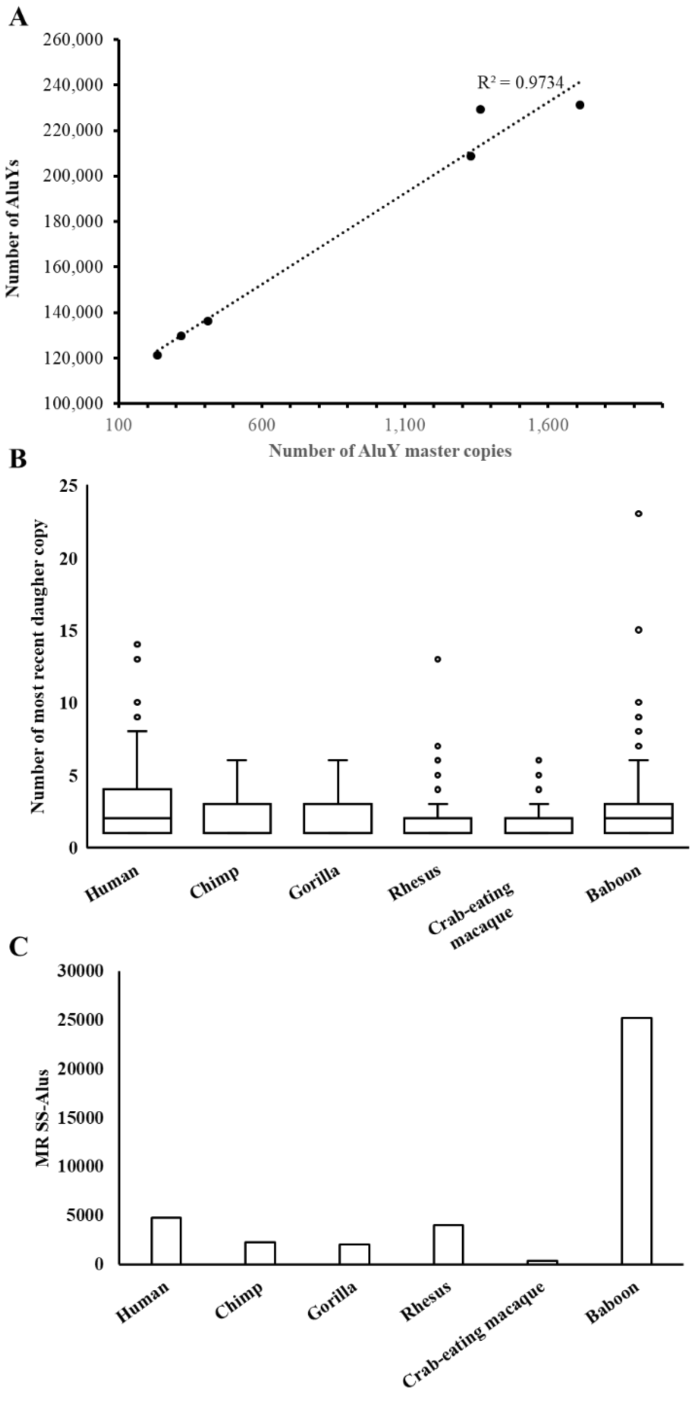


Figure 4.4 Alu master copies in the six primate genomes

A, an XY scatter plot between the number of AluY master copies and the total number of AluY elements in the genomes. B, boxplots showing the distribution pattern of recent daughter copy numbers for the master copies in the six genomes. C, bar plots showing the numbers of most recent species-specific Alus in the six primate genomes based on data from Tang and Liang, 2019 with some modifications.

More importantly, the number of AluY master copies in the six primate genomes showed a very strong positive correlation to the total number of AluYs (R^2=0.9734), and such correlation is statistically significant (P-value < 0.000001) (Fig. 4.4A). This indicates that the number of Alu master copies directly contributes to the differential AluY transposition in these genomes.

We also examined a subgroup of these Alu master copies as the most recent Alu master copies in the six primate genomes, which generated at least one most recent daughter copy. A most recent daughter copy is defined as daughter copy sharing 100% sequence similarity with its parent copy, which is the same sequence similarity cut-off we used to identify most recent species-specific Alu elements in a previous study (Tang and Liang 2019).

Table 4.4 Numbers of most recent Alu master copies and most recent SS-Alus in the six primate genomes

Genome	**Most recent Alu master copies**				**MR SS-Alu**
	All	**Top 50%**	**Top 25%**	**Top 10%**	
Human	153	99	62	16	4772
Chimpanzee	65	30	15	2	2296
Gorilla	63	25	16	3	2079
Rhesus	432	337	139	29	4076
Crab-eating macaque	718	187	84	9	412
Baboon	589	332	189	42	25247
Total	**2020**	**1010**	**505**	**101**	**38882**

As shown in Table 4.4, 2,020 (~37.4%) of the total 5,401 Alu master copies in the primate genomes have generated at least one most recent daughter copy. Similar to the total numbers of Alu master copies, the numbers of most recent Alu master copies are much higher in the three monkey genomes (~580 copies per genome) than in the three ape genomes (~93 copies per genome) (Table 4.4). Further, we analyzed the distribution of the most recent daughter copies based on their numbers of daughter copies in each of the genomes (Fig. 4.4B). While crab-eating macaque had the highest number (718) of Alu master copies with most recent daughter copies in its genome compared to the other five primates (Table 4.4), the majority of these Alu master copies generated less than five most recent daughter copies as indicated by its lowest distribution (Fig. 4.4B). This was consistent with our previous finding that crab-eating macaque had the least number of most recent SS-Alus in its genome (Fig. 4.4C). In contrast, the baboon genome, which had the largest number of most recent SS-Alus (Fig. 4.4C), has the highest middle quartile compared to the other primate genome (Fig. 4.4B). Additionally, the baboon genome has the highest number of outliers at the top, with the largest number of most recent daughter copies being 23 (Fig. 4.4B), consistent with the highest number of SS-Alus and most recent SS-Alus in this genome (Fig. 4.4C).

We further investigated the subgroups of most recent Alu master copies in the six primate genomes by sorting the most recent Alu master copies based on the numbers of most recent daughter copies from high to low. We then grouped these most recent Alu master copies as top 50%, top 25% and top 10% copies and examined their relationship with the number of most recent SS-Alus. As shown in Table 4.4, the number of most recent Alu master copies are higher in the three genomes with higher numbers of most recent SS-Alus (human, rhesus and baboon) than in the three genomes with lower number of most recent SS-Alus (chimpanzee, gorilla and crab-eating macaque), especially for those in the top 10% group. The baboon genome, which has 25,247 most recent SS-Alu entries, has the largest number of most recent Alu master copies within both the top 25% and top 10% (Table 4.3). The crab-eating macaque, despite having the largest number of most recent Alu master copies (718) in its genome, only had nine entries (~1.3%) ranked among the top 10% group, which is approximately five-fold lower than the average of the other five genomes (Table 4.4). The poor performance of these master copies in the crab-eating macaque genome was consistent with our observation that it also had the lowest number of most recent SS-Alus (Fig. 4.4C), suggesting the existence of a potential mechanism suppressing ME transposition.

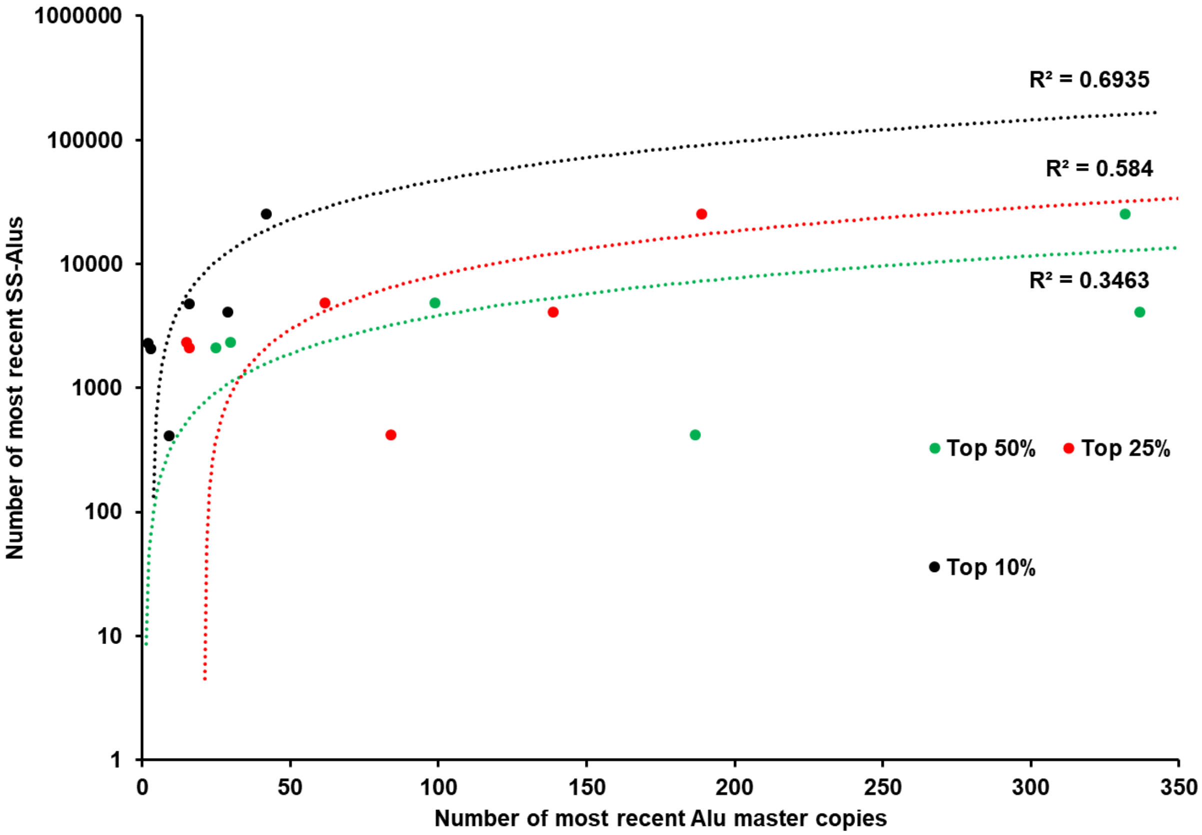


Figure 4.5 Relationship between the number of most recent master copies and the number of most recent species-specific Alus (SS-Alus) in the six primate genomes

Most recent Alu master copies are defined as Alu master copies which have generated at least one most recent daughter copy, and they are divided into percentiles based on the distribution of their numbers of most recent daughter copies with the top representing the larger numbers. Three plots were made based on the top 50% (green), top 25% (red), and top 10% (black) of most recent Alu master copies. A trendline with the calculated Pearson coefficient (R^2) was generated for each data group.

As shown in Fig. 4.5, the numbers of most recent Alu master copies in each primate genome were positively correlated with the numbers of most recent SS-Alu, and such positive correlations were observed for all three groups, but the top 10% group showed the strongest (R^2

=0.6935) and the top 25% group being the next (R^2=0.5840), and the top 50% group showed the weakest correlation (R^2=0.3463; Fig. 4.5). This suggests that the overall success of Alus in recent primate genomes seems to have depended on a few more successful master copies.

4.4.3 The expression pattern of Alu master copies in the primate genomes

To test whether the expression level of Alu master copies in these primate genomes also played a role in the differential Alu transposition, we analyzed the expression of the Alu master copies utilizing RNA-seq data from the Non-Human Primate Reference Transcriptome Resource (NHPRTR) (Pipes et al. 2013), as well as human transcriptome data. In total, five non-human primate RNA-seq data, which were generated using mixed (generic) samples, were collected. Besides, 24 human RNA-seq samples across 12 different tissues were downloaded from the NCBI Short Read Archive database (Leinonen et al. 2011; Shin et al. 2014) and merged using samtools to create a simulated mixed sample (Table 4.1).

Table 4.5 Summary statistics for RNA-seq data from the six primate transcriptomes

Species	Total number of Reads	Total number of mapped reads	Mappable reads (%)
Human	3,095,615,672	2,834,591,035	91.6
Chimpanzee	1,673,728,164	1,366,258,623	81.6
Gorilla	1,772,522,826	1,407,304,071	79.4
Rhesus	1,408,986,794	1,201,943,687	85.3
Crab-eating macaque	1,788,735,188	1,386,239,226	77.5
Baboon	1,837,471,794	1,561,157,333	85.0

The six primate RNA-seq data were then mapped to their corresponding reference genome assembly using TopHat2 (Kim et al. 2013)(Table 4.5). The simulated human transcriptome data had the highest mappable ratio (91.6%) probably due to its best genome assembly quality among the six primate reference genomes.

Table 4.6 The average expression level in fpkm of master copies and full-length copies of major Alu subfamilies in the six primate genomes

Species	Subfamily	Copy number		Average fpkm	
		full-length Alus	Alu master copies	full-length Alus	Alu master copies
Human	AluJ	90,559	0	0.00035	0
	AluS	359,038	3	0.00090	0
	AluY	90,781	410	0.00225	0.00285
Chimpanzee	AluJ	85,929	0	0.00013	0
	AluS	358,638	5	0.00015	0
	AluY	84,143	319	0.00032	0.00180
Gorilla	AluJ	81,826	1	0.00018	0
	AluS	328,832	4	0.00020	0
	AluY	71,927	235	0.00050	0.00223
Rhesus	AluJ	77,635	0	0.00020	0
	AluS	316,928	6	0.00007	0
	AluY	158,802	1710	0.00007	0.00012
Crab-eating macaque	AluJ	80,247	0	0.00023	0
	AluS	307,915	7	0.00018	0.00102
	AluY	131,585	1330	0.00099	0.00166
Baboon	AluJ	78,747	0	0.00013	0
	AluS	303,047	7	0.00005	0
	AluY	149,884	1364	0.00007	0.00010

In collecting the RNA-seq reads to represent the expression level of Alu master copies, we exercised extreme caution to reduce the erroneous assignment of reads between different Alus due to their extremely high sequence similarity. These cautionary protocols include requiring very high sequence similarity and using a read only once as well as calculating the fpkm values for all Alu master copies at the subfamily level instead of at the individual copy level. Specifically, the Alu master copies were grouped based on their subfamilies, and the average fpkm values for the subfamilies were calculated (Table 4.6). Additionally, the average fpkm value was calculated for each subfamily based on full-length Alus (>= 95% of the consensus sequence) and used as a baseline value for comparison.

As shown in Table 4.6, for the full length Alus, the average fpkm values for the AluY subfamily were generally higher than that for the AluS and AluJ subfamilies, averaging ~2.7 and ~3.5 times higher, respectively. This observation was consistent with the fact that the AluY subfamily is the youngest and most active among the three major subfamilies (Batzer et al. 1996). Exceptions to this general trend are the rhesus and baboon genomes, in which the average of fpkm value of AluY was similar to that of AluS but lower than that of AluJ (Table 4.6). This could be contributed by the fact that both the rhesus and baboon genomes have a higher number of full length AluY elements (158,802 and 149,884) than the other four genomes (Table 4.6), which might have triggered more suppression for expression from the host genome's defensive mechanism in general at the subfamily level. Apparently, the average expression level may not be an accurate reflection of the expression level for the individual master copies. In fact, the expression level of Alu master copies was always higher than that of full length Alus from any subfamily (Table 4.6), suggesting that the high activity level of Alu master copies was supported by their higher expression levels.

We further examined the relationship of different most recent SS-Alu subfamilies' success and the expression levels of their master copies.

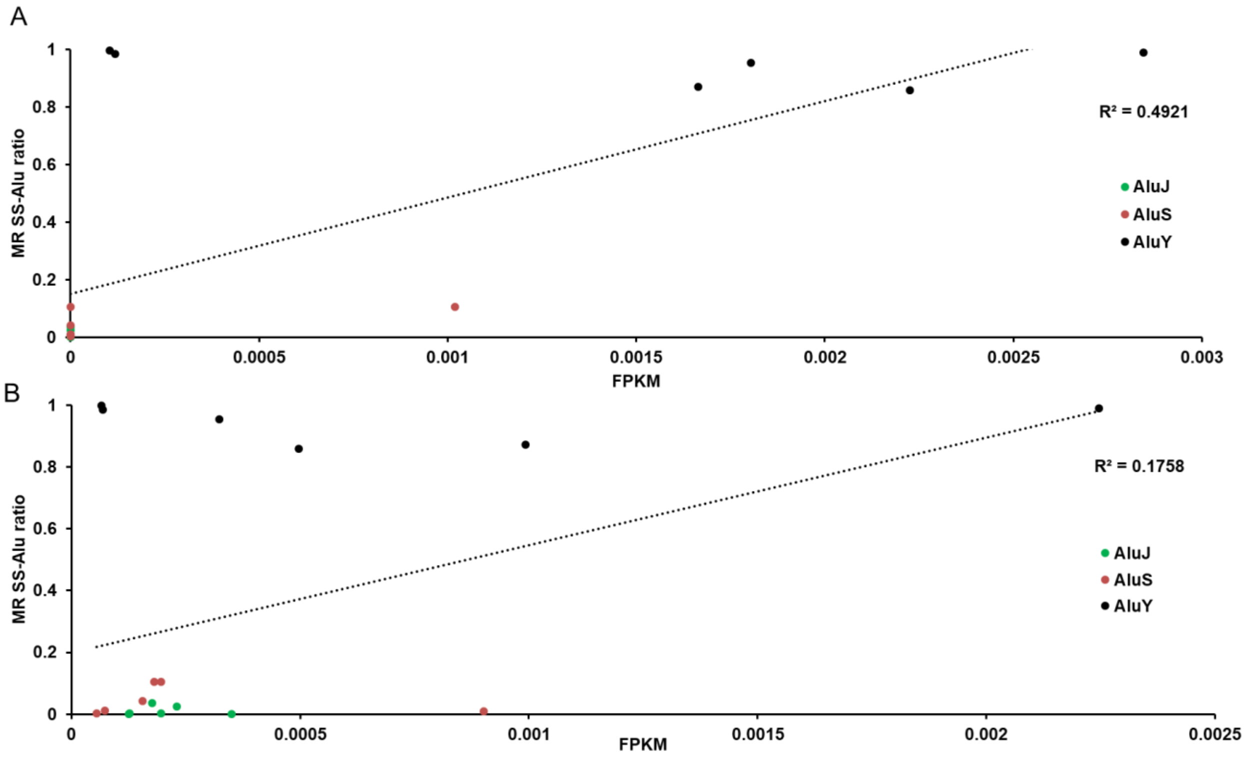


Figure 4.6 Relationship between the numbers of most recent species-specific Alus (MR SS-Alu) and the average expression levels in fpkm of both master copies or full-length Alu copies for each major Alu families

A. An XY scatter plot showing the relationship between the numbers of MR SS-Alus and the average fpkm values of master copies. B. A scatter plot showing the relationship between the numbers of most recent species-specific Alus (MR SS-Alu) and the average fpkm values of full-length Alu copies. Red markers represent the AluS family. A trend line with the calculated Pearson coefficient (R^2) was generated for each XY scatter plot.

As shown in Fig. 4.6, there is a positive correlation (R^2=0.4921) between the average fpkm values of Alu master copies and the ratio of most recent SS-Alus in the subfamily, and the strength of this correlation is much stronger than the background signal (i.e. the full-length Alus) (R^2=0.1758) (Fig. 4.6). In addition, we performed a Spearman’s rank correlation using the most recent SS-Alus ratios and the average fpkm values of master copies at more defined subfamily levels for all six primates. The data suggested that the two values showed a moderate positive correlation (Spearman's correlation coefficient $R^2 = 0.4327$) and such correlation is significant ($p=0.0006 <0.001$) (detail data not shown). The positive correlation between the expression level of Alu master copies and their success in Alu transposition is expected as Alu transcripts are required for Alu retrotransposition.

4.5 Discussions

4.5.1 Lineage and species-specific differential Alu transposition in recent primate genomes

Alu elements, as a type of SINEs uniquely found in primate genomes, experienced an explosion in the early phases of primate evolution by generating over one million copies and constituting more than 10% of the genome from the AluS and AluJ subfamilies (Britten et al. 1988; Deininger and Batzer 1999; Deininger 2011; Kazazian 2004). It made Alus the exclusive type of SINEs and a major type of MEs in these genomes, implying their important roles in primate evolution.

Recent studies focusing on Alu elements have suggested that the Alu proliferation profile differs greatly among primate genomes. For example, Steely et al. recently illustrated a burst of most recent Alu activity in the baboon genome by identifying 28,114 baboon-specific Alu elements (Steely et al. 2018). This finding has been confirmed by our observation in a separate recent study focusing on SS-MEs in eight primate genomes covering the *Hominidae* and the *Cercopithecidae* families, which demonstrates that the baboon genome has the largest number of SS-Alus compared to seven other primates (Tang and Liang 2019). The same study also shows that in contrast to the baboon genome, the crab-eating macaque genome has a sustained low transposition for all types of MEs including Alus. Furthermore, the study reveals differences in recent Alu transposition between closely related species. For example, the human genome has a substantially higher very recent acceleration of SINE transposition than the chimpanzee genome, due to the emergence of a few very young and active AluY subfamilies, including AluYa5, AluYb8 and AluYb9. This led to the number of most recent SS-SINEs in the human genome being much larger than that in the chimpanzee genome, despite the latter having a larger number of total SS-SINEs (Tang and Liang 2019).

Similarly, in the current study, we compared the Alu transposition profile among the three closely related monkey genomes (rhesus, crab-eating macaque, and baboon) but in a spectrum beyond the period since their speciation. This allowed us to look back more into their evolution paths and gain a more complete picture of the differential Alu transposition in this lineage. Our analysis revealed that the extreme success of Alu transposition in the baboon genome was contributed by a sustained higher level of transposition for the entire period since its divergence from the common ancestor of the two macaque species via multiple waves of accelerations with a clear uptrend towards to the most recent period (Fig. 4.4). In contrary, the crab-eating macaque genome showed a sustained lower level of Alu transposition since its divergence from rhesus monkey (Fig. 4.4), leading to its low level of Alu transposition, apparently as a species-specific phenomenon, since the extremely low SINE transposition is not seen in its closely related rhesus genome (Tang and Liang 2019). In comparison with our previous study, the current study narrows down the differential SINE transposition to the AluY differential transposition.

At a broader evolutionary spectrum, lineage differences in Alu transposition have also been documented. For example, our recent study showed that SINE transposition ~4 times higher on average in the genomes of the *Cercopithecidae* family than that for the *Hominidae* family (Tang and Liang 2019). In the current study, we compared the Alu age profile of the genomes from the *Hominidae* and the *Cercopithecidae* families in more detail covering the entire Alu evolutionary spectrum. Our results showed that, while there are no significant differences for the older AluS and AluJ subfamilies between the genomes of the two primate families, the profile of the younger AluY subfamily varies substantially between the two families, leading to the total number of AluY elements in the *Cercopithecidae* genomes being almost double of that in the genomes of the *Hominidae* genomes (Table 4.2).

When examined into more detail, the two primate families started to deviate from each other in Alu transposition from the middle point of the AluY evolutionary history, or specifically from the 7% Alu sequence divergence onwards (Fig. 4.2). This converts roughly to ~25 My of evolutionary time, which points roughly to the divergence point for the two primate families based on data from the TimeTree database (http://timetree.org). Interestingly, while Alu transposition in the *Hominidae* genomes showed a steady slowing down from this point onward (Fig. 4.2A), it had a steady increase and peaked in a later stage in the *Cercopithecidae* genomes (Fig. 4.2C). This clearly explains the reason behind the higher AluY transposition in the monkey genomes than the ape genomes.

4.5.2 Alu master copies: the driver of differential Alu transposition in the primate genomes

It was initially proposed that all young Alu elements are capable of generating daughter copies as they all contain internal RNA polymerase III promoters (Jagadeeswaran et al. 1981). This turned out to be false, as we would expect to see the Alu amplification showing an exponential style increase since their numbers kept growing, which is not what we observed in the primate genomes. Later, with more Alu sequences become available, we came to recognize that Alu elements can be categorized into multiple levels of hierarchical or lineal subfamilies based on the specific sequence mutations with members from the younger subfamilies showing higher sequence similarities within the same subfamilies (Batzer et al. 1990; Batzer and

Deininger 1991; Britten et al. 1988). Based on this, researchers hypothesized that there are only a small number of Alu elements or even a single copy as the master copy (copies) responsible for generating new daughter copies (Batzer and Deininger 1991; Britten et al. 1988; Shen et al. 1991). A later study suggested that it is highly unlikely, if not at all impossible, for any Alu subfamilies to have accumulated from one single master copy; rather a small group of Alu master copies might be responsible for Alu amplification in the human genome (Johnson and Brookfield 2006).

So far, a large-scale analysis focusing on the Alu master copies in the primate genomes beyond the human genome is still lacking. In this study, we identified a total of 5,401 Alu master copies by surveying the entire 6.4 million Alu elements in the six primate genomes from the *Hominidae* and the *Cercopithecidae* families (Table 4.3). Our results are consistent with the current theory that only a small group of Alu master genes exist in a primate genome, as our identified Alu master copies only represent ~0.08% of the entire Alu population in these genomes. As expected, the majority of these Alu master copies belong to the younger AluY subfamily, accounting for ~99.4% of all identified Alu master copies (Table 4.3). The AluJ and AluS subfamilies only have 1 and 32 master copies in the current primate genomes, which is consistent with their reduced activity levels in recent primate evolution (Table 4.3). We reason that a large number of master copies should have existed for AluJ and AluS subfamilies, which were responsible for producing the extremely large numbers of AluJ and AluS elements in the genomes. However, most of them are no longer detected as master copies based on our criteria, among which is the requirement for a minimum of 95% sequence similarity, excluding older master copies due to accumulation of mutations.

While the total numbers of Alu master copies do not vary much among genomes from either of the two primate families, the numbers of AluY master copies in the *Cercopithecidae* family are ~4 times more than that in the *Hominidae* family. The number of AluY master copies shows a strong positive linear correlation with the total number of AluY elements in the genomes (Table 4.3, Fig. 4.4A). Further, the activity level of the master copies as measured by the number of daughter copies also seems to have a positive correlation with the overall SS-Alu transposition as measured by the number of most recent SS-Alus (Fig. 4.4B & C, Fig. 4.5, Table 4.4).

Overall our results indicate that not only the total number of Alu master copies, but also their activity level determine the overall Alu transition activity in the genomes. This extends to explain that the extremely high level of most recent Alu transposition in the baboon genome is due to the existence of several highly active recent Alu master copies, while the extremely low level of recent Alu transposition in the crab-eating macaque genome is due to lack of any highly active Alu master copies.

4.5.3 The relationship between the expression level of Alu master copies and their efficiency in Alu transposition

Since transcription is a required step in retrotransposition, it is interesting to test whether the transcriptional level of Alu master copies is directly related to their retrotransposition activity. However, there is a technical challenge for doing this at the individual Alu level due to the very high sequence similarity between master copies from the same subfamily, which may be 100% identical. This issue is more challenging compared to similar analysis for L1s, which are 20 times longer in sequence with a much better chance to carry sequence divergence (Bennett et al. 2008; Iskow et al. 2010; Mills et al. 2007). For this reason, we surveyed the expression level of master copies at the major Alu subfamilies level (i.e. AluS, AluJ, and AluY) as a whole rather than at the individual Alu level. Another issue associated with the analysis of Alu expression is the need to distinguish between the active expression driven by the Alu internal promoter vs. passive transcription driven by nearby or host genes for Alus locating in the transcribed regions (both exons and introns) of genes. While the chance for passively transcribed Alu RNAs to be used in retrotransposition cannot be completely ruled out, these Alus are unlikely to be able to function as master copies. Therefore, only the expression level represented by active Alu transcription would be meaningful for our purpose. We used a relatively simple and straightforward strategy to deal with this issue, and it is to ignore all Alus and the transcripts mapped to them if any passive reads in the pattern shown in Fig. 4.1 were identified for them. Further, for each Alu read is only used once at its primary mapping position for calculating the fpkm values to minimize the impact of random mapping for equally good matches by most aligners.

As expected, a positive correlation is seen between the total expression level of Alu master copies divided into the three major subfamilies and the total number of most recent SS-Alus in the corresponding subfamilies, and this correlation is much stronger than the background level, which is based on the expression level of all full-length Alus in the subfamily (Fig. 4.6). Even though we were unable to generate the relevant data, we can speculate that at the individual Alu master copy level, the correlation between the expression level of Alu master copies and their transposition activity would be much more direct and stronger. Certainly, we are facing another challenge in addressing this relationship, and it is the need to be able to find the germline tissues or germline cells for such analysis. At the current stage, the best tissues available to us for the primates are mixed tissue samples, and this might be one of the reasons that we were only able to see a relatively weak correlation between the expression level of the Alu master copies and the Alu transposition activity.

4.6 **Conclusions and future perspectives**

By using a comparative genomic approach, we examined Alu age and composition profiles and identified Alu master copies in six primate genomes from the *Hominidae* and *Cercopithecidae* families to better understand the differential Alu transposition in primate genomes. Our results indicate that the differential Alu transposition in primate genomes was mainly contributed by the different activities of the younger AluY subfamily. Different primate lineages, as well as between closely related species, are revealed to have Alu activity profiles, which differ not only by the overall AluY transposition level but also by the number of Alu transposition waves and their relative activity levels. Our data for the six primate genomes supports the current model of Alu transposition by a very small number master copies with the number of Alu master copies directly correlated with the total Alu transposition output in the corresponding subfamily across genomes. Furthermore, species with extremely successful recent Alu transposition, such as baboon, tend to have master copies that have high transposition activity. Albert limited, our data suggest a positive correlation between the expression level of Alu master copies and the transposition activity.

Future studies on differential Alu transposition in primates may be directed to the analysis and identification of the sequence features which enabled the success of the Alu master copies to better understand the detailed mechanism of TPRT-based ME transposition and how it may differ between Alus and L1s transposition. In the meantime, an experimental study on the functional impact of individual lineage- and species-specific Alu elements in the primate genomes would be very valuable. Last, but not least, investigating into the host genomes' mechanisms in regulating ME transposition would help explain some unusual phenomenon, such as the global suppression of ME transposition in the crab-eating macaque genome.

Chapter 5　General Discussions

It has been almost two decades since the completion of the human genome project, which lasted for 15 years and cost billions of dollars before producing the first draft of the human genome (Lander et al. 2001). Owing to the numerous new technologies and tools that emerged after, researchers have been able to sequence and assemble a genome for a much more reasonable cost and in a much shorter period. Therefore, more and more primate genomes have since been sequenced (Carbone et al. 2014; Chimpanzee Sequencing and Analysis 2005; Locke et al. 2011; Rhesus Macaque Genome Sequencing and Analysis et al. 2007; Scally et al. 2012; Yan et al. 2011). The amount of primate genome data has greatly expanded our knowledge of MEs. More and more studies have shown that MEs played important roles in both genome evolution and gene function by generating both inter- and intra-species genomic structure variations in the primate genomes. My Ph.D. thesis has focused on the MEs in several primates including human by taking advantage of these recently sequenced primate genomes and transcriptomes.

Despite the availability of more and more primate genomes, past and ongoing studies on MEs have been focusing on either limited numbers of young and active members in the human genome or species-specific mobile elements (SS-MEs) using a few limited primate genomes. In this thesis, by utilizing the recently available primate genomes, a comprehensive analysis of SS-MEs in eight primate genomes from the families of *Hominidae* and *Cercopithecidae* was performed, mainly focusing on the retrotransposons. The study has also identified a small number of DNA transposons, despite considered to be currently inactive in the primate genomes, that appear to be species-specific. The efforts to understand the mechanisms underlying these species-specific DNA transposons have resulted in the identification of a new type of non-LTR retrotransposons derived from DNA transposons is reported for the first time. These DNA transposons share sequence features characteristic of L1-based retrotransposons, and we, therefore, name them as retro-DNA, adding them as the fifth subclass of non-LTR retrotransposons after LINEs, SINEs, SVAs, and processed pseudogenes. The thesis also describes a comprehensive analysis of Alu master copies at both genome and transcriptome levels, as part of efforts to understand the differential Alu amplification in recent primate genomes identified in the first study.

Overall, the thesis has been focusing on mobile elements in the recent primate genomes and their impact on genome evolution. The results presented in the three data chapters are valuable for the field of mobile elements as well as primate evolution. Chapter 2 has revealed remarkable differential levels of ME transposition among primate genomes from the top two primate families, *Hominidae* and *Cercopithecidae*. Notably, the ME transposition seems to be lowered to a ground level for all ME classes in the crab-eating macaque genome, likely due to a genome-wide suppression of ME transposition, while it is highly active in the baboon and human genome, each due to the existence of several unique highly active ME subfamilies. The results from Chapter 2, including the unpublished parts, inspired follow-up studies that resulted in the other two data chapters. Chapter 3 has reported a new type of non-autonomous non-LTR retrotransposons, which derived from DNA transposon sequences. Named as “retro-DNA”, these elements represent the 5th type of non-LTR retrotransposons after LINE, SINE, SVA, and processed pseudogene, very likely using the same L1-based TPRT mechanism. They serve to propagate DNA transposon sequences in the absence of the canonical DNA transposon activity. These retro-DNA elements, albeit being smaller in number when compared to the other types of non-LTR retrotransposons, do contribute to the genetic diversity among primate species. Moreover, the discovery of these retro-DNA elements suggests that the L1 TRPT machinery may have been used by more diverse types of RNA transcripts than what we currently know. Chapter 4 has shown that the differential Alu transposition in primate genomes was mainly contributed by the different activities of the younger AluY subfamily. Different primate lineages as well as between closely related species are revealed have Alu activity profiles, which differ not only by the overall level but also by the number of Alu transposition waves and their relative activity levels. Our data for the six primate genomes supports the current model of Alu transposition by a very small number of master copies with the number of Alu master copies directly correlated with the total Alu transposition output in the corresponding subfamily across genomes.

One of the biggest challenges for studying MEs in the primate genomes is their high content and similarity in the primate genomes. Averaging ~3.3 to ~ 3.6 million copies, MEs have contributed to almost half of the primate genomes. The abundance of MEs in the primate genomes, combined with the high sequence similarities between copies especially for the ones coming from the same subfamilies, have made it a challenge for studying MEs using comparative genomics methods, particularly while using the short next-generation sequence (NGS) reads. Therefore, despite having the advantage of being homoplasy-free as well as associations with cancer and other types of diseases, MEs have yet to receive wide applications in the clinical fields. Furthermore, despite many genomes having become available over recent years, owing to the new technology and tools emerged since the initial human genome project, there are still many issues with these resources. For example, most reference genomes are still incomplete and may contain assembly errors, especially for the non-human primate genomes. The gap regions in these genomes are usually biased towards the repeat sequence regions and therefore can affect comparative genomics studies. Additionally, the read lengths for the NGS data are usually shorter than the MEs, making it extremely difficult to assemble and more errorprone in regions rich of MEs. Although MEs are no longer being considered as “junk DNA”, many mysteries remain to be uncovered in order to fully understand their roles in primate genome evolution and function. With the constantly improving sequencing technologies and bioinformatics tools, more primate genome and transcriptome data with better quality will become available, and this should be able to greatly benefit future research on MEs using comparative genomics methods. Furthermore, the results presented in this thesis, despite being limited to primate genomes, can have implications for understanding the general trend regarding the mechanism and function of MEs in other groups of organisms.

References for general introduction and discussions

Ahmed, M. and Liang, P. (2012), 'Transposable elements are a significant contributor to tandem repeats in the human genome', *Comp Funct Genomics,* 2012, 947089.

Ahmed, M., Li, W., and Liang, P. (2013), 'Identification of three new Alu Yb subfamilies by source tracking of recently integrated Alu Yb elements', *Mobile DNA,* 4 (1), 25.

Allet, B. (1979), 'Mu insertion duplicates a 5 base pair sequence at the host inserted site', *Cell,* 16 (1), 123-29.

Altschul, S. F., et al. (1990), 'Basic local alignment search tool', *J Mol Biol,* 215 (3), 403-10.

Anwar, S. L., Wulaningsih, W., and Lehmann, U. (2017), 'Transposable Elements in Human Cancer: Causes and Consequences of Deregulation', *Int J Mol Sci,* 18 (5).

Battilana, J., et al. (2006), 'Alu insertion polymorphisms in Native Americans and related Asian populations', *Annals of Human Biology,* 33 (2), 142-60.

Batzer, M. A. and Deininger, P. L. (1991), 'A human-specific subfamily of Alu sequences', *Genomics,* 9 (3), 481-87.

Batzer, M. A., et al. (1990), 'Structure and variability of recently inserted Alu family members', *Nucleic Acids Res,* 18 (23), 6793-8.

Batzer, M. A., et al. (1996), 'Standardized nomenclature for Alu repeats', *J Mol Evol,* 42 (1), 3-6.

Baucom, R. S., et al. (2009), 'Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome', *PLoS Genet,* 5 (11), e1000732.

Beck, C. R., et al. (2010), 'LINE-1 retrotransposition activity in human genomes', *Cell,* 141 (7), 1159-70.

Benit, L., Calteau, A., and Heidmann, T. (2003), 'Characterization of the low-copy HERV-Fc family: evidence for recent integrations in primates of elements with coding envelope genes', *Virology,* 312 (1), 159-68.

Bennett, E. A., et al. (2008), 'Active Alu retrotransposons in the human genome', *Genome Res,* 18 (12), 1875-83.

Bourque, G., et al. (2018), 'Ten things you should know about transposable elements', *Genome Biol,* 19 (1), 199.

Britten, R. J., et al. (1988), 'Sources and evolution of human Alu repeated sequences', *Proc Natl Acad Sci U S A,* 85 (13), 4770-4.

Callinan, P. A., et al. (2005), 'Alu retrotransposition-mediated deletion', *Journal of Molecular Biology,* 348 (4), 791-800.

Carbone, L., et al. (2014), 'Gibbon genome and the fast karyotype evolution of small apes', *Nature,* 513 (7517), 195-201.

Chenna, R., et al. (2003), 'Multiple sequence alignment with the Clustal series of programs', *Nucleic Acids Res,* 31 (13), 3497-500.

Chimpanzee Sequencing and Analysis, Consortium (2005), 'Initial sequence of the chimpanzee genome and comparison with the human genome', *Nature,* 437 (7055), 69-87.

Chuong, E. B., Elde, N. C., and Feschotte, C. (2016), 'Regulatory evolution of innate immunity through co-option of endogenous retroviruses', *Science,* 351 (6277), 1083-7.

Cordaux, R. and Batzer, M. A. (2009), 'The impact of retrotransposons on human genome evolution', *Nature reviews.Genetics,* 10 (10), 691-703.

Cost, G. J. and Boeke, J. D. (1998), 'Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure', *Biochemistry,* 37 (51), 18081-93.

Deininger, P. (2011), 'Alu elements: know the SINEs', *Genome Biol,* 12 (12), 236.

Deininger, P. L. and Batzer, M. A. (1999), 'Alu repeats and human disease', *Molecular genetics and metabolism,* 67 (3), 183-93.

Deininger, P. L., et al. (2003), 'Mobile elements and mammalian genome evolution', *Curr Opin Genet Dev,* 13 (6), 651-8.

Dewannieux, M., Esnault, C., and Heidmann, T. (2003), 'LINE-mediated retrotransposition of marked Alu sequences', *Nature genetics,* 35 (1), 41-48.

Ewing, A. D. and Kazazian, H. H., Jr. (2011), 'Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans', *Genome research,* 21 (6), 985-90.

Feschotte, C. and Pritham, E. J. (2007), 'DNA transposons and the evolution of eukaryotic genomes', *Annu Rev Genet,* 41, 331-68.

Feschotte, C., Swamy, L., and Wessler, S. R. (2003), 'Genome-wide analysis of mariner-like transposable elements in rice reveals complex relationships with stowaway miniature inverted repeat transposable elements (MITEs)', *Genetics,* 163 (2), 747-58.

Goodier, J. L. (2016), 'Restricting retrotransposons: a review', *Mob DNA,* 7, 16.

Grindley, N. D. (1978), 'IS1 insertion generates duplication of a nine base pair sequence at its target site', *Cell,* 13 (3), 419-26.

Han, J. S., Szak, S. T., and Boeke, J. D. (2004), 'Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes', *Nature,* 429 (6989), 268-74.

Han, K., et al. (2005), 'Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages', *Nucleic acids research,* 33 (13), 4040-52.

Han, K., et al. (2007), 'Alu recombination-mediated structural deletions in the chimpanzee genome', *PLoS genetics,* 3 (10), 1939-49.

Harris, Robert S. (2007), 'Improved pairwise alignment of genomic dna', (Pennsylvania State University).

Harrow, J., et al. (2012), 'GENCODE: the reference human genome annotation for The ENCODE Project', *Genome Res,* 22 (9), 1760-74.

Hedges, S. B., Dudley, J., and Kumar, S. (2006), 'TimeTree: a public knowledge-base of divergence times among organisms', *Bioinformatics,* 22 (23), 2971-2.

Herron, P. R. (2004), 'Mobile DNA II', *Heredity,* 92 (5), 476-76.

Higashino, A., et al. (2012), 'Whole-genome sequencing and analysis of the Malaysian cynomolgus macaque (Macaca fascicularis) genome', *Genome Biol,* 13 (7), R58.

Hinrichs, A. S., et al. (2006), 'The UCSC Genome Browser Database: update 2006', *Nucleic Acids Res,* 34 (Database issue), D590-8.

Hu, J., Zheng, Y., and Shang, X. (2018), 'MiteFinderII: a novel tool to identify miniature inverted-repeat transposable elements hidden in eukaryotic genomes', *BMC Med Genomics,* 11 (Suppl 5), 101.

Iskow, R. C., et al. (2010), 'Natural mutagenesis of human genomes by endogenous retrotransposons', *Cell,* 141 (7), 1253-61.

Jagadeeswaran, P., Forget, B. G., and Weissman, S. M. (1981), 'Short interspersed repetitive DNA elements in eucaryotes: transposable DNA elements generated by reverse transcription of RNA pol III transcripts?', *Cell,* 26 (2 Pt 2), 141-2.

Jasinska, A. J., et al. (2013), 'Systems biology of the vervet monkey', *ILAR J,* 54 (2), 122-43.

Jasinska, Anna J., et al. (2017), 'Genetic variation and gene expression across multiple tissues and developmental stages in a nonhuman primate', *Nature genetics,* 49 (12), 1714-21.

Jha, A. R., et al. (2009), 'Cross-sectional dating of novel haplotypes of HERV-K 113 and HERV-K 115 indicate these proviruses originated in Africa before Homo sapiens', *Molecular biology and evolution,* 26 (11), 2617-26.

Jiao, Y., et al. (2017), 'Improved maize reference genome with single-molecule technologies', *Nature,* 546 (7659), 524-27.

Johnson, L. J. and Brookfield, J. F. (2006), 'A test of the master gene hypothesis for interspersed repetitive DNA sequences', *Mol Biol Evol,* 23 (2), 235-9.

Jordan, V. E., et al. (2018), 'A computational reconstruction of *Papio* phylogeny using *Alu* insertion polymorphisms', *Mob DNA,* 9, 13.

Jukes, Thomas H. and Cantor, Charles R. (1969), 'CHAPTER 24 - Evolution of Protein Molecules', in H. N. Munro (ed.), *Mammalian Protein Metabolism* (Academic Press), 21-132.

Jurka, J. (1997), 'Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons', *Proceedings of the National Academy of Sciences of the United States of America,* 94 (5), 1872-77.

Jurka, J., et al. (2005), 'Repbase Update, a database of eukaryotic repetitive elements', *Cytogenet Genome Res,* 110 (1-4), 462-7.

Kapitonov, V. V. and Jurka, J. (2001), 'Rolling-circle transposons in eukaryotes', *Proc Natl Acad Sci U S A,* 98 (15), 8714-9.

Kazazian, H. H., Jr. (2004), 'Mobile elements: drivers of genome evolution', *Science,* 303 (5664), 1626-32.

Kazazian, H. H., Jr. and Goodier, J. L. (2002), 'LINE drive. retrotransposition and genome instability', *Cell,* 110 (3), 277-80.

Kent, W. J. (2002), 'BLAT--the BLAST-like alignment tool', *Genome research,* 12 (4), 656-64.

Kidwell, M. G. (2002), 'Transposable elements and the evolution of genome size in eukaryotes', *Genetica,* 115 (1), 49-63.

Kim, D., et al. (2013), 'TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions', *Genome Biol,* 14 (4), R36.

Konkel, M. K. and Batzer, M. A. (2010), 'A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome', *Seminars in cancer biology*.

Lander, Eric S., et al. (2001), 'Initial sequencing and analysis of the human genome', *Nature,* 409 (6822), 860.

Law, J. A. and Jacobsen, S. E. (2010), 'Establishing, maintaining and modifying DNA methylation patterns in plants and animals', *Nat Rev Genet,* 11 (3), 204-20.

Lee, S. I. and Kim, N. S. (2014), 'Transposable elements and genome size variations in plants', *Genomics Inform,* 12 (3), 87-97.

Leinonen, R., et al. (2011), 'The sequence read archive', *Nucleic Acids Res,* 39 (Database issue), D19-21.

Liang, P. and Tang, W. (2012), 'Database documentation of retrotransposon insertion polymorphisms', *Frontiers in bioscience (Elite edition),* 4, 1542-55.

Locke, D. P., et al. (2011), 'Comparative and demographic analysis of orang-utan genomes', *Nature,* 469 (7331), 529-33.

Madeira, F., et al. (2019), 'The EMBL-EBI search and sequence analysis tools APIs in 2019', *Nucleic Acids Res,* 47 (W1), W636-W41.

McClintock, B. (1950), 'The origin and behavior of mutable loci in maize', *Proc Natl Acad Sci U S A,* 36 (6), 344-55.

Mills, R. E., et al. (2007), 'Which transposable elements are active in the human genome?', *Trends in genetics : TIG,* 23 (4), 183-91.

Mills, R. E., et al. (2006), 'Recently mobilized transposons in the human and chimpanzee genomes', *Am J Hum Genet,* 78 (4), 671-9.

Mita, P. and Boeke, J. D. (2016), 'How retrotransposons shape genome regulation', *Curr Opin Genet Dev,* 37, 90-100.

Moran, J. V., et al. (1996), 'High frequency retrotransposition in cultured mammalian cells', *Cell,* 87 (5), 917-27.

Navarro, F. C. and Galante, P. A. (2015), 'A Genome-Wide Landscape of Retrocopies in Primate Genomes', *Genome Biol Evol,* 7 (8), 2265-75.

Oliver, K. R. and Greene, W. K. (2011), 'Mobile DNA and the TE-Thrust hypothesis: supporting evidence from the primates', *Mob DNA,* 2 (1), 8.

Ostertag, E. M. and Kazazian, H. H., Jr. (2001), 'Biology of mammalian L1 retrotransposons', *Annu Rev Genet,* 35, 501-38.

Pace Ii, John K. and Feschotte, Cé (2007), 'The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage', *Genome research,* 17 (4), 4-4.

Page, R. D. (1996), 'TreeView: an application to display phylogenetic trees on personal computers', *Comput Appl Biosci,* 12 (4), 357-8.

Penzkofer, T., et al. (2017), 'L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes', *Nucleic Acids Res,* 45 (D1), D68-D73.

Pipes, L., et al. (2013), 'The non-human primate reference transcriptome resource (NHPRTR) for comparative functional genomics', *Nucleic Acids Res,* 41 (Database issue), D906-14.

Pritham, E. J., Putliwala, T., and Feschotte, C. (2007), 'Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses', *Gene,* 390 (1-2), 3-17.

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007), 'NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins', *Nucleic Acids Res,* 35 (Database issue), D61-5.

Quinn, J. P. and Bubb, V. J. (2014), 'SVA retrotransposons as modulators of gene expression', *Mob Genet Elements,* 4, e32102.

Ramsay, L., et al. (1999), 'Intimate association of microsatellite repeats with retrotransposons and other dispersed repetitive elements in barley', *Plant J,* 17 (4), 415-25.

Ray, D. A., et al. (2005), 'Inference of human geographic origins using Alu insertion polymorphisms', *Forensic science international,* 153 (2-3), 117-24.

Rhesus Macaque Genome Sequencing and Analysis, Consortium, et al. (2007), 'Evolutionary and biomedical insights from the rhesus macaque genome', *Science (New York, N.Y.),* 316 (5822), 222-34.

Roberts, R. J., Carneiro, M. O., and Schatz, M. C. (2017), 'Erratum to: The advantages of SMRT sequencing', *Genome Biol,* 18 (1), 156.

Rogers, J., et al. (2019), 'The comparative genomics and complex population history of Papio baboons', *Sci Adv,* 5 (1), eaau6947.

Scally, A. and Durbin, R. (2012), 'Revising the human mutation rate: implications for understanding human evolution', *Nat Rev Genet,* 13 (10), 745-53.

Scally, A., et al. (2012), 'Insights into hominid evolution from the gorilla genome sequence', *Nature,* 483 (7388), 169-75.

Schneider, G. F. and Dekker, C. (2012), 'DNA sequencing with nanopores', *Nat Biotechnol,* 30 (4), 326-8.

Seleme, M. C., et al. (2006), 'Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity', *Proceedings of the National Academy of Sciences of the United States of America,* 103 (17), 6611-16.

Sen, S. K., et al. (2006), 'Human genomic deletions mediated by recombination between Alu elements', *Am J Hum Genet,* 79 (1), 41-53.

Shen, M. R., Batzer, M. A., and Deininger, P. L. (1991), 'Evolution of the master Alu gene(s)', *J Mol Evol,* 33 (4), 311-20.

Shin, H., et al. (2014), 'Variation in RNA-Seq transcriptome profiles of peripheral whole blood from healthy individuals with and without globin depletion', *PLoS One,* 9 (3), e91041.

Smit, A. F. and Riggs, A. D. (1996), 'Tiggers and DNA transposon fossils in the human genome', *Proc Natl Acad Sci U S A,* 93 (4), 1443-8.

Smit, A. F. A., Hubley, R., and Green, P. (2004), 'RepeatMasker Open-3.0'.

Smit, A. F. A., Hubley, R., and Green, P. 'RepeatMasker Open-4.0.', <http://www.repeatmasker.org>, accessed.

Smith, S. A. and Donoghue, M. J. (2008), 'Rates of molecular evolution are linked to life history in flowering plants', *Science,* 322 (5898), 86-9.

Steely, C. J., et al. (2018), 'Analysis of lineage-specific *Alu* subfamilies in the genome of the olive baboon, *Papio anubis*', *Mob DNA,* 9, 10.

Stewart, C., et al. (2011), 'A comprehensive map of mobile element insertion polymorphisms in humans', *PLoS genetics,* 7 (8), e1002236.

Symer, D. E., et al. (2002), 'Human l1 retrotransposition is associated with genetic instability in vivo', *Cell,* 110 (3), 327-38.

Szak, S. T., et al. (2003), 'Identifying related L1 retrotransposons by analyzing 3' transduced sequences', *Genome Biol,* 4 (5), R30.

Tang, Wanxiangfu and Liang, Ping (2019), 'Comparative genomics analysis reveals high levels of differential retrotransposition among primates from the Hominidae and the Cercopithecidae families', *Genome Biology and Evolution*.

Tang, Wanxiangfu, et al. (2018), 'Mobile elements contribute to the uniqueness of human genome with 15,000 human-specific insertions and 14 Mbp sequence increase', *DNA Research*, Pages 521–33.

Tarailo-Graovac, M. and Chen, N. (2009), 'Using RepeatMasker to identify repetitive elements in genomic sequences', *Curr Protoc Bioinformatics,* Chapter 4, Unit 4 10.

Tarasov, A., et al. (2015), 'Sambamba: fast processing of NGS alignment formats', *Bioinformatics,* 31 (12), 2032-4.

Thomas, J., Perron, H., and Feschotte, C. (2018), 'Variation in proviral content among human genomes mediated by LTR recombination', *Mob DNA,* 9, 36.

Trizzino, M., et al. (2017), 'Transposable elements are the primary source of novelty in primate gene regulation', *Genome Res,* 27 (10), 1623-33.

Tutar, Y. (2012), 'Pseudogenes', *Comp Funct Genomics,* 2012, 424526.

Wallace, N., et al. (2008), 'LINE-1 ORF1 protein enhances Alu SINE retrotransposition', *Gene,* 419 (1-2), 1-6.

Wang, H., et al. (2005), 'SVA elements: a hominid-specific retroposon family', *J Mol Biol,* 354 (4), 994-1007.

Wang, J., et al. (2006), 'Whole genome computational comparative genomics: A fruitful approach for ascertaining Alu insertion polymorphisms', *Gene,* 365, 11-20.

Ward, M. C., et al. (2013), 'Latent regulatory potential of human-specific repetitive elements', *Mol Cell,* 49 (2), 262-72.

Warren, W. C., et al. (2015), 'The genome of the vervet (Chlorocebus aethiops sabaeus)', *Genome Res,* 25 (12), 1921-33.

Wheelan, S. J., et al. (2005), 'Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution', *Genome Res,* 15 (8), 1073-8.

Wilder, J. and Hollocher, H. (2001), 'Mobile elements and the genesis of microsatellites in dipterans', *Mol Biol Evol,* 18 (3), 384-92.

Xing, J., et al. (2006), 'Emergence of primate genes by retrotransposon-mediated sequence transduction', *Proceedings of the National Academy of Sciences of the United States of America,* 103 (47), 17608-13.

Yan, G., et al. (2011), 'Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques', *Nat Biotechnol,* 29 (11), 1019-23.

Zerbino, D. R., et al. (2018), 'Ensembl 2018', *Nucleic Acids Res,* 46 (D1), D754-D61.

Zhang, Q., Arbuckle, J., and Wessler, S. R. (2000), 'Recent, extensive, and preferential insertion of members of the miniature inverted-repeat transposable element family Heartbreaker into genic regions of maize', *Proc Natl Acad Sci U S A,* 97 (3), 1160-5.

Appendix

Please see the attached excel file for the appendix tables due to their large sizes

Chapter 2:

I. A list of species-specific mobile elements (SS-MEs) in potential protein coding genes in the eight primate genomes

Chapter 3:

II. Detailed list of retro-DNA elements located in genic regions in the 10 primate genomes

III. Detailed list of retro-DNA elements and their potential parent sites in the 10 primate genomes

IV. Expression level of retro-DNA elements and potential parent sites in the 7 primate transcriptomes