


# Comparative Genomics Analysis Reveals High Levels of Differential Retrotransposition among Primates from the *Hominidae* and the *Cercopithecidae* Families

Wanxiangfu Tang and Ping Liang  \*

Department of Biological Sciences, Brock University, St. Catharines, Ontario, Canada

\*Corresponding author: E-mail: [pliang@brocku.ca](mailto:pliang@brocku.ca).

Accepted: October 24, 2019

**Data deposition:** The complete lists of SS-MEs for the eight genomes with the genome coordinates and ME classification are freely available as delimited text files at [http://genomics.brocku.ca/PublicationData/Primate\\_SS\\_ME/](http://genomics.brocku.ca/PublicationData/Primate_SS_ME/).

## 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; and 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.

**Key words:** primates, comparative genomics, ME transposition, evolution.

## 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 (Lander et al. 2001; Deininger et al. 2003; Chimpanzee Sequencing and Analysis Consortium 2005; Rhesus Macaque Genome Sequencing and Analysis Consortium et al. 2007; Cordaux and Batzer 2009; Locke et al. 2011; Yan et al. 2011; Scally et al. 2012; Carbone et al. 2014). This percentage is expected to increase slightly in these genomes from further

improvements of the genome sequences and repeat annotation, especially for the nonhuman primate genomes.

By the mechanism of their transposition, MEs can be divided into two major classes: DNA transposons and retrotransposons (Stewart et al. 2011). DNA transposons move in the genome in a “cut and paste” style, for which they were initially called “jumping genes” (McClintock 1950; Deininger et al. 2003). 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 et al. 2007). DNA transposons constitute ~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 an 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 (Deininger et al. 2003; Cordaux and Batzer 2009). 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-RV/NTR/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) (Grindley 1978; Allet 1979). 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 (Symer et al. 2002; Szak et al. 2003; Han et al. 2004, 2005, 2007; Callinan et al. 2005; Wheelan et al. 2005; Sen et al. 2006; Konkel and Batzer 2010; Quinn and Bubb 2014; Chuong et al. 2016; Mita and Boeke 2016; Trizzino et al. 2017; Bourque et al. 2018). MEs also contribute to genetic diseases in human via both germline and somatic insertions (Goodier 2016; Anwar et al. 2017).

Furthermore, MEs have intimate associations with other repetitive elements such as microsatellite repeats/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/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 (SS-MEs) can provide novel regulatory sites in the genome, which can potentially regulate nearby genes’ expression, and ultimately lead to 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 (Ray et al. 2005; Battilana et al. 2006; Seleme et al. 2006; Wang et al. 2006; Jha et al. 2009; Ewing and Kazazian 2011; Stewart et al. 2011). 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 (Benit et al. 2003; Wang et al. 2005; Mills et al. 2007; Beck et al. 2010; Ahmed et al. 2013; Thomas et al. 2018). Besides these, limited analyses of SS-MEs have also been performed in a few primate genomes. The first of such study was done by Mills et al. (2006), who analyzed SS-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. However, these early studies of SS-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 multiway comparative genomic approach, leading to the identification of 14,870 human-specific MEs, which contribute to 14.2-Mb net genome sequence increase (Tang et al. 2018). Other studies focused on SS-MEs target on either one particular ME type and/or a few primate genomes. For example, Navarro and Galante performed comparative analysis of retrogenes (processed pseudogenes) in seven primate genomes (Navarro and Galante 2015), while Steely et al. (2018) recently ascertained 28,114 baboon-specific Alu elements by comparing the genomic sequences of baboon to both rhesus macaque and human genomes.

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 multiway comparative genomic approach used for identifying human-specific MEs to analyze SS-MEs in eight primate genomes, representing the *Hominidae* family and the

*Cercopithecidae* family of the primates. Our analysis identified a total of 230,855 SS-MEs in these genomes, which collectively contribute to ~82 Mb genome sequences, revealing significant differential ME transposition among primate species.

## Materials and Methods

### 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>; last accessed November 12, 2019) 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 (December 2014, NCBI project 31265/gorGor4.1), and orangutan genome (July 2007, WUSTL version Pongo\_albelii-2.0.2/ponAbe2). For gorilla genome, there is a newer version (March 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 (March 2014, VGC Chlorocebus\_sabeus-1.1/chlSab2), crab-eating macaque genome (June 2013, WashU Macaca\_fascicularis\_5.0/macFas5), rhesus monkey genome (November 2015, BCM Mmul\_8.0.1/rheMac8), and baboon (Anubis) genome (March 2012, Baylor Panu\_2.0/papAnu2). The information regarding the sequencing platforms and the genome assembly quality is provided in [supplementary table S1, Supplementary Material](#) online.

### Identification of 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/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 (Kent 2002; Hinrichs et al. 2006).

### 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 nonhuman primate

genomes, were generated on a local server using a modified version of UCSC pipeline RunLastzChain (<http://genome.ucsc.edu>; last accessed November 12, 2019).

### Preprocessing 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/deletions as individual ME entries, we performed a preprocess 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.

### 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>; last accessed November 12, 2019), 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.

### 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/>; last accessed November 12, 2019). The GenBank accessions for the nine *ACTB* sequences used in the analysis include NM\_0011101.5 (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 ([supplementary fig. S1, Supplementary Material](#) online). 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×(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.

### Identification of TSDs, Transductions, and Insertion-Mediated Deletions

The TSDs, as well as transductions and insertion-mediated deletions (IMDs) for all SS-MEs, were identified using in-house Perl scripts as described previously (Tang et al. 2018). For each of those with TSDs successfully identified, a 30-bp sequence centered at each insertion site in the predicted preintegration alleles was 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 were 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 preintegration site in the outgroup genomes, they were considered candidates for IMDs, which were subject to further validation.

### 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 the genome using BLAT (minScore  $\geq 100$ ; minIdentity  $\geq 97\%$ ). Those showing a 100% sequence identity with another copy of SS-ME (nonself-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).

### 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 nonhuman primate genomes. The sequences of each genome were divided into a nonredundant list of categorized regions in gene context, including CDS, noncoding 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.

### 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://compu-tecanada.ca>).

## Results

### 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 [supplementary table S2, Supplementary Material](#) online, 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 ([supplementary table S2, Supplementary Material](#) online). By copy number from low to high among the genomes, SINEs as the most successful MEs have 1,631,626 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 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 have 2,328 copies in the orangutan to 4,931 and 4,933 copies in chimpanzee and human, respectively ([supplementary table S2, Supplementary Material](#) online). 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

**Table 1**

Species-Specific Mobile Elements (SS-MEs) and Most Recent SS-MEs in Eight Primate Genomes

ME Class	Human			Chimp			Gorilla			Orangutan		
	Raw	Normalized	MR SS-MEs	Raw	Normalized	MR SS-MEs	Raw	Normalized <sup>a</sup>	MR SS-MEs	Raw	Normalized <sup>a</sup>	MR SS-MEs
SINE	8,844	7,175	4,775	10,612	8,610	2,309	6,324	3,399	2,105	9,630	2,556	172
LINE	3,946	3,201	2,736	7,288	5,913	3,595	4,085	2,196	2,197	21,711	11,717	11,717
SVA	1,571	1,275	658	1,597	1,296	564	877	471	397	1,180	313	242
LTR	530	430	110	1,924	1,561	175	689	370	147	2,933	779	107
Total	14,891	12,081	8,279	21,421	17,379	6,643	11,975	6,437	4,846	35,454	15,365	12,238
		Rhesus			Crab-Eating Macaque			Baboon			Green Monkey	
SINE	22,069	22,069	4,083	2,257	2,257	416	56,247	54,969	25,292	34,277	15,515	7,922
LINE	3,016	3,016	1,217	782	782	411	8,407	8,216	6,376	11,981	5,928	5,928
SVA	2	2	0	0	0	0	0	0	0	3	1	0
LTR	1,346	1,346	107	234	234	50	1,764	1,724	268	2,324	1,052	198
Total	26,433	26,433	5,407	3,273	3,273	877	66,418	64,909	31,936	48,585	22,496	14,048

<sup>a</sup>Normalized numbers in gray highlights were lower but manually adjusted to be the same as the most recent SS-MEs.

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 (supplementary table S2, Supplementary Material online). Collectively, MEs from these five major classes constitute from 46.8% (green monkey) to 50.7% (baboon) to the genomes (supplementary table S2, Supplementary Material online). All retrotransposons together contribute from 43.3% in the green monkey genome to 47.1% in the baboon genome (supplementary table S2, Supplementary Material online). DNA transposons were excluded from further analyses in this study due to their smaller percentages and very low-activity levels in these genomes.

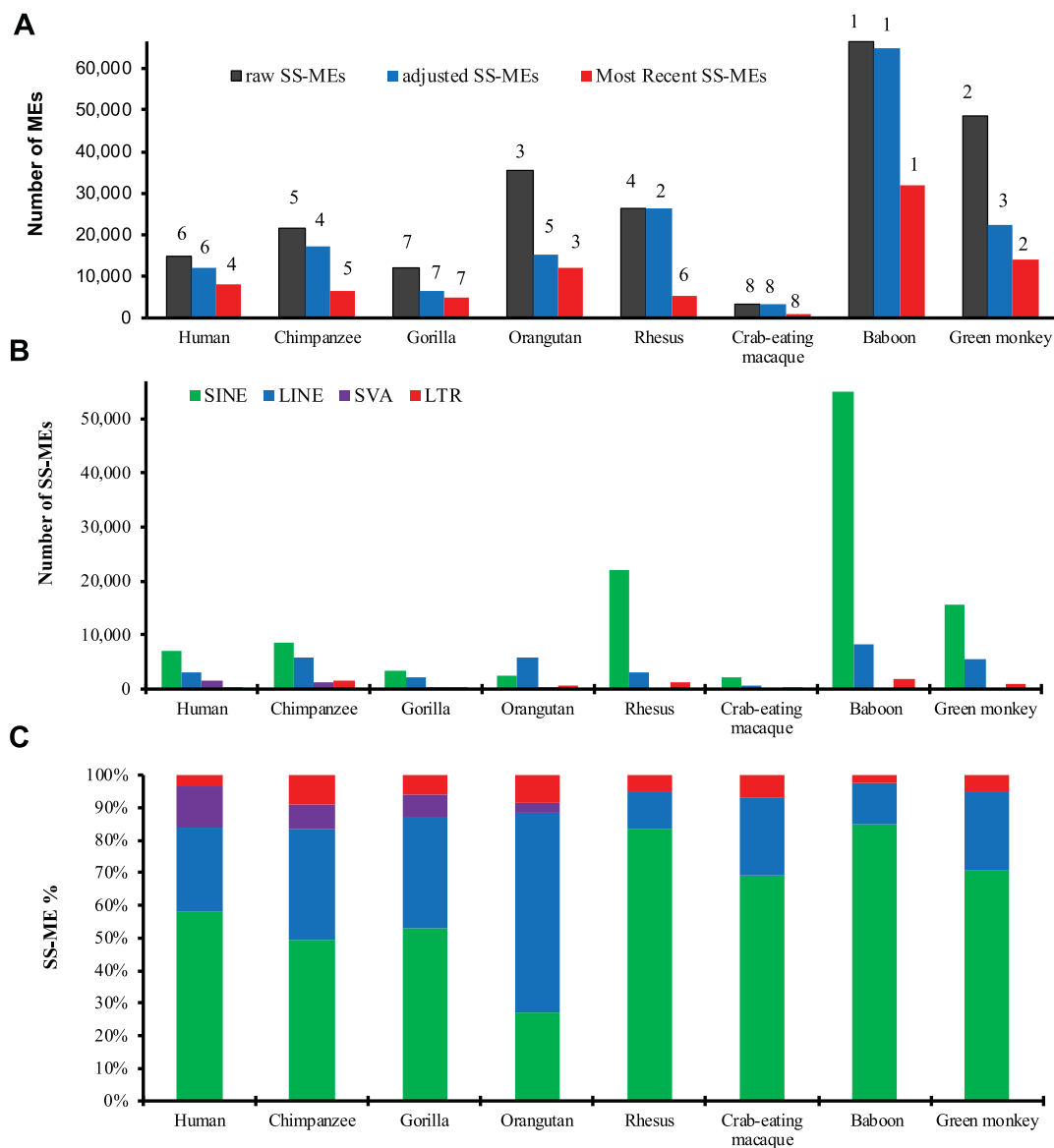
### Differential Level of SS-MEs in Primate Genomes

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 multiway 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 (supplementary table S3, Supplementary Material online). 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.

As seen in table 1 and figure 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 four 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 1).

It is worth mentioning here that while a few factors associated with the variable quality of the genome assemblies and ME annotation 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%, supplementary table S1, Supplementary Material online), and as one would expect, the total numbers of MEs (after integration) in these genomes are quite similar to each other (supplementary table S2, Supplementary Material online) with variation <7% (data not shown), further confirming the qualities of genome



**Fig. 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 one being the highest and eight 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).

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 third 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 1 and fig. 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. This 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*) CDS collected from NCBI

(supplementary fig. S1, Supplementary Material online). As shown in table 1, 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. Although 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 third to fifth due to its largest distance from the other genomes and the latter moved up by two from the fourth to the second due to its shortest evolutionary distance, while the chimpanzee genome moved up by one position (fig. 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 second least number of SS-MEs, respectively, while the human genome remains as the sixth. Further analyses from this point on were based on normalized SS-MEs unless otherwise specified.

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 second highest genome (rhesus, 22,069) and more than three times higher than all genome average (14,569) (table 1; supplementary table S3, Supplementary Material online; and fig. 1B and 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 (Steely et al. 2018; Rogers et al. 2019). 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 1 and fig. 1A). In contrast, the number of SS-SINEs in orangutan is significantly lower than that of all other genomes (2,556 vs.  $\geq 3,399$ ) except for crab-eating macaque 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) seemingly in negative correlation with the evolutionary ages. Although between 100 and 200 MacSVAs are present in the

*Cercopithecidae* genomes, no more than three or zero SS-MacSVAs are detected (supplementary table S2, Supplementary Material online and table 1), and thus they were excluded from further analysis.

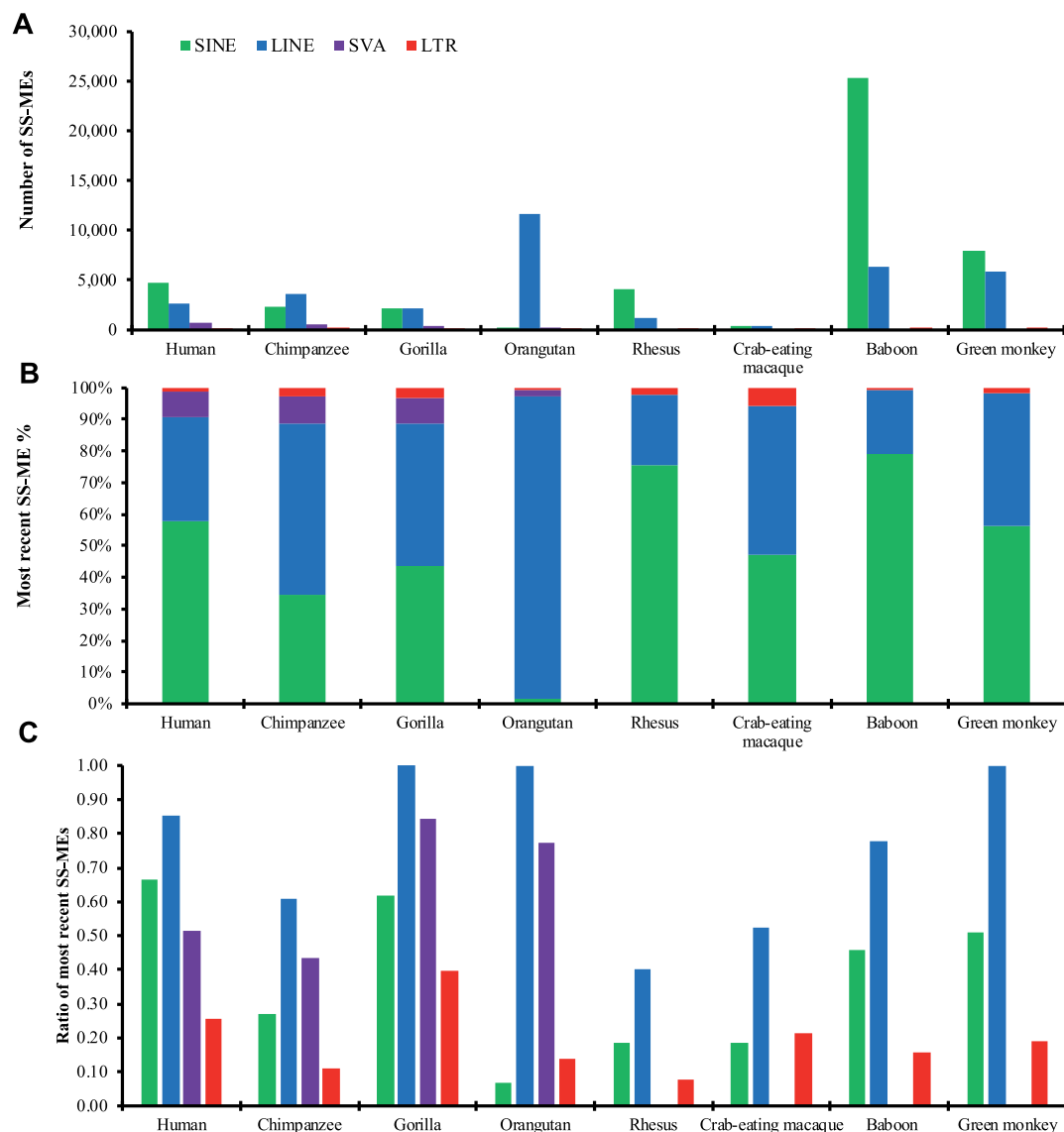
It is worth noting that for all genomes except for crab-eating macaque have one or more ME class being very successful (e.g., baboon for SINE and LTR, orangutan for LINE, and human for SVA) or moderately successful (e.g., rhesus and green monkey genomes for SINE and chimpanzee for LTR). In contrast the extreme low number of SS-MEs applies to all ME classes in the crab-eating macaque genome (table 1 and 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 four 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) (supplementary table S3, Supplementary Material online). The slightly higher level of SS-LTRs in the *Cercopithecidae* family (1,089 vs. 785 for *Hominidae*) also contributes to these differences. Interestingly, while the level of ME accumulation seems to be more or less similar (within one order of differences) among the *Hominidae* genomes, it differs dramatically (more than one 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 1 and fig. 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 figure 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 figure 1B.

#### 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 ( $\geq 100$  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



**Fig. 2.**—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.

focusing on the SS-MEs resulted from the narrowest window (compared with if a lower stringency, e.g., 98% sequence similarity, was used) of species evolution toward 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, seemingly 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.

The overall trend for the total number of most recent SS-MEs among the genomes is similar to that of normalized SS-MEs (table 1 and fig. 1A). Like for the raw and normalized SS-MEs, the baboon genome keeps its first position as having the highest number of most recent SS-MEs (31,936), while the crab-eating macaque genome has the lowest number



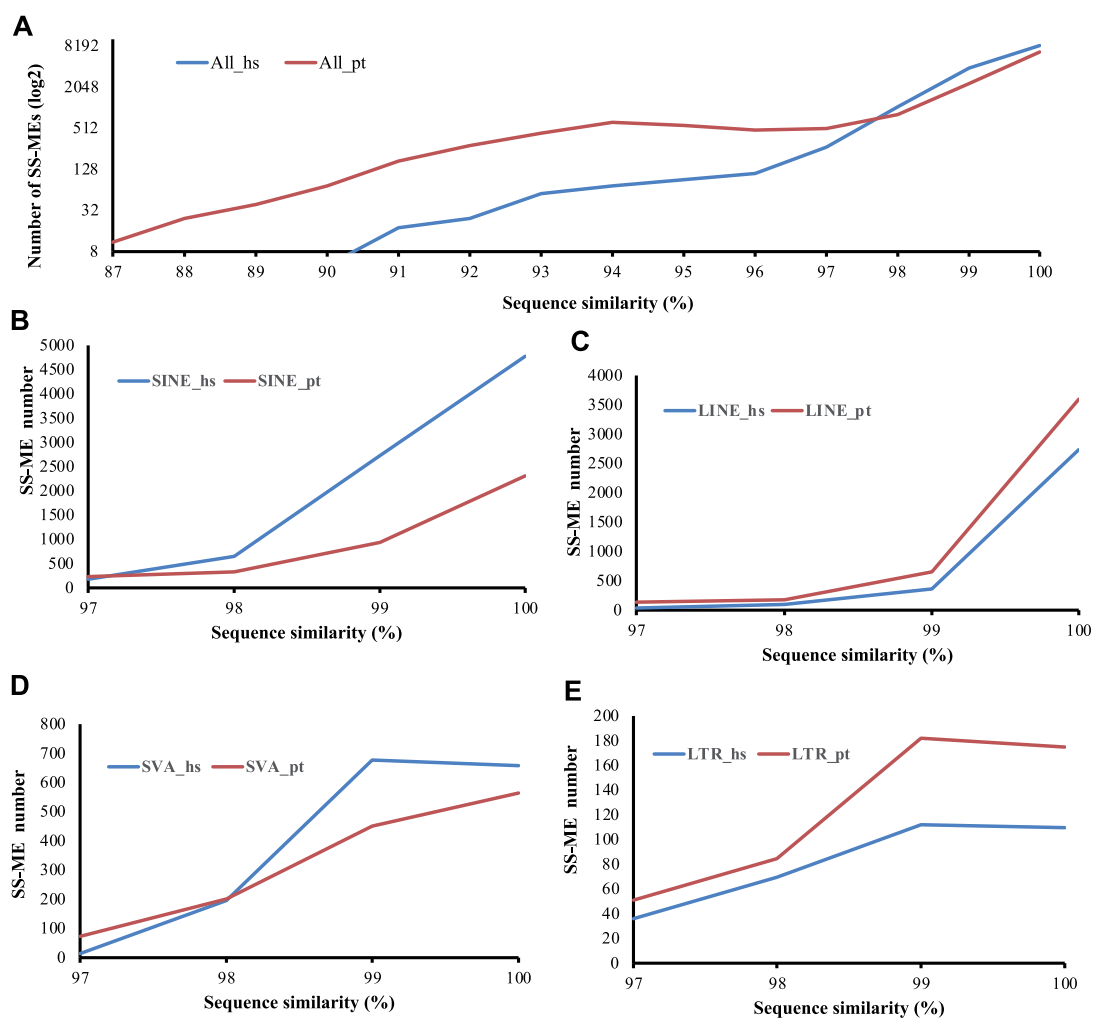
(877), and gorilla genome has the second least number (4,846), making the ranking of these three genomes being the same by all three sets of SS-ME numbers (table 1 and fig. 1A). The overall patterns of the most recent SS-ME profiles by ME class in number and percentage are also more or less similar to these of the normalized SS-MEs (fig. 2A vs. fig. 1B for numbers and fig. 2B vs. fig. 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. 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. 2B) as in the case of SS-MEs (fig. 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 numbers of SS-MEs and most recent SS-MEs both with a strong bias for SS-SINEs (figs. 1A, 2A, and 1B). The rhesus genome had the largest drop in ranking from the second for normalized SS-MEs to the sixth position by the number of most recent SS-MEs, while the orangutan and human genomes had the largest increase from the fifth to the third and from the sixth to the fourth, respectively. It is also worth noting that between human and chimpanzee, which are mutually the closest among the eight genomes, the ranking moved up two positions for human, but moved down one position for chimpanzee. Although 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 1 and fig. 1A). Another interesting difference is the much stronger dominance of LINEs in the most recent SS-MEs (~99%) (fig. 2B) than in the SS-MEs (~85%) (fig. 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 (117,171 vs. 4,722) (table 1 and supplementary table S3, Supplementary Material online). 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 1). 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

figure 2C, 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. 2A). 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 1). These results indicate that the high success of SINEs and other non-LTR retrotransposons always requires the support of activity of LINEs or L1s to be more specific.

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 figure 3, 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 toward the more recent period as reflected by the higher ratios of SS-MEs at high-sequence similarity levels (fig. 3A). 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 than the chimpanzee genome (fig. 3B and D). The chimpanzee genome showed a higher most recent activity for LINEs and LTRs (fig. 3C and 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 toward 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. 3D). 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 (supplementary fig. S2 and table S4, Supplementary Material online), supporting SVA-E and SVA-F being human-specific (Wang et al. 2005).



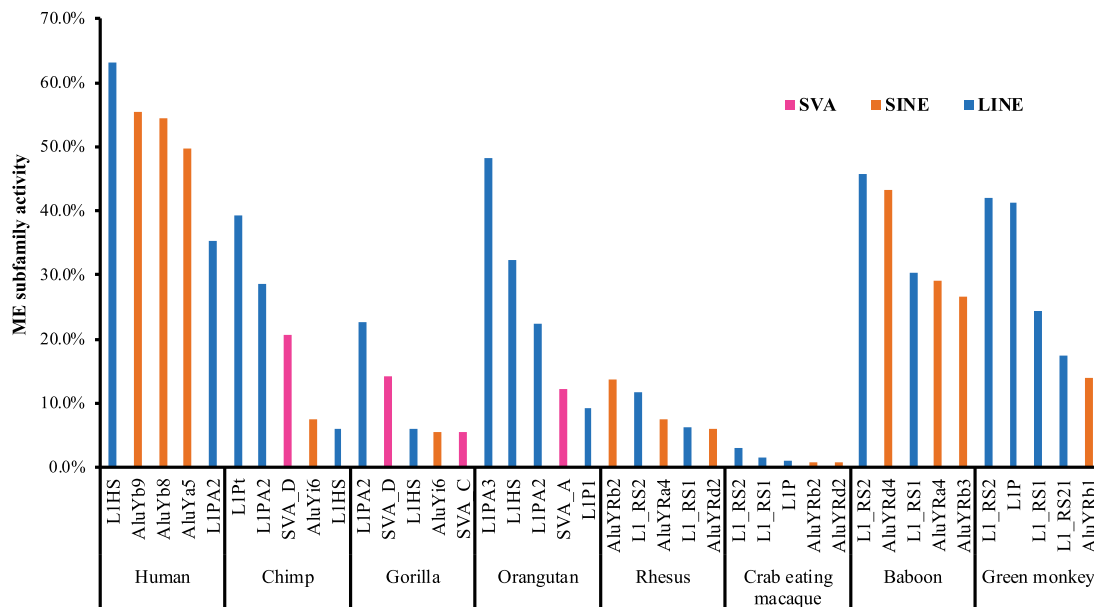
**Fig. 3.**—The comparison of activity profiles of species-specific mobile elements (SS-MEs) in the human and chimpanzee genomes. (A) The numbers of SS-MEs with sequence similarity at 87% or more to another copy of SS-MEs in the human and chimpanzee genomes with  $y$  axis shown in log<sub>2</sub> scale. (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 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. [Supplementary table S4, Supplementary Material online](#), 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 ([supplementary table S4, Supplementary Material online](#)). A visual representation for active ME subfamilies and their relative activity levels in the

eight genomes is shown as a heatmap ([supplementary fig. S2, Supplementary Material online](#)), while the top five active ME subfamilies in each genome were also shown in [figure 4](#). As shown in [supplementary figure S2, Supplementary Material online](#) and [figure 3](#), each genome has a unique profile of active MEs that differ not only by ME subfamilies but also by their relative activity levels.

In consistent with having the largest number of SS-SINEs, the baboon genome has the largest number of Alu subfamilies at high activities (ten subfamilies at 10% or more) despite none being the highest among all genomes ([supplementary table S4, Supplementary Material online](#) and [fig. 4](#)). Similarly, the orangutan genome has the largest number of recently active L1 subfamilies (4 at 17% or more) and with four of its top five active ME subfamilies being from LINES, all at relatively high activity, explaining its largest number of SS-LINES.



**Fig. 4.**—Most active subfamilies of mobile elements (MEs) in the eight primate genomes. The top five 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.

Next to the orangutan genome, the green monkey genome also seems to have a high level of recent L1 activity by having four of the five top active ME subfamilies from L1, all with relatively high levels of activity, supporting its high number of most recent SS-LINEs (table 1; supplementary table S4, Supplementary Material online; and figs. 2A and 4). 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 fifth ME subfamily (L1PA2) and three SVA subfamilies also have the highest among the same subfamilies from all *Hominidae* genomes (supplementary table S4 and fig. S2, Supplementary Material online). 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 six subfamilies got onto the list of active ME subfamilies with activity >1% (supplementary table S4 and fig. S2, Supplementary Material online). 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 (supplementary table S4 and fig. S2, Supplementary Material online). 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 with the non-LTR retrotransposons (supplementary table S4, Supplementary Material online).

It is worth to note that, in contrast with all other genomes, the crab-eating macaque genome lacks a single highly active ME subfamily (supplementary table S4, Supplementary Material online, fig. 4, and supplementary fig. S2, Supplementary Material online) 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.

#### 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 IMDs 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. As shown in table 2, in all eight genomes, SS-MEs have led to a net genome size increase. Collectively, SS-MEs have contributed to a combined ~82.3-Mb increase in the eight genomes or on an average ~10 Mb per genome or ~7 Mb 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 (~1.1 Mb), which is directly correlated

**Table 2**

Impact of Species-Specific Mobile Elements (SS-MEs) on Genome Size (kb)

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

TSD, target site duplications; IMD, insertion-mediated deletions.

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 IMDs contributing to a significant amount of size loss offsetting the size increases from the insertions (table 2).

### 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 on the gene annotation data in human from the GENCODE project (Release July 23, 2015) (Harrow et al. 2012) combined with the NCBI RefGene annotation set (Pruitt et al. 2007) and ENSEMBL gene annotation data for the nonhuman 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.

As shown in [supplementary table S5, Supplementary Material](#) online, a total of 76,646 SS-MEs, representing ~33.5% of all SS-MEs, are located in genic regions, which include protein-coding genes, noncoding 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 ([supplementary tables S6 and S7, Supplementary Material](#) online).

## 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 ([supplementary table S3, Supplementary Material](#) online). 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.

### 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 nonhuman 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; Tarailo-Graovac and Chen 2009; Smit et al. 2013), and 5) variable mutation rate across species (Sally and Durbin 2012), which could have an impact on the analysis of the most-recent SS-MEs based on a sequence similarity cutoff. For nonhuman primate genomes, the second and third issues are larger than for human genome due to the generally lower quality of the reference genome assemblies ([supplementary table S1, Supplementary Material](#) online). The gap regions are usually biased toward 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 fourth 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 <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 <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 four 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 nonhuman 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 (Schneider and Dekker 2012; Roberts et al. 2017). 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).

### The Differential ME Transposition among Primate Genomes

Despite more and more nonhuman 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 (Ray et al. 2005; Battilana et al. 2006; Mills et al. 2006; Wang et al. 2006; Jha et al. 2009; Ewing and Kazazian 2011; Stewart et al. 2011; Jordan et al. 2018; Steely et al. 2018; Tang et al. 2018). So far, only very limited comparative analyses involving a small number of genomes have been reported. Among these, the work by 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 earlier, 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. 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 gorilla genomes (table 1 and fig. 1A). Although 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/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 (supplementary table S4 and fig. S2, Supplementary Material online). 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 (Steely et al. 2018; Rogers et al. 2019). The fact that, despite the great success of *Alu* transposition in the baboon genome, none of the active *Alu* subfamilies has 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 acceleration for SINES/*Alus* (fig. 3B).

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,  $\sim 1/16$  of *Cercopithecidae* family average, and  $\sim 1/35$  of that for the baboon genome (table 1 and supplementary table S3, Supplementary Material online). 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 1 and figs. 1, 2, and 4) strongly suggests 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 hypermethylation 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 based on 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 estimation 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>; last accessed November 12, 2019) (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 Myr among the 36 studies collected in the TimeTree database, and TimeTree provides 12.4 Myr as the estimate for the distance between the two species. This is larger than the distance between gorilla and human (9.06 Myr from TimeTree). Although 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 (supplementary fig. S1A, Supplementary Material online), with the distance between baboon and the two macaque species being much closer. Although 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. 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 are 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 toward 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 the profiles 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 1 and fig. 2C). 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 1). The ratio of the most recent SS-SINEs is 67% for human genome compared with ~27% for chimpanzee genome (fig. 2C). Higher ratios of most recent SS-MEs in the human genome are also seen for LINES and LTRs (fig. 2C), despite their numbers being lower than the counterparts in the chimpanzee genome (table 1). These data may suggest that, relatively speaking between the two genomes as shown in figure 3A, 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. 4 and supplementary table S4, Supplementary Material online). 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 to the larger total number of most recent SS-MEs in the human genome than in the chimpanzee genome (fig. 3A, B, and 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 numbers of SS-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 numbers 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 multiway 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. 4 and supplementary fig. S2 and table S4, Supplementary Material online), and this is in good agreement with results of two recent studies (Steely et al. 2018; Rogers et al. 2019). 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. 4), 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 J, et al., 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. 2C) regardless of the overall level of SS-MEs. This is expected, as the functional L1 retrotransposition machinery is required 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.

### 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-Mb total net increase in these genomes (table 2) with a net increase in each genome, ranging from ~1.2 Mb in the crab-eating macaque genome to ~25.5 Mb 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 addition 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 recent review by 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, noncoding RNAs, and transcribed pseudogenes (supplementary table S5, Supplementary Material online). 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 nonhuman 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 CDS in a transcript. As shown in supplementary table S7, Supplementary Material online, most of these CDS SS-MEs are SS-SINEs (119/251), even more so in 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 nonhuman 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 nonhuman 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).

### 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 being more successful. Remarkable differences in ME transposition are also seen among closely related genomes, as observed between human and chimpanzee genomes with ME transposition showing a later and quicker acceleration in the human genome compared with 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.

## Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

## Acknowledgments

This work is in part supported by grants from the Canadian Research Chair program, Canadian Foundation of Innovation, Ontario Ministry of Research and Innovation, Canadian Natural Science and Engineering Research Council (NSERC), and Brock University to P.L., and was made possible by Compute Canada/SHARCNET high-performance computing facilities.

## Literature Cited

- Ahmed M, Liang P. 2012. Transposable elements are a significant contributor to tandem repeats in the human genome. *Comp Funct Genomics*. 2012:1.
- Ahmed M, Li W, Liang P. 2013. Identification of three new Alu Yb sub-families by source tracking of recently integrated Alu Yb elements. *Mob DNA*. 4(1):25.
- Allet B. 1979. Mu insertion duplicates a 5 base pair sequence at the host inserted site. *Cell* 16(1):123–129.
- Anwar SL, Wulaningsih W, Lehmann U. 2017. Transposable elements in human cancer: causes and consequences of deregulation. *Int J Mol Sci*. 18(5):974.
- Battilana J, et al. 2006. Alu insertion polymorphisms in Native Americans and related Asian populations. *Ann Hum Biol*. 33(2):142–160.
- Baucom RS, et al. 2009. Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genet*. 5(11):e1000732.
- Beck CR, et al. 2010. LINE-1 retrotransposition activity in human genomes. *Cell* 141(7):1159–1170.
- Benit L, Calteau A, 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):159–168.
- Bourque G, et al. 2018. Ten things you should know about transposable elements. *Genome Biol*. 19(1):199.
- Callinan PA, et al. 2005. Alu retrotransposition-mediated deletion. *J Mol Biol*. 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):3497–3500.
- Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437(7055):69–87.
- Chuong EB, Elde NC, Feschotte C. 2016. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* 351(6277):1083–1087.
- Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. *Nat Rev Genet*. 10(10):691–703.
- Deininger PL, et al. 2003. Mobile elements and mammalian genome evolution. *Curr Opin Genet Dev*. 13(6):651–658.
- Ewing AD, Kazazian HH Jr. 2011. Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. *Genome Res*. 21(6):985–990.
- Goodier JL. 2016. Restricting retrotransposons: a review. *Mob DNA*. 7:16.
- Grindley ND. 1978. IS1 insertion generates duplication of a nine base pair sequence at its target site. *Cell* 13(3):419–426.
- Han JS, Szak ST, Boeke JD. 2004. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* 429(6989):268–274.
- Han K, et al. 2005. Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. *Nucleic Acids Res*. 33(13):4040–4052.
- Han K, et al. 2007. Alu recombination-mediated structural deletions in the chimpanzee genome. *PLoS Genet*. 3(10):e184–e249.
- Harris RS. 2007. Improved pairwise alignment of genomic DNA. Pennsylvania State University. [PhD thesis]. [http://www.bx.psu.edu/~rsharris/rsharris\\_phd\\_thesis\\_2007.pdf](http://www.bx.psu.edu/~rsharris/rsharris_phd_thesis_2007.pdf)
- Harrow J, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*. 22(9):1760–1774.
- Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* 22(23):2971–2972.
- Herron PR. 2004. Mobile DNA II. *Heredity* 92(5):476–476.
- Hinrichs AS, et al. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res*. 34(9):D590–D598.
- Jha AR, 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*. *Mol Biol Evol*. 26(11):2617–2626.
- Jiao Y, et al. 2017. Improved maize reference genome with single-molecule technologies. *Nature* 546(7659):524–527.
- Jordan VE, et al. 2018. A computational reconstruction of *Papio* phylogeny using *Alu* insertion polymorphisms. *Mob DNA*. 9:13.
- Jurka J, et al. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*. 110(1–4):462–467.
- Kazazian HH Jr. 2004. Mobile elements: drivers of genome evolution. *Science* 303(5664):1626–1632.
- Kent WJ. 2002. BLAT—the BLAST-like alignment tool. *Genome Res*. 12(4):656–664.
- Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* 115(1):49–63.
- Konkel MK, Batzer MA. 2010. A mobile threat to genome stability: the impact of non-LTR retrotransposons upon the human genome. *Semin Cancer Biol*. 20(4):211.
- Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409(6822):860.
- Law JA, Jacobsen SE. 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet*. 11(3):204–220.
- Lee SI, Kim NS. 2014. Transposable elements and genome size variations in plants. *Genomics Inform*. 12(3):87–97.
- Locke DP, et al. 2011. Comparative and demographic analysis of orangutan genomes. *Nature* 469(7331):529–533.
- McClintock B. 1950. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A*. 36(6):344–355.
- Mills RE, et al. 2007. Which transposable elements are active in the human genome? *Trends Genet*. 23(4):183–191.
- Mills RE, et al. 2006. Recently mobilized transposons in the human and chimpanzee genomes. *Am J Hum Genet*. 78(4):671–679.
- Mita P, Boeke JD. 2016. How retrotransposons shape genome regulation. *Curr Opin Genet Dev*. 37:90–100.



- Navarro FC, Galante PA. 2015. A genome-wide landscape of retrocopies in primate genomes. *Genome Biol Evol.* 7(8):2265–2275.
- Oliver KR, Greene WK. 2011. Mobile DNA and the TE-Thrust hypothesis: supporting evidence from the primates. *Mob DNA.* 2(1):8.
- Pace I, John K, Feschotte C. 2007. The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. *Genome Res.* 17(4):422–424.
- Penzkofer T, et al. 2017. L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes. *Nucleic Acids Res.* 45(D1):D68–D73.
- Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 35(Database):D61–D65.
- Quinn JP, Bubb VJ. 2014. SVA retrotransposons as modulators of gene expression. *Mob Genet Elem.* 4(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):415–425.
- Ray DA, et al. 2005. Inference of human geographic origins using Alu insertion polymorphisms. *Forensic Sci Int.* 153(2–3):117–124.
- Rhesus Macaque Genome Sequencing and Analysis Consortium, et al. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316(5822):222–234.
- Roberts RJ, Carneiro MO, Schatz MC. 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.
- Sally A, Durbin R. 2012. Revising the human mutation rate: implications for understanding human evolution. *Nat Rev Genet.* 13(10):745–753.
- Sally A, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* 483(7388):169–175.
- Schneider GF, Dekker C. 2012. DNA sequencing with nanopores. *Nat Biotechnol.* 30(4):326–328.
- Seleme MC, et al. 2006. Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. *Proc Natl Acad Sci U S A.* 103(17):6611–6616.
- Sen SK, et al. 2006. Human genomic deletions mediated by recombination between Alu elements. *Am J Hum Genet.* 79(1):41–53.
- Smit AFA, Hubley R, Green P. 2013. RepeatMasker Open-4.0. Available from: <http://www.repeatmasker.org>.
- Smith SA, Donoghue MJ. 2008. Rates of molecular evolution are linked to life history in flowering plants. *Science* 322(5898):86–89.
- Steely CJ, 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 Genet.* 7(8):e1002236.
- Symer DE, et al. 2002. Human L1 retrotransposition is associated with genetic instability in vivo. *Cell* 110(3):327–338.
- Szak ST, et al. 2003. Identifying related L1 retrotransposons by analyzing 3' transduced sequences. *Genome Biol.* 4(5):R30.
- Tang W, et al. 2018. Mobile elements contribute to the uniqueness of human genome with 15,000 human-specific insertions and 14 Mbp sequence increase. *DNA Res.* 25(5):521–533.
- Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics.* Chapter 4, Unit 4.10.
- Thomas J, Perron H, 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):1623–1633.
- 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 MC, et al. 2013. Latent regulatory potential of human-specific repetitive elements. *Mol Cell.* 49(2):262–272.
- Wheelan SJ, et al. 2005. Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution. *Genome Res.* 15(8):1073–1078.
- Wilder J, Hollocher H. 2001. Mobile elements and the genesis of microsatellites in dipterans. *Mol Biol Evol.* 18(3):384–392.
- 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):1019–1023.
- Zerbino DR, et al. 2018. Ensembl 2018. *Nucleic Acids Res.* 46(D1):D754–D761.

Associate editor: Richard Cordaux