Cryptic endogenous retrovirus subfamilies in the primate lineage

Xun Chen[1,#], Zicong Zhang[1], Yizhi Yan[1], Clement Goubert[2], Guillaume Bourque[1,2,3,4,#], Fumitaka Inoue[1,#]

[1] Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto 606-8501, Japan
[2] Department of Human Genetics, McGill University, Montréal, QC H3A 0C7, Canada
[3] Victor Phillip Dahdaleh Institute of Genomic Medicine at McGill University, Montréal, QC H3A 0G1, Canada
[4] Canadian Center for Computational Genomics, McGill University, Montréal, QC H3A 0G1, Canada

[#] Correspondence: Xun Chen (chen.xun.3r@kyoto-u.ac.jp), Guillaume Bourque (guil.bourque@mcgill.ca) and Fumitaka Inoue (inoue.fumitaka.7a@kyoto-u.ac.jp)

ABSTRACT

Many endogenous retroviruses (ERVs) in the human genome are primate-specific and have contributed novel cis-regulatory elements and transcripts. However, current approaches for classifying and annotating ERVs and their long terminal repeats (LTRs) have limited resolution and are inaccurate. Here, we developed a new annotation based on phylogenetic analysis and cross-species conservation. Focusing on the evolutionary young MER11A/B/C subfamilies, we revealed the presence of 4 'phyletic groups', that better explained the epigenetic heterogeneity observed within these subfamilies, suggesting a new annotation for 412 (19.8%) of the MER11 instances. Furthermore, we functionally validated the regulatory potential of these four phyletic groups using a massively parallel reporter assay (MPRA), which also identified motifs associated with their differential activities. Combining MPRA with phyletic groups across primates revealed an apes-specific gain of SOX related motifs through a single-nucleotide deletion. Lastly, by applying our approach across 53 primate-specific LTR subfamilies, we determined the presence of 75 phyletic groups and found that 3,807 (30.0%) instances from 26 LTR subfamilies could be categorized into a novel phyletic group, many of which with a distinct epigenetic profile. Thus, with our refined annotation of primate-specific LTRs, it will be possible to better understand the evolution in primate genomes and potentially identify new roles for ERV/LTRs in their hosts.

INTRODUCTION

Transposable elements (TEs) occupy nearly half of the human genome, and recent genomic and epigenomic analyses have revealed that many have been co-opted by the host (Fueyo et al. 2022; Bourque et al. 2018). In particular, at least 8% of the human genome finds its origin in endogenous retroviruses (ERVs) (Lander et al. 2001; Hoyt et al. 2022), which are often found in open chromatin and have the potential to act as genomic regulatory elements (Wang et al. 2007; Bourque et al. 2008; Sundaram et al. 2014; Jacques, Jeyakani, and Bourque 2013; Chuong, Elde, and Feschotte 2016). ERVs originate from retrovirus infections, where the viral fragment interacts with transcription factors in the host cell via its long terminal repeats (LTRs) to express viral RNA, then spreads throughout the genome by a copy-and-paste mechanism (Feschotte and Gilbert 2012). To limit the deleterious effects of uncontrolled transposition, host cells have evolved multiple defense mechanisms to silence ERVs including CpG methylation, m6A RNA methylation, RNA interference (PIWI-associated small RNA), KRAB-associated repressors genes and H3K9me3 modification (Molaro and Malik 2016; Almeida et al. 2022; Chelmicki et al. 2021). Along with this active silencing, most TE sequences, including LTRs, accumulate mutations, which eventually result in their inactivation (Jerzy Jurka 1998). Nevertheless, some LTRs may retain their regulatory activity or acquire beneficial mutations within transcription factor (TF) binding motifs (Fueyo et al. 2022). This can impact the regulatory activity of nearby gene expression, and contribute to the adaptation of the LTRs in the host genome (Grow et al. 2015; Chuong, Elde, and Feschotte 2017; Fueyo et al. 2022). Additionally, ERVs have a higher chance of survival in the next generation when they are expressed in germline or pluripotent cells during early embryonic development (Gerdes et al. 2016; Ma et al. 2022; Göke et al. 2015). In fact, several LTRs contain binding sequences for pluripotency transcription factors such as POU5F1 and SOX2 (Kunarso et al. 2010; Fueyo et al. 2022).

Many ERVs integrated into the primate genomes after the divergence from other mammals (Mayer and Meese 2005). In the process, these relatively young ERV elements have contributed a substantial number of regulatory sequences to the human genome (Jacques, Jeyakani, and Bourque 2013) and are associated with the evolution of TF binding sites during primate evolution (Andrews et al. 2023). Some subfamilies retain regulatory or transcriptional activity and have influenced human transcriptional networks (Patoori et al. 2022; Xiang et al. 2022; Bourque et al. 2008; Wang et al. 2007; Trizzino et al. 2017; Fuentes, Swigut, and Wysocka 2018; Chuong, Elde, and Feschotte 2016; Sexton, Tillett, and Han 2021). HERVH-LTR7, for instance, is considered endogenized in the primate lineage, since many of its instances (copies) have been co-opted by the host as gene regulatory elements in pluripotent stem cells (Sexton, Tillett, and Han 2021). LTR5_Hs has also been shown, using a chimeric array of gRNA oligos (CARGO) and CRISPR, to act as enhancers and potentially regulate hundreds of human genes (Fuentes, Swigut, and Wysocka 2018). Furthermore, the divergent expansion of ERVs/LTRs within individual branches of the primate lineage, provided a wealth of species-specific

enhancers, significantly influenced the regulatory network of these genomes during speciation (Senft and Macfarlan 2021).

The proper classification and annotation of ERV/LTR instances is critical to understand their evolution and potential impact on the host (Sotero-Caio et al. 2017). The standard approach used to characterize TE instances relies on homology between genomic sequences and curated TE libraries, and aims to attribute a unique family or subfamily name to a group of monophyletic sequences (Wicker et al. 2007). Although an important effort of manual curation has been applied to the TEs in the human lineage, a correct classification and annotation of these repeat elements remains a challenging problem (Hoen et al. 2015; Arkhipova 2017; Carey et al., 2021; Hassan and Adelson 2023). Historical nomenclature discrepancy, a high degree of sequence similarity between related yet distinct monophyletic groups, as well as recombination events within ERVs or between ERVs and exogenous viruses has led to various misclassifications (Blomberg et al. 2009). Furthermore, instances of an established ERV/LTR subfamily continuously evolve further into divergent sub-lineages, introducing another layer of complexity (Carter et al. 2022; Le Rouzic, Boutin, and Capy 2007). Phylogenetic analyses have been used to study ERV/LTR sequence evolution with either a small set of full length ERV or solo LTR sequences (Grandi et al. 2021; Scognamiglio et al. 2023), or to interpret their regulatory heterogeneity (Ito et al. 2017). Unfortunately, most of these studies relied on the current ERV/LTR classification. More recently, phylo-regulatory approaches, combining phylogenetic reconstruction of LTR subfamilies and layering of epigenetic data, have helped overcoming some of these issues for specific LTR subfamilies, e.g., LTR7 (Carter et al. 2022).

Considering the nature of ERVs/LTRs, we hypothesized that the proper classification and annotation of LTRs would require investigating their phylogenetic relationships in order to infer their evolution and function. In this work, we present an improved primate-specific ERV/LTR annotation in the human genome using a phylogenetic approach that combines sequence analysis and the presence of orthologous instances in other species.

RESULTS

The LTR subfamilies spreading in the primate lineage are heterogenous

We wanted to investigate the evolution of primate-specific ERVs/LTRs and selected 179 LTR subfamilies present in the human genome with copies in the marmoset or more closely related primate species, but absent from lemur (hence putatively integrated since the simian ancestor ~40 million years ago) (**Figure 1A**). We focused on LTRs since they usually contain most of the sequences driving ERVs regulatory and transcriptional activity. Among them, we further selected 35 subfamilies with limited shared repeat instances (≤ 60%) in the macaque genome (**Figure S1A**). Based on a network analysis of repeat consensus sequence similarity (Atkinson et al. 2009), we also identified 26 closely related LTR

subfamilies for a total of 61 putatively simian-specific subfamilies organized in 19 groups (**Figure S2**). For instance, MER9a1, MER9a2, MER9a3, and MER9B were clustered together and were distinct from other subfamilies. Next, by examining the distribution of divergence rates of instances within the 61 LTR subfamilies (**Figure 1B** and **Figure S3A**), we found that 28 (46%) had a non-normal distribution (chi-square *p* value ≤ 0.001), including LTR12B, LTR5B, LTR7Y, LTR61, LTR5_Hs and others showing a bimodal distribution. This is consistent with a recent report characterizing multiple subgroups within the LTR7 subfamilies (Carter et al. 2022), and suggests that many simian-specific LTR subfamilies are heterogeneous.

From this list, the MER11 subfamilies are of particular interest, because they were amongst the youngest and displayed non-normal divergence rate distributions (**Figure 1B** and **Figure S3B**). To further explore the variability in this group, we built an unrooted tree separately for MER11A, MER11B and MER11C based on the multiple sequence alignment of all the repeat instances. Another MER11 subfamily, MER11D, was also analyzed to confirm its distal relationship. Following a method described previously (Carter et al. 2022), we grouped instances into 66 clusters based on internal branch length: 18 for MER11A (labeled 11A_c1-18), 18 for MER11B (labeled 11B_c1-18), 27 for MER11C (labeled 11C_c1-27) and 3 clusters for MER11D (labeled 11D_c1-3) (**Figure 1C** and **Figure S3C**). To understand the relationship among the 66 MER11 clusters, we performed a median-joining network analysis using the cluster consensus sequences - a method used to infer intraspecific phylogenies (Bandelt, Forster, and Röhl 1999) (**Figure 1D**). As expected, MER11D clusters were grouped as an independent branch, and we found that most MER11A and MER11C clusters were grouped together. Notably, many MER11B clusters were dispersed between MER11A and MER11C clusters except 11B_c3/c5/c9/c18 that were found to be on a separate branch.

To further understand the evolution among these MER11A/B/C subfamilies, we inferred rooted trees using the non-reversible model without the use of an outgroup (Naser-Khdour, Quang Minh, and Lanfear 2022) (see Methods). After lifting over to other primate genomes the instances from each cluster, we selected the root that was most consistent with the proportions of shared instances across species (**Figure S3D**). This rooted tree was also found to be consistent with the network we constructed, further confirming the expansion progress of these repeat instances (**Figure S3E**). Based on the rooted tree and internal branch lengths between cluster consensus sequences, we observed four groups, which we defined as “phyletic groups” (see Methods, **Figure 1E**). MER11_G1, MER11_G2, and MER11_G3 were paraphyletic groups and MER11_G4 was a monophyletic group. As expected, instances in these phyletic groups displayed more homogenous divergence rates compared to the original subfamilies (**Figures S3B** and **S3F**). Notably, some clusters of instances from the evolutionary old MER11A were put in MER11_G1; while others were grouped with clusters from MER11B/C to form MER11_G2. Moreover, half of MER11B clusters and two MER11A and three MER11C clusters were grouped into MER11_G3, and

most MER11C clusters and another half of MER11B clusters were grouped into MER11_G4. We also found that clusters with a relative higher or lower liftOver rate to macaques compared to other clusters (e.g., 11B_c17, 11C_c27, and 11B_c1) were more likely to be reassigned to different phyletic groups (**Figure 1F**). Notably, if we were to select the top phyletic group to represent each subfamily, we found that a total of 412 (19.8%) MER11 instances would be annotated differently (**Figure 1G** and **Table S1**). Taken together, detailed sequence analysis revealed four phyletic groups with many MER11 instances that contrasted with the current subfamily classification.

MER11 phyletic groups display more consistent epigenetic profiles as compared to MER11 subfamilies

Given that the reconstituted MER11 phyletic groups were more homogeneous based on divergence rates (**Figure S3F**), we hypothesized that they would also have more consistent epigenetic profiles in human cells. As endogenous retrovirus expression and endogenization often occurs in early developmental stages (Hermant and Torres-Padilla 2021; Grow et al. 2015), we first compared chromatin accessibility and H3K27ac active histone mark in human embryonic stem cells (hESCs) and in hESC-derived neural progenitor cells (NPCs) using two published datasets (Inoue et al., 2019; Xie et al. 2013). From this, we identified 18 LTR subfamilies that were significantly enriched in hESCs (**Figures S4A-S4C**). In particular, we found that MER11B and MER11C subfamilies showed relatively high cell-type specificity in hESCs and mesendoderm cells (**Figure S4D**). Next, we performed an integrative analysis combining the phylogenetic trees built from MER11A/B/C clusters and their epigenetic profiles in hESCs using a panel of 27 histone marks and 65 transcription factor (TF) ChIP-seq datasets from the ENCODE project (**Figure 2A**). Notably, we observed that specific clusters within the subfamily trees were enriched for specific histone marks and TF peaks. For example, the younger MER11A clusters, with a long evolutionary distance to the root, were enriched for active histone marks and TF peaks. In contrast, the more ancient MER11B and MER11C clusters were enriched for active histone marks and TF peaks. The youngest MER11C clusters were enriched for active histone marks but less enriched for TF peaks. Thus, each subfamily had a high epigenetic heterogeneity.

To inspect whether our four reconstituted MER11 phyletic groups displayed a more consistent epigenetic profile compared to the original subfamily annotations, we rearranged the epigenetic data, using the rooted tree defined in the previous section, and observed distinct patterns of epigenetic states between groups (**Figure 2B**). For instance, MER11_G1 lacked TF peaks and active histone marks; MER11_G3 was significantly enriched for open chromatin and most TF peaks, followed by MER11_G2; Among MER11_G4 clusters, only the youngest ones were enriched for active histone marks. Because of the balance between chromatin accessibility and repression over TEs (Bourque et al. 2018), we also looked at the enrichment of KRAB-ZNFs and KAP1 (**Figures S4E-S4F**) binding in HEK293T cells. We further observed the sequential loss of ZNF440, ZNF433, ZNF468, ZNF611, ZNF33A, ZNF808 binding

and the gain of ZNF525 binding along the evolution of the MER11 clusters (**Figure 2B**). KAP1 binding was mostly enriched in MER11_G3 and relatively old MER11_G4 clusters, which was consistent with the enrichment of ZNF808 binding. Finally, compared to the original annotation of the MER11 subfamilies, we found that the four phyletic groups achieved a higher specificity across these active marks (**Figure 2C** and **Figure S4G**). For example, TEAD4 enrichment was 29.6% using phyletic groups (it overlapped 29.6% of instances in MER11_G3 but 0% of instances in MER11_G1) which is much higher than using subfamilies (the enrichment was 8% since it overlapped 11% of MER11B instances but 3% instances in MER11A). Taken together, the four reconstituted phyletic groups appear to resolve the epigenetic heterogeneity within MER11 instances, and we found that relative age was associated with distinct regulatory profiles.

MPRA confirms the regulatory potential of MER11 phyletic groups and reveals associated TF binding motifs

To further assess the biological relevance of the reconstituted MER11 phyletic groups, we leveraged a massively parallel reporter assay (MPRA). First, we identified two peaks of accessible regions within the MER11A/B/C instances, and extracted their sequences (~250-bp) as the frames - putative functional sequences of a suitable length - to be analyzed by MPRA (**Figure 3A** and **Figures S5A-S5B**). As controls, we analyzed two older LTR subfamilies, MER34 and MER52 (**Figure 3B**); MER34A1/C_ subfamilies were enriched for both ATAC-seq and H3K27ac peaks in hESCs compared to NPCs, while the enrichment of MER52C subfamily was slightly increased during the NPC differentiation (**Figures S4B-S4C**). We then retrieved homologous sequences in the human, chimpanzee and macaque genomes to characterize the regulatory potential of a large fraction of the observed evolutionary variants (**Figure S5C**). Some sequences had to be excluded due to the high number of mutations and truncations relative to the core frames used for the MPRA experiment (Methods). In the end, we analyzed 16,929 unique LTR sequences, including 6,912 MER11s, 5,751 MER34s, and 4,266 MER52s sequences, together with 100 positive and 100 negative control sequences (**Figure 3C** and **Tables S2-S3**), in both human iPSCs and iPSC-derived NPCs and in triplicates. We analyzed only high-quality sequences that were observed with at least 10 barcodes associated in the library (> 80% of MER11/52 and 30-75% of MER34, **Figure S5D**). Lower quality observed for MER34 was probably due to low GC content and long insert length (data not shown). The RNA/DNA ratios between replicates were strongly correlated (R^2 = 0.83), including for the positive and negative controls, indicating the accuracy of the MPRA measurements (**Figure S5E**).

After normalization, we found that MER11 frame 2 sequences showed higher MPRA activities compared to frame 1 sequences, which was consistent with the chromatin accessibility data (**Figure 3D** and **Figure S5F** and **Table S3**). Specifically, half of the MER11 frame 2 sequences were highly active (z-scaled activity ≥ 2) in human iPSCs, while the proportions stayed around 10% across MER11 frame 1 sequences. We next examined the MPRA activity among the human MER11 clusters and phyletic groups.

Notably, even though the overall activity of frame 1 remained low, the activity levels were increased for the more recent MER11_G4 clusters (**Figure S5G**). For frame 2, the activity varied between phyletic groups, specifically some clusters from MER11_G2/3 and young clusters from MER11_G4 had the highest activity levels (**Figure 3E**, **Table S1** and **Table S3**). Most instances from the oldest phyletic group, MER11_G1, could not be analyzed by MPRA due to mutations and truncations; however, among the remaining sequences, the reported activity levels were low.

One of the reasons for generating MPRA data was to implement a TE-wide motif association analysis approach to identify TF binding motifs contributing to the activity detected (Kheradpour et al. 2013). Among all frame 1 genomic sequences, we identified SP3 and related motifs (e.g., KLF12) followed by ZICs, POU::SOXs and SOXs motifs (**Figure S6**). We then looked at the MER11 frame 2 sequences and identified many motifs that were significantly associated with the MPRA activity including SOXs, POU2F1::SOX2, PITXs, ZICs, and TEADs. The motifs found to be enriched in MER34 (e.g., SPs and POU::SOXs) and MER52 (e.g., SPs, KLFs) were quite different. As expected, many TF motifs were strongly correlated with each other, such as RFX6, DMBX1, FOXF2, NR5A1, INSM1, and RARA::RXRG in MER11 frame 2 sequences (**Figure S7**). For each motif group, we only kept the most strongly associated motif ($p \leq 1 \times 10^{-10}$). Next, we looked at the proportion of MER11 frame 2 sequences containing each motif (**Figure 3F**). Notably, we observed a clear clustering of most motifs amongst the different phyletic groups. For instance, we observed the unique enrichment of NR5A1, FOXF2, and RFX6 related motifs in MER11_G3; ZNF136, ZIC3, and TEAD4 in both MER11_G3/G4; POU2F1::SOX2 and SOX17 in MER11_G4.

Finally, we investigated whether these TF motifs overlapped with nucleotides associated with the MPRA activity based on TE-wide nucleotide association analysis (Du et al. 2022). Focusing on MER11 frame 2, we identified 15 single nucleotide variants and 32 indels that were significantly associated with the MPRA activity ($p \leq 1 \times 10^{-10}$) (**Figures S8A-S8B**). For instance, we observed an overlap between strongly associated motifs, e.g., ZIC3, ZNF136, RFX6, and CRX, and nucleotides in the human frame 2 multiple sequence alignment associated with MPRA activity (**Figure S8C**). We also observed an enrichment of GATA5 and ZNF317 in both MER11_G1/4 which were due to apparent motif turnovers at different locations. When we compared the specificity of enriched motifs, we consistently observed a higher motif enrichment amongst MER11 phyletic groups relative to the originally assigned subfamilies (**Figure 3G** and **Figure S5H**). For example, the highest proportion of frame 2 sequences containing TEAD1 was 66.5% amongst phyletic groups and 40.5% amongst subfamilies, while the lowest proportion was 2.4% amongst phyletic groups and 15.3% amongst subfamilies. Taken together, we concluded that the four reconstituted MER11 phyletic groups were distinguishable based on distinct sets of motifs associated with MPRA activity.

The human MER11 phyletic groups are conserved in the primate lineage but spreading independently

To further characterize the evolution of MER11 phyletic groups, we examined the conservation of instances across the human, chimpanzee, and macaque genomes. We found that MER11 subfamilies had expanded in a lineage-specific fashion since the human-macaque ancestor with, for example, more than 80% of the macaque MER11A sequences absent from the human genome (**Figure 4A**). In contrast, only 7.4% of the chimpanzee MER11A instances are absent in the human genome (**Figure S9A**). Next, using the approach described above, we built the unrooted tree amongst MER11A macaque instances and identified 33 clusters (11A_m_c1 to c33, **Figure 4B** and **Figure S9B**). Some clusters (e.g., 11A_m_c2/c16/c17/c18) containing > 40% of instances shared with humans were clustered together, while other clusters with a low proportion of instances shared with humans (e.g., 11A_m_c1/c3/c10) were clustered with MER11B/C consensus sequences as labeled in Repbase, suggesting that these instances may be mis-annotated.

Next, we inferred the rooted tree of the macaque MER11A clusters (**Figure 4C**, left) and validated that evolutionary old clusters had amongst the highest proportions of instances shared with humans. We found that the MER11 rooted trees had a high consistency between two lineages (**Figure 4C**, right). Notably, based on the sequence similarity patterns and the rooted trees, we could also mostly recapture the four phyletic groups we had defined based on the human instances. We further compared the features of each phyletic group between human and macaque lineages. We found that MER11_G1 and MER11_G2 in both species consistently had the highest proportions of instances shared with each other (**Figure 4D** and **Figure S9C**). Even though MER11_G4 was the least conserved, we also observed that most instances were in the oldest and youngest groups in both species (**Figure 4E**). Finally, we compared the MPRA activities of the frame 1/2 sequences obtained from the two species (**Figure 4F**). Overall, the activities of each phyletic group were comparable between two species, with macaque having relatively lower activities (4.4-15.0% for highly active MER11_G2/G3/G4). Taken together, we observed a high conservation in both lineages of the phyletic groups with respect to sequence and MPRA activity. This is especially notable in the youngest phyletic group (G4), in spite of an independent spread following the species divergence.

The gain of SOX-related motifs within a phyletic group recently occurred in humans and chimpanzees but not in macaques

The gain and loss of TF motifs might help explain the different expansions of MER11 in the primate lineages. To look for such motifs, we focused on MER11_G4 frame 2 sequences in humans, chimpanzees and macaques. Applying the approach described previously, we identified SOX15, POU2F1::SOX2, HSF1, ZBED2, ZNF317, and IKZF1 to be the most associated motifs across three species (**Figure 5A**). We then examined the association between the motif combination and MPRA

activities amongst MER11_G4 sequences (**Figure 5B**). We found that MER11_G4 frame sequences containing either or both POU2F1::SOX2 and SOX15 motifs had the highest activity levels compared to others.

Next, we inspected whether the gain of POU::SOX and SOX motifs only occurred in specific primate lineages. We performed the motif association analysis for each species separately. We observed a strong association between the POU2F1::SOX2 motif and MPRA activity consistently in humans, chimpanzees, and macaques; however, SOX related motifs were observed to be significantly associated in humans ($p \leq 1 \times 10^{-10}$) and chimpanzees ($p \leq 1 \times 10^{-10}$) but were missing in macaques (**Figure 5C**). We further analyzed the prevalence of the POU2F1::SOX2 and SOX15/17 motifs amongst the four phyletic groups frame 2 sequences across species. The proportion of sequences containing POU2F1::SOX2 remained below 10% in MER11_G1/2/3 and was significantly increased in MER11_G4, which was consistent across the three species. In contrast, the proportions of SOX15/17 were significantly enriched in humans and chimpanzees but remained low in macaque MER11_G4 (**Figure S10A**). Only a few other enriched motifs showed a divergence between species with most (e.g., INSM1, DMBX1, and CRX) being consistently enriched (**Figure S10B**). Nucleotide association further revealed that a human- and chimpanzee-specific single nucleotide deletion at position 158 in the frame alignment was significantly associated with the MPRA activity, which was located within SOX related motifs (**Figure 5D** and **Figure S10C**).

Finally, we inspected the POU::SOX and SOX related motifs across the re-constructed cluster consensus sequences at the target region (**Figure 5E**). Human and macaque cluster consensus sequences were compared. We observed a single nucleotide deletion at position 144 occurred between MER11_G3 and MER11_G4 leading to the gain of POU2F1::SOX2 related motifs; another 158-bp deletion occurred between 11B_c7 and 11C_c5 within MER11_G4 leading to the gain of SOX related motifs. Moreover, we compared the MPRA activities between MER11_G4 frame 2 sequences containing POU2F1::SOX2 and SOX15/17 motifs in these species (**Figure 5F**). As expected, we observed significantly higher activities in the sequences containing SOX15/SOX17 motifs relative to the sequences containing the POU2F1::SOX2 motif only. Taken together, the phylogenetic analysis of the MER11 family revealed a single nucleotide deletion leading to the gain of SOX-related motifs which was species-specific and significantly increased regulatory potential of the instances in this evolutionary young phyletic group.

Cryptic endogenous retroviruses subfamilies in the primate lineage with distinct epigenetic profiles

Having shown the usefulness of defining MER11 phyletic groups to understand the evolution of this family in primate lineages, we wanted to apply the same approach to examine other simian-specific LTR subfamilies (**Figure 1B and Figure S3A**). Specifically, we first built the unrooted trees and then selected

the best representative rooted tree for all analyzed 19 subfamily groups except for the one containing LTR12C and related subfamilies due to practical issues (Methods). In this way, we identified 75 phyletic groups from 18 subfamily groups (containing 53 LTR subfamilies) (**Figure 6A** and **Table S4**). Among them, the LTR7 subfamily group had the most (*N* = 12) phyletic groups and LTR66 only had one phyletic group. Moreover, 26 of the individual LTR subfamilies could be subdivided into multiple phyletic groups with a maximum of seven for LTR7C, supporting their high sequence heterogeneity (**Figure 6B**). For each LTR subfamily, we selected the phyletic group with the most instances to be the representative for that subfamily. With this, a total of 3,807 (30.0%) instances from these 26 LTR subfamilies were classified into a different phyletic group (**Figure S11A**). For instance, a total of 258 LTR5_Hs instances (42.7%) were now classified in a non-primary phyletic group.

To validate the approach, we reanalyzed the LTR7 subfamily which was carefully studied through phylo-regulatory analysis (Carter et al. 2022). Here, we also included other related subfamilies from the LTR7 subfamily group (i.e., LTR7B, LTR7A, LTR7Y) to depict their full evolutionary history (**Figure S11B**). As expected, we observed a high consistency between the 12 phyletic groups we identified and the sequence clusters previously reported by Carter et al. (2022) (**Figure S11C**). Moreover, by looking at the epigenetic profiles across different cell types, we found that the evolutionary young clusters from LTR7_G12, similar to LTR7up1/up2/up3 from (Carter et al. 2022), were more active and enriched for several TFBSs as compared to other phyletic groups in ESCs (**Figure S11C**). Additionally, LTR7_G4/G5 were found to be enriched for accessibility and H3K27ac peaks in trophoblast cells (TBL) compared to other cell types, while LTR7_G9 was enriched for accessibility for mesendoderm cells compared to others. We also observed the enrichment of ZNF808 binding sites in LTR7_G3/G11 and KAP1 binding sites in LTR7_G7/G8. These results highlight that the LTR7 phyletic groups have distinct cell-type specific epigenetic profiles.

Finally, we explored the epigenetic properties of the 75 newly identified phyletic groups (**Figure 6C** and **Table S5**). As expected, we found that phyletic groups of many subfamily groups were significantly active during the differentiation from human ESC to NPCs, such as LTR5s, LTR6s, LTR7s, LTR13s, LTR22s, and MER11s. LTR phyletic groups were also active in a cell-specific manner, such as LTR6_G4 in mesendoderm cells (ME). We further looked at the epigenetic states in ESCs and found that they were also enriched for different sets of histone marks and TFBSs between phyletic groups. For instance, LTR13_G9/G12 were enriched for RAD21, CTCF, and ZNF143 except for LTR13_G1. We also observed a distinct enrichment of KRAB-ZNF and KAP1 binding sites in HEK293T cells between phyletic groups within certain subfamily groups, such as MER9s. Focusing on the six LTR5 phyletic groups revealed that LTR5_G5 had amongst the highest epigenetic enrichment (**Figure S11D**). LTR5_G1/G2/G3, which were evolutionarily older, were also overrepresented for multiple active histone marks. As expected, we also observed a higher TF specificity amongst phyletic groups relative to the

subfamilies, such as LTR6 and LTR22 subfamily groups (**Figure 6D**). Taken together, the phyletic groups we characterized in simian-specific LTR subfamilies were enriched for different sets of active histone marks, TFs, and KRAB-ZNFs, suggesting differential evolutionary trajectories.

DISCUSSION

The MER11 family of endogenous retroviruses was previously analyzed for its phylogeny and TF binding sites and motifs (Ito et al. 2017; Scognamiglio et al. 2023; Imbeault, Helleboid, and Trono 2017; De Franco et al. 2023). However, these studies lacked the assays demonstrating the regulatory activity of distinct genomic variants, nor elucidate the evolutionary process of their regulatory function. Here, based on a critical assessment of the available annotations of MER11 instances, we demonstrate the importance of analyzing the regulatory activity of thousands of genomic instances (copies) using properly reconstituted phyletic groups. We found such an annotation critical to understand the expansion process of MER11 across diverse simian genomes. Moreover, such a phylo-regulatory approach allowed us to detect stronger and new associations between phyletic groups of MER11 and their epigenetic profiles. Following these observations, an MPRA helped us measure the regulatory activity of MER11 variants at an unprecedented scale. MPRAs were previously used to analyze the regulatory activity and evolution of other LTR families, such as LTR18A (Du et al. 2022), however none of these studies focused on refining the annotation of the instances themselves.

Among the four MER11 phyletic groups, we found that the intermediate-aged groups (MER11_G2/G3) but not MER11_G1 (oldest) contained multiple TF motifs, such as ZIC and TEAD (**Figure S8C**). These TF motifs were conserved between human and macaque (**Figure S10B**), suggesting a functional role during the early expansion of MER11, and before the divergence between these two species. It remains unclear whether these motifs were originally present in the ancestral MER11 sequence but subsequently lost in G1, or if they were gained in G2/G3. In contrast, MER11_G4 did not have the ZIC and TEAD motifs, but gained the POU::SOX motif following a single nucleotide deletion. This change might have played a role in the expansion of MER11_G4 in the simian ancestral genome. Notably, in human MER11_G4, another deletion occurred within the same motif region and was detected to be a sole SOX motif in our analysis (**Figure 5E**). This suggests an increase in binding affinity for the SOX proteins, supported by the observed increase in the MPRA activity (**Figure 5F**). This deletion emerged after the divergence between humans and macaques, potentially contributing to the human-specific expansion of this younger group of MER11_G4. These results demonstrate that using the reconstituted MER11 phyletic groups combined with MPRA can shed light into the process of LTR expansion and divergence at the single nucleotide level. The enhancer function of the MER11 enhancer in relation to the host genome awaits validation in future studies.

The enhancer function of LTRs is influenced not only by transcription activators but also by the de-repression of KRAB zinc-finger transcription repressors that play a role in the defense mechanism of the host genome (Friedli and Trono 2015; Rosspopoff and Trono 2023). In our study, we observed that various KRAB repressor binding sequences were enriched in different LTR phyletic groups (**Figure 2B**). We showed that proper classification based on phyletic groups could also be useful to understand the arms race between transcriptional activation and repressive mechanisms. Previous research reported ZNF808 binding to MER11 during pancreatic development (De Franco et al. 2023). It remains uncertain which ZNF protein(s) potentially bind to MER11 in iPSCs, although we found a motif for ZNF136, which is expressed in iPSCs, in the frame 2 region of MER11_G2/G3. However, it is important to note that in our MPRA experiment, we primarily focused on the core region of the chromatin accessibility (around 250 bp), which may have limited our ability to fully detect ZNF motifs or mutations that negatively correlate with MPRA activity and contribute to the de-repression. To understand the molecular mechanism of the arms race between transcriptional activation and repressive mechanisms (e.g. mutations in ZNF motifs), a broader range of ERV/LTR sequences should be explored.

Our approach with MER11 highlights the significance of the phylogenetic classification and annotation of LTRs to understand their functional evolution. Unfortunately, the current annotation of LTR subfamilies groups together heterogeneous sets of sequences (**Figure 1B**). We also observed that macaque MER11B/C/D instances were all mis-annotated as MER11A (Figures S5A-S5B), which suggests that non-human species may be even more problematic. This issue is likely particularly prevalent in primate LTRs due to their high degree of sequence similarity. To highlight this phenomenon beyond MER11 subfamilies, we used a similar approach across the 53 simian-specific LTR subfamilies and observed that nearly half (26 out of 53) encompasses distinct phyletic groups. This analysis suggests a new annotation for 30% of the genomic instances from these 26 simian-specific LTR subfamilies (**Table S1**). As for MER11, these refined annotations could reveal signals that were missed previously and foster new discoveries relative to the contribution of TEs to primate genome evolution.

Finally, the classification and annotation of TEs across species has been a challenging problem because of the lack of ground truth (Hoen et al. 2015). Going forward, we argue that using repeat instances epigenetic and functional profiles, as we have done here, could be an effective strategy to evaluate alternative methods for TE classification and annotation.

METHODS

LTR phylogenetic analysis

LTR orthology and sequence divergence analysis: To study the expansion of LTR subfamilies in primate lineages, we lifted over the human LTR instances to representative primate species and the

mouse. The human (hg19) TE annotation file "hg19.fa.out" was obtained from the UCSC database (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/), which also contains the divergence rates (% substitutions) of instances relative to each subfamily consensus sequence. After renaming few subfamilies, we then converted the downloaded ".out" file to BED format using makeTEgtf.pl (https://github.com/mhammell-laboratory/TEtranscripts/issues/83) script. Chain files from human (hg19) to chimpanzee (panTro6), gorilla (gorGor3), orangutan (ponAbe2), gibbon (nomLeu3), macaque (macFas5), baboon (papAnu2), marmoset (calJac3), lemur (micMur1), and mouse (mm10) were downloaded from the UCSC database (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/liftOver/). We also lifted over hg19 to hg38 as a control. BnMapper (https://github.com/bxlab/bx-python) with the parameters "*-k -t 0.5*" was used for the liftOver analysis. Simian-specific LTR subfamilies were detected as the LTR subfamilies with a minimum of 100 instances (≥ 200 bp) and a maximum of 20% instances that were shared with the lemur genome. Evolutionary ages were computed based on the divergence rates as we previously described (Chen et al. 2023). Briefly, the divergence rates were first divided by the substitution rate for the human genome (2.2×10^{-9}) and then averaged across instances from each subfamily as the evolutionary age. To identify subfamilies with potential distinct expansion in primate lineages, we further kept the subfamilies with a maximum of 60% instances that were also present in the macaque genome.

LTR consensus sequence similarity and network analysis: We first retrieved the TE consensus sequences in FASTA format from the Repbase database (J. Jurka et al. 2005), which was used to annotate the human (hg19) and macaque (macFas5) genomes used here. We then calculated the sequence similarity score by comparing the sequences amongst themselves using Blastn (BLAST 2.13.0+) (Camacho et al. 2009) with the parameters of "*-task dc-megablast -outfmt 6 -num_threads 4*" with the default cut-off (E-value < 10).

After that, the resulting bit scores between each pair of sequences were used as the input of Cytoscape (v3.10.0) for the network analysis (Shannon et al. 2003; Atkinson et al. 2009). Subfamilies with similar consensus sequences were categorized into a subfamily group for the following analysis. Specifically, we first used the edge score of 200 to identify confident subfamily groups containing each candidate simian-specific LTR subfamily. We then used all edges to recover closely related subfamilies for groups containing a single candidate LTR subfamily. SVA subfamilies are homologous to both LTR5_Hs and Alu sequences in different regions, thus we only kept LTR5 subfamilies within this subfamily group for the following analysis.

Human MER11 unrooted trees reconstruction: We first extracted the coordinates of instances (≥ 200 bp) from each MER11 subfamily from the TE annotation BED file separately. We then used Bedtools2 *getfasta* function to extract sequences from the human reference genome (hg19) with the parameter "*-nameOnly -s*". To reconstruct the evolutionary tree, we first performed the multiple sequence alignment of

instances from each subfamily using MAFFT (v7.5.05) (Nakamura et al. 2018) with the parameters "*--localpair --maxiterate 1000*". The sequence alignment was further refined using PRANK (v170427) (Löytynoja 2014) with the parameters "*-showanc -njtree -uselogs -prunetree -F -showevents*". We then used trimAL (v1.4.1) (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009) to remove gaps that were present in less than 10% of sequences. Lastly, the evolutionary tree was obtained by IQ-TREE 2 (v2.1.2) (Minh et al. 2020) with the parameters "*-nt AUTO -m MFP -bb 6000 -asr -minsup .95 -T 4*".

Human MER11 clusters detection and consensus sequences median-joining network analysis: To determine MER11 clusters per subfamily, we identified branches supported by > 95% ultrafast bootstrap and a minimum internal branch length of 0.02 to any other instances and contained ≥ 10 instances. The original MER11A/B/C/D consensus sequences were also included as references and instances from the identified branches containing < 10 instances were excluded in the following analyses. We then used the *consensusString* function from Biostrings (v2.64.1) R package (https://bioconductor.org/packages/Biostrings) to get the consensus sequences with the majority rule (> 0.51). The cluster consensus sequences were then submitted to PopART (v1.7) (Leigh and Bryant 2015) for the median-joining network analysis. MER11D clusters were also included to confirm their distal relationships with other MER11A/B/C clusters.

Human MER11 cluster consensus sequences divergence rate analysis: We used MAFFT with the parameters "*--globalpair –maxiterate 1000*" to align MER11 cluster consensus sequences with the default parameters. We then used RAxML (v8.2.12) with the parameters "*raxmlHPC-PTHREADS-AVX -f x -p 12345 -m GTRGAMMA*" to compute the divergence rates (maximum likelihood distances) between each pair of human cluster consensus sequences. We also re-calculated the divergence rates of every instance versus their corresponding subfamily and phyletic group consensus sequences separately. We used the consensus sequence of relatively ancient cluster (i.e., 11A_c7, 11A_c8, 11B_c11, and 11C_c2) to represent each phyletic group. We first ran RepeatMasker with all MER11A/B/C instances and each consensus sequence as the inputs and parameters "-e rmblast -pa 4 -s -no_is". We then used the "one_code_to_find_them_all_but_sanely.pl" (https://github.com/mptrsen/mobilome/blob/master/code/Onecodetofindthemall/) script to combine adjacent partial hits. We used ggplot2 for the visualization.

Human MER11 phyletic groups determination: We next want to infer the best rooted tree among all MER11A/B/C cluster consensus sequences. Firstly, we performed the multiple sequence alignment analysis using MAFFT with the parameters "*--localpair --maxiterate 1000*". Secondly, the alignment was refined using PRANK with the parameters "*-showanc -njtree -uselogs -prunetree -F -showevents*". After that, we constructed the rooted trees using IQ-TREE 2 with the parameters "*--model-joint 12.12 -B 1000 -T AUTO --root-test -zb 1000 -au*". It also performed the statistical tests for rooting positions on every

branch. We then selected the best tree based on the ranking, different statistical tests, and the liftOver rates to other primate species. Specifically, we selected from the top-ranked rooted tree, which rooted cluster has among the highest liftOver rates in the most ancient primate species we used above. We also prioritized the rooted trees having the highest values amongst different statistical tests. The top-selected rooted tree was then used in the following analyses. The branch length (bootstrap value) from each cluster to the root referred to the evolutionary age.

We next determined the phyletic groups based on the internal branch lengths of the top-selected rooted tree. Since the branch lengths varied between subfamily groups, we grouped clusters as a phyletic group which have amongst the top branch lengths to others. We also confirmed the phyletic groups by examining the pair-wide divergence rates to look at extreme values between every adjacent clusters. We lastly kept the phyletic groups containing a minimum of 25 instances.

Macaque MER11A phylogenetic analysis: We performed the same phylogenetic analysis for macaque (macFas5) MER11A instances (≥ 200 bp). After we subdivided them into clusters, we also inferred the cluster consensus sequences rooted tree using the same approach. We then determined phyletic groups for the macaque MER11A subfamily while two phyletic groups were named “G4-1” and “G4-2” according to their close relationship with human “MER11_G4” consensus sequences. We also lifted over the instances from each cluster to human (hg19) using bnMapper with the parameters “*-k -t 0.5*”. The chain file “macFas5ToHg19” was downloaded from the UCSC database (https://hgdownload.soe.ucsc.edu/gbdb/macFas5/liftOver/). The divergence rates between human and macaque MER11 cluster consensus sequences were also computed using the same approach above. We used ggplot2 for the visualization.

***Simian-specific LTR subfamily groups phylogenetic analysis*:** We performed a similar analysis for other simian-specific LTR subfamily groups. Briefly, we first constructed the unrooted trees of instances (≥ 200 bps) for every subfamily from a group. After identifying the clusters per subfamily, we constructed the rooted trees and selected the best one to determine phyletic groups for each subfamily group. Using the same approach, we kept phyletic groups with a minimum of 25 instances except LTR61 subfamily which has a small number of instances.

LTR epigenetic analysis

Chromatin accessibility permutation analysis at the TE subfamily level: We downloaded the ATAC-seq peaks from the NCBI Gene Expression Omnibus (GEO) database (GSE115046) (Inoue et al. 2019). After that, we used the same approach as we previously described to evaluate the enrichment level relative to the random genomic background per TE subfamily (Chen et al. 2023). Briefly, we first shuffled the peak regions 1000 times relative to the distribution of peaks. We then computed the number of

instances per subfamily that overlapped with the actual and shuffled peak summits separately using Bedtools2 (Quinlan and Hall 2010) *intersect* function with the parameters "*intersect -wa -u -a*". After that, we counted the number of instances that were associated with peaks per phyletic group. Fold enrichment was computed as the number of actual peaks-associated instances divided by the average number of shuffled peaks-associated instances, and the permutation test was used to compute the *p* values. We also performed the same analyses on another H1 ESC (E003) DNase-seq dataset (Xie et al. 2013).

Differential accessibility and H3K27ac activity between ESCs and NPCs: To identify LTR subfamilies with the accessibility and H3K27ac activity changes during the cell differentiation, we re-analyzed the accessibility and H3K27ac activity changes in TEs between ESCs and NPCs. The additional ATAC-seq and H3K27ac peak files were downloaded from the same sources. Fold change per subfamily between ESCs and NPCs was computed as we described previously (Chen et al. 2023). We then kept subfamilies with a minimum of two-fold enrichment of chromatin accessibility in ESCs compared to NPCs. We further validated the results in other differentiated cells including mesendoderm cells (ME, E004), trophoblast-like cells (TBL, E005), mesenchymal stem cells (MSC, E006), and neural progenitor cells (NPC, E007) (Xie et al. 2013). We kept the LTR subfamilies that were significantly enriched in the accessibility and H3K27ac in ESCs from two independent sources.

Permutation test of other epigenetic marks overlapped each MER11 subfamily: Except the above datasets, we further obtained additional H1 ESC histone and TF Chip-seq peaks (https://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/), and HEK293T KRAB-ZNF and KAP1 Chip-seq peaks (Imbeault, Helleboid, and Trono 2017). We then examined the enrichment of each epigenetic mark overlapped with MER11 subfamilies using the same approach. The number of peaks overlapped with each subfamily was normalized by the total number of peaks per mark. We then kept epigenetic marks overlapped with a minimum of 20 instances per subfamily that were significantly enriched (fold enrichment ≥ 2 and *p* value ≤ 0.05).

Hypergeometric test of histone marks and TFBSs at the cluster level: We selected the significantly enriched epigenetic marks in any MER11 subfamilies as well as other well-known active (H3K4me1/2) and repressive marks (H3K27me3). We then inspected whether these marks were overrepresented within specific MER11 clusters. To do it, we intersected the actual peaks with MER11 clusters using Bedtools2 *intersect* function with parameters "*-wa -e -f 0.5 -F 0.5 -u*". After that, we computed the proportion of instances per cluster were overlapped with each peak. Moreover, we used the hypergeometric test (R *phyper* function) to compute the *p* value for the enrichment of peaks-associated instances per cluster relative to each subfamily. The *p* values were adjusted using R *p.adjust* function with the *Benjamini & Hochberg* method.

Permutation test of epigenetic marks overlapped with phyletic groups of each subfamily group: We examined the enrichment of the above epigenetic marks overlapped with each determined MER11 phyletic group. Specifically, we first shuffled the peak regions 100 times relative to the distribution of peaks using our previous approach (Chen et al. 2023). We then intersected the actual and shuffled peaks with instances from each phyletic group using Bedtools2 *intersect* function with parameters "*-wa -e -f 0.5 -F 0.5 -u*". After that, we counted the number of peaks-associated instances per phyletic group. Fold enrichment was computed as the number of actual peaks-associated instances divided by the average number of shuffled peaks-associated instances. A permutation test was used to compute the *p* values. Phyletic groups with a minimum of two-fold enrichment (log2((actual counts+1)/(mean shuffled counts + 1)) ≥ 1) with *p* value ≤ 0.05 were kept. We also filtered out epigenetic marks overlapped with less than five instances from phyletic groups containing < 100 instances and less than 5% for phyletic groups containing ≥ 100 instances separately.

LTR lentiMPRA library design

Detection of sequence frames along the LTR consensus sequences: The computed reads per million (RPM) distribution of accessible MER11B instances were first aligned to each subfamily consensus sequence using the downloaded ".align" file as we previously described (Chen et al. 2023). We determined the consensus accessible regions based on the aggregated RPM distribution along the consensus sequences. After that, we extracted the consensus accessible sequences (frame regions) centered at the peak summits at around 250 bp long. We also retrieved the homologous MER11B and MER11C consensus sequences based on the multiple sequence alignment using ClustalW2 with the default parameters (https://www.ebi.ac.uk/Tools/msa/clustalw2/). Similarly, we determined the frame regions on MER34, and MER52 subfamilies separately.

Extraction of human, chimpanzee, and macaque LTR genomic sequences: We used the in-house Python script "*Organize_seqFile_to_consensus.py*" to retrieve annotated MER11, MER34, and MER52 sequences that were homologous to each frame sequence. We then kept homologous sequences with a maximum of 270 bp and without ambiguous nucleotides "N". Sequences that were shorter than 70% of the frame sequences were removed. Sequences with a reverse alignment against the frame sequences were converted to the reverse complementary sequences. Similarly, we obtained the chimpanzee and macaque genomic sequences homologous to each frame sequence. Chimpanzee and macaque TE annotation files (Repeat library 20140131) were downloaded from http://www.repeatmasker.org/species/macFas.html and http://www.repeatmasker.org/species/panTro.html.

Negative and positive controls: We first used the shuffleFasta tool with the parameter "*-n 100*" to randomly select 100 sequences from the extracted genomic sequences

(https://krishna.gs.washington.edu/content/members/vagar/forTaka/). We then used the Python script "*kMerFilter_fromMartin.py*" also from the same link with the parameters "*-k 8 --inclMinOverlap*" to further shuffle the nucleotides. The obtained sequences were used as the negative controls without activities. Sequences with activity (Inoue et al. 2019) were also used as positive controls in this study.

Consensus sequences: MER11/34/52 subfamily consensus sequences were also included in the library. We also reconstructed the consensus sequences amongst the retrieved genomic sequences per species based on the ".align" file. Consensus sequences with every nucleotide at ambiguous nucleotides "N" that were different from the above genomic sequences were kept.

Adding adapter sequence: After the removal of redundant sequences, we added the upstream and downstream adapter sequences "AGGACCGGATCAACT" and "CATTGCGTGAACCGA" to the examined sequences using an in-house Python script.

lentiMPRA experiment

lentiMPRA library cloning and sequence-barcode association: Designed sequence oligos were synthesized by Twist Bioscience. The lentiMPRA library construction was performed as previously described with modifications (Gordon et al. 2020). In brief, the synthesized oligo pool was amplified by 7-cycle PCR using forward primer (5BC-AG-f01, **Table S6**) and reverse primer (5BC-AG-r01, **Table S6**) that added mP and spacer sequences downstream of the sequence. The amplified fragments were purified with 1.8x AMPure XP (Beckman coulter), and proceeded to the second round 9-cycle PCR using forward primer (5BC-AG-f02) and reverse primer (5BC-AG-r02, **Table S6**) to add 15-nt random sequence that serves as a barcode. The amplified fragments were then inserted into SbfI/AgeI site of the pLS-SceI vector (Addgene, 137725) using NEBuilder HiFi DNA Assembly mix (NEB, E2621L), followed by transformation into 10beta competent cells (NEB, C3020) using the Gemini X2 machine (BTX). Colonies were allowed to grow overnight on Carbenicillin plates and midiprepped (Qiagen, 12945). We collected approximately 1.2 million colonies, so that on average 70 barcodes were associated with each sequence. To determine the sequences of the random barcodes and their association with each sequence, the sequence-mP-barcode region was amplified from the plasmid library using primers that contain flowcell adapters (P7-pLSmP-ass-gfp and P5-pLSmP-ass-i#, **Table S6**). The PCR fragment was then sequenced with a NextSeq mid-output 300-cycle kit using custom primers (pLSmP-ass-seq-R1, pLSmP-ass-seq-ind1 (index read), pLSmP-ass-seq-R2, **Table S6**).

Cell culture, lentiviral infection and barcode sequencing: WTC11 human iPSCs (Coriell Institute, RRID:CVCL_Y803) were cultured on matrigel (Corning, 354277) in mTeSR plus medium (Stemcell technologies, 100-0276) and passaged using ReLeSR (Stemcell technologies, 100-0484), according to manufacturer's instruction. WTC11 cells were used for the MPRA experiments at passage 43. Lentivirus

packaging was performed as previously described with modifications (Gordon et al. 2020). Briefly, 50,000 cells/cm2 293T cells (ATCC, CRL-3216) were seeded in four T175 flasks and cultured for 48 hours. The cells were co-transfected with 7.5 µg/flask of plasmid libraries, 2.5 µg/flask of pMD2.G (Addgene 12259) and 5 µg/flask of psPAX2 (Addgene 12260) using EndoFectin Lenti transfection reagent (GeneCopoeia, EF002) according to manufacturer's instruction. After 8 hours, cell culture media was refreshed and ViralBoost reagent (Alstem, VB100) was added. The transfected cells were cultured for 2 days and lentivirus was filtered through a 0.45um PES filter system (Thermo Scientific, 165-0045) and concentrated by Lenti-X concentrator (Takara Bio, 631232) according to manufacturer's protocol. We obtained in total 1.2 mL lentivirus solution from the four T175 flasks (100x concentration).

For lentiviral infection, approximately 4 million WTC11 cells per replicate were seeded in mTeSR plus medium supplemented with Y-27632 (Cayman, 10005583) in a 10 cm dish. After 24 hours, the cell culture medium was replaced by fresh mTeSR plus medium without Y-27632. To perform magnetofection, 100 µL/dish of the concentrated lentivirus library, 150 µL/dish ViroMag R/L reagent (OZ Biosciences, RL41000), and 750 µL/dish mTeSR plus medium were mixed and incubated at room temperature for 20 minutes. WTC11 cells in a 10 cm dish were added with the virus-ViroMag mixture and placed on a magnetic plate for 30 minutes. The cells were removed from the magnetic plate, and incubated for 24 hours. The infected cells were cultured for additional 3 days with a daily change of the mTeSR plus media, or induced into neural lineage for 3 days with dual-Smad inhibitors as described previously (Inoue et al. 2019). For each experiment, three independent infections were performed to obtain three biological replicates.

DNA/RNA extraction and barcode sequencing were all performed as previously described (Gordon et al. 2020). Briefly, genomic DNA and total RNA were purified from the infected cells using an AllPrep DNA/RNA mini kit (Qiagen, 80204). 120 µg RNA was treated with Turbo DNase (Thermo Fisher Scientific, AM1907) to remove contaminating DNA, and reverse-transcribed with SuperScript II (Invitrogen, 18064022) using a barcode-specific primer (P7-pLSmp-assUMI-gfp, **Table S6**), which has a 16-bp unique molecular identifier (UMI). The cDNA and 48 µg genomic DNA from each sample were used for 3-cycle PCR with specific primers (P7-pLSmp-assUMI-gfp and P5-pLSmP-5bc-i#, **Table S6**) to add sample index and UMI. A second round PCR (21 and 26 cycles for DNA and RNA barcode samples, respectively) was performed using P5 and P7 primers (P5, P7, **Table S6**). The PCR fragments were purified and sequenced with a NextSeq high-output 75-cycle kit (15-cycle paired-end reads, 16-cycle index read1 for UMI, and 10-cycle index read2 for sample index), using custom primers (pLSmP-ass-seq-ind1, pLSmP-bc-seq, pLSmP-UMI-seq, pLSmP-5bc-seqR2, **Table S6**).

MPRA activity measurement

Association analysis: To ensure the accurate association between barcodes and LTR inserts with variable lengths, we revised the Python script "*nf_ori_map_barcodes.py*" implemented in MPRAflow (v2.3.1) (Gordon et al. 2020). Specifically, we kept reads that were fully matched to the designed oligos nucleotide sequences with the cigar "M" at the extract insert lengths. We then ran the optimized MPRAflow pipeline with the following command and parameters "*nextflow run association.nf –mapq 5 –min-frac 0.5*" to associate each LTR insert sequence with multiple barcodes. Paired-end insert DNA FASTQ files, barcode FASTQ files, and designed library FASTA file were used as the inputs.

DNA/RNA count analysis: After that, we ran the MPRAflow command "*nextflow run count.nf –bc-length 15 –umi-length 15 -thresh 5 –merge_intersect FALSE*" to achieve the normalized number of DNA and RNA reads per barcode, and the RNA/DNA ratio associated with each LTR insert sequence. The designed library FASTA, the output file from the above association analysis, and the list of DNA/RNA barcode and UMI FASTQ files were used as the inputs.

MPRA activity (alpha value) calculation: We used the MPRAnalyze R package (v1.18.0) to compute the MPRA activity. The DNA and RNA count matrices of three replicates were used as the inputs. We first estimated the library size correction factors (i.e., batch and condition factors) using the *estimateDepthFactors* function with the parameters "*which.lib = "both", depth.estimator = "uq'"*". After the quant model fitting using the *analyzeQuantification* function, the MPRA activity (alpha value) was obtained using the *getAlpha* function with the parameters "*by.factor = "batch'"*". We then kept LTR insert sequences that were associated with ≥ 10 barcodes in more than two DNA libraries.

MPRA activity normalization: We normalized the activity (alpha value) by the negative controls following the same approach here (Inoue et al. 2019). Briefly, we first computed the mean absolute deviation of negative control sequences. After we extracted the high quality negative control sequences, we then computed the *z*-scaled MPRA activity using the same formula we previously reported. *P* values were computed based on the standard normal distribution and were further adjusted using the Benjamini–Hochberg method. LTR insert sequences with adjusted *p* value ≤ 0.05 were classified as sequences with MPRA activity.

Motif and nucleotide association analyses

De novo motif discovery and summary analysis: We searched each MER11 frame sequence for known motifs from the JASPAR 2022 database using MEME (v5.2.0) *fimo* function (Bailey et al. 2009). After the removal of redundant motifs per frame sequence, we computed the proportion of sequences containing each motif at the cluster level. Similarly, we computed the proportion for each MER11 phyletic group. Human, chimpanzee, and macaque sequences were analyzed separately.

Motif association analysis: We implemented a new TE motif association approach to identify motifs that contributed to the activity. The non-redundant motifs per frame sequence were converted to pseudo genotypes for each motif (i.e., presence denoted as A/B and absence denoted as A/A) in ".ped" format. Each row referred to a frame sequence and every two columns (starting from the 7th column) referred to the two pseudo alleles of a motif. The z-scaled MPRA activity was used as the quantitative phenotypes (6th column of the ".ped" file). We also prepared a corresponding ".map" file containing the list of motifs (same order as the ".ped" file). We lastly used Plink2 for the association analysis (https://www.cog-genomics.org/plink/2.0/) (Chang et al. 2015). Specifically, we filtered the genotypes with the parameters "*--maf 0.05 --make-bed --input-missing-phenotype 999*". After we computed the allele frequency, the generalized linear model was used for the association analysis with the parameter "*--glm allow-no-covars*". *P* values and BETA values were visualized by ggplot2. Analyses were done amongst the frame sequences from three species (human, chimpanzee, and macaque) or each species separately.

Nucleotide association analysis: We performed the multiple sequence alignment of MER11 frame 2 sequences using MAFFT with the parameters "*--localpair --maxiterate 1000*". We then converted the variants across each frame sequence into pseudo-genotypes (i.e., minor variant as A/B and major variant as A/A) for each nucleotide along the alignment. Gaps (or indels) and nucleotide changes were analyzed separately. We then used them as the inputs for the association analysis with the *z*-scaled MPRA activity in iPSCs as the quantitative phenotypes. Plink2 was used for the association analysis as we described above. *P* values and BETA values were visualized by ggplot2. Analyses were also done with the inclusion of MER11_G4 frame 2 sequences retrieved from three species or each species, separately.

ACKNOWLEDGMENTS

We would like to thank Drs. Gracie Gordon and Tal Ashuach for their help with the use of MPRAflow and MPRAnalyze tools. We also acknowledge Institute for the Advanced Study of Human Biology (ASHBi), Calcul Québec and the Digital Research Alliance of Canada for access to computing resources. We thank the Single-Cell Genome Information Analysis Core (SignAC) at WPI-ASHBi, Kyoto University, for their support. We thank Dr. Spyros Goulas for critical reading and suggestions on the manuscript.

Funding: This work was supported by the World Premier International Research Center Initiative (WPI), MEXT, Japan, the JSPS KAKENHI under Grant Numbers JP21K06119 (F.I.), JP21K15066 (X. C.), the Takeda Science Foundation, Bioscience Research Grants (F.I.), and by the Mitsubishi Foundation, Research Grants in the Natural Sciences (F.I.). This work was also supported by a Canadian Institute of Health Research (CIHR) program grant (CEE-151618). G.B. is supported by a Canada Research Chair Tier 1 award, an FRQ-S, and a Distinguished Research Scholar award. This research was enabled in part by support provided by Calcul Quebec and the Digital Research Alliance of Canada.

Author contributions: Conceptualization: X.C., G.B., and F.I. Methodology: X.C., Z.Z., G.B., and F.I. Formal Analysis: X.C. and Z.Z. Investigation: X.C., Z.Z., Y.Y., and F.I. Resources: G.B. and F.I. Visualization: X.C. Writing – original draft: X.C. G.B., and F.I. Writing – review & editing: X.C., Z.Z., C.G. G.B., and F.I. Funding Acquisition: X.C., G.B., and F.I. Overall supervision: G.B.

Competing Interests: F.I. receives funding from Relation Therapeutics.

Data and materials availability: The datasets generated in this study are available at the NCBI Gene Expression Omnibus (GEO) as accession number GEO: GSE245662. Scripts for main analyses are available at https://github.com/xunchen85/TE-MPRA-and-phylogenetic-analysis and will also be submitted to Zenodo at DOI:10.5281/zenodo.10016500 upon acceptance.

REFERENCES

Almeida, Miguel Vasconcelos, Grégoire Vernaz, Audrey L. K. Putman, and Eric A. Miska. 2022. "Taming Transposable Elements in Vertebrates: From Epigenetic Silencing to Domestication." Trends in Genetics 38 (6): 529–553. https://doi.org/10.1016/j.tig.2022.02.009.

Andrews, Gregory, Kaili Fan, Henry E. Pratt, Nishigandha Phalke, ZOONOMIA CONSORTIUM, Elinor K. Karlsson, Kerstin Lindblad-Toh, Steven Gazal, Jill E. Moore, and Zhiping Weng. 2023. "Mammalian Evolution of Human Cis-Regulatory Elements and Transcription Factor Binding Sites." Science 380 (6643): eabn7930. https://doi.org/10.1126/science.abn7930.

Arkhipova, Irina R. 2017. "Using Bioinformatic and Phylogenetic Approaches to Classify Transposable Elements and Understand Their Complex Evolutionary Histories." Mobile DNA 8 (1): 19. https://doi.org/10.1186/s13100-017-0103-2.

Atkinson, Holly J., John H. Morris, Thomas E. Ferrin, and Patricia C. Babbitt. 2009. "Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies." PLOS ONE 4 (2): e4345. https://doi.org/10.1371/journal.pone.0004345.

Bailey, Timothy L., Mikael Boden, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, and William S. Noble. 2009. "MEME Suite: Tools for Motif Discovery and Searching." Nucleic Acids Research 37 (suppl_2): W202–8. https://doi.org/10.1093/nar/gkp335.

Bandelt, H. J., P. Forster, and A. Röhl. 1999. "Median-Joining Networks for Inferring Intraspecific Phylogenies." Molecular Biology and Evolution 16 (1): 37–48. https://doi.org/10.1093/oxfordjournals.molbev.a026036.

Blomberg, Jonas, Farid Benachenhou, Vidar Blikstad, Göran Sperber, and Jens Mayer. 2009. "Classification and Nomenclature of Endogenous Retroviral Sequences (ERVs): Problems and Recommendations." Gene, Genomic Impact of Eukaryotic Transposable Elements, 448 (2): 115–23. https://doi.org/10.1016/j.gene.2009.06.007.

Bourque, Guillaume, Kathleen H. Burns, Mary Gehring, Vera Gorbunova, Andrei Seluanov, Molly Hammell, Michaël Imbeault, et al. 2018. "Ten Things You Should Know about Transposable Elements." Genome Biology 19 (1): 199. https://doi.org/10.1186/s13059-018-1577-z.

Bourque, Guillaume, Bernard Leong, Vinsensius B. Vega, Xi Chen, Ling Lee Yen, Kandhadayar G. Srinivasan, Joon Lin Chew, et al. 2008. "Evolution of the Mammalian Transcription Factor Binding Repertoire via Transposable Elements." Genome Research 18 (11): 1752–62. https://doi.org/10.1101/gr.080663.108.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. "BLAST+: Architecture and Applications." BMC Bioinformatics 10 (1): 421. https://doi.org/10.1186/1471-2105-10-421.

Carey, Kaitlin M., Gilia Patterson, and Travis J. Wheeler. 2021. "Transposable Element Subfamily Annotation Has a Reproducibility Problem." *Mobile DNA* 12 (1): 4. https://doi.org/10.1186/s13100-021-00232-4.

Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. "trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses." Bioinformatics 25 (15): 1972–73. https://doi.org/10.1093/bioinformatics/btp348.

Carter, Thomas A, Manvendra Singh, Gabrijela Dumbović, Jason D Chobirko, John L Rinn, and Cédric Feschotte. 2022. "Mosaic Cis-Regulatory Evolution Drives Transcriptional Partitioning of HERVH Endogenous Retrovirus in the Human Embryo." Edited by Mia T Levine and Detlef Weigel. eLife 11 (February): e76257. https://doi.org/10.7554/eLife.76257.

Chang, Christopher C, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. 2015. "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets." GigaScience 4 (1): s13742-015-0047–0048. https://doi.org/10.1186/s13742-015-0047-8.

Chelmicki, Tomasz, Emeline Roger, Aurélie Teissandier, Mathilde Dura, Lorraine Bonneville, Sofia Rucli, François Dossin, Camille Fouassier, Sonia Lameiras, and Deborah Bourc'his. 2021. "m6A RNA Methylation Regulates the Fate of Endogenous Retroviruses." Nature 591 (7849): 312–16. https://doi.org/10.1038/s41586-020-03135-1.

Chen, Xun, Alain Pacis, Katherine A. Aracena, Saideep Gona, Tony Kwan, Cristian Groza, Yen Lung Lin,

et al. 2023. “Transposable Elements Are Associated with the Variable Response to Influenza Infection.” Cell Genomics 3 (5): 100292. https://doi.org/10.1016/j.xgen.2023.100292.

Chuong, Edward B., Nels C. Elde, and Cédric Feschotte. 2016. “Regulatory Evolution of Innate Immunity through Co-Option of Endogenous Retroviruses.” Science 351 (6277): 1083–87. https://doi.org/10.1126/science.aad5497.

Chuong, Edward B., Nels C. Elde, and Cédric Feschotte. 2017. “Regulatory Activities of Transposable Elements: From Conflicts to Benefits.” Nature Reviews Genetics 18 (2): 71–86. https://doi.org/10.1038/nrg.2016.139.

De Franco, Elisa, Nick D. L. Owens, Hossam Montaser, Matthew N. Wakeling, Jonna Saarimäki-Vire, Athina Triantou, Hazem Ibrahim, et al. 2023. “Primate-Specific ZNF808 Is Essential for Pancreatic Development in Humans.” *Nature Genetics*, November, 1–7. https://doi.org/10.1038/s41588-023-01565-x.

Du, Alan Y., Xiaoyu Zhuo, Vasavi Sundaram, Nicholas O. Jensen, Hemangi G. Chaudhari, Nancy L. Saccone, Barak A. Cohen, and Ting Wang. 2022. “Functional Characterization of Enhancer Activity during a Long Terminal Repeat’s Evolution.” Genome Research 32 (10): 1840–1851. https://doi.org/10.1101/gr.276863.122.

Feschotte, Cédric, and Clément Gilbert. 2012. “Endogenous Viruses: Insights into Viral Evolution and Impact on Host Biology.” Nature Reviews Genetics 13 (4): 283–96. https://doi.org/10.1038/nrg3199.Friedli, Marc, and Didier Trono. 2015. “The Developmental Control of Transposable Elements and the Evolution of Higher Species.” Annual Review of Cell and Developmental Biology 31 (1): 429–51. https://doi.org/10.1146/annurev-cellbio-100814-125514.

Fuentes, Daniel R, Tomek Swigut, and Joanna Wysocka. 2018. “Systematic Perturbation of Retroviral LTRs Reveals Widespread Long-Range Effects on Human Gene Regulation.” Edited by Edith Heard and Detlef Weigel. eLife 7 (August): e35989. https://doi.org/10.7554/eLife.35989.

Fueyo, Raquel, Julius Judd, Cedric Feschotte, and Joanna Wysocka. 2022. “Roles of Transposable Elements in the Regulation of Mammalian Transcription.” Nature Reviews Molecular Cell Biology, February, 1–17. https://doi.org/10.1038/s41580-022-00457-y.

Gerdes, Patricia, Sandra R. Richardson, Dixie L. Mager, and Geoffrey J. Faulkner. 2016. “Transposable Elements in the Mammalian Embryo: Pioneers Surviving through Stealth and Service.” Genome Biology 17 (1): 100. https://doi.org/10.1186/s13059-016-0965-5.

Göke, Jonathan, Xinyi Lu, Yun-Shen Chan, Huck-Hui Ng, Lam-Ha Ly, Friedrich Sachs, and Iwona Szczerbinska. 2015. “Dynamic Transcription of Distinct Classes of Endogenous Retroviral Elements Marks Specific Populations of Early Human Embryonic Cells.” Cell Stem Cell 16 (2): 135–41. https://doi.org/10.1016/j.stem.2015.01.005.

Gordon, M. Grace, Fumitaka Inoue, Beth Martin, Max Schubach, Vikram Agarwal, Sean Whalen, Shiyun Feng, et al. 2020. “lentiMPRA and MPRAflow for High-Throughput Functional Characterization of Gene Regulatory Elements.” Nature Protocols 15 (8): 2387–2412. https://doi.org/10.1038/s41596-020-0333-5.

Grandi, Nicole, Maria Paola Pisano, Eleonora Pessiu, Sante Scognamiglio, and Enzo Tramontano. 2021. “HERV-K(HML7) Integrations in the Human Genome: Comprehensive Characterization and Comparative Analysis in Non-Human Primates.” Biology 10 (5): 439. https://doi.org/10.3390/biology10050439.

Grow, Edward J., Ryan A. Flynn, Shawn L. Chavez, Nicholas L. Bayless, Mark Wossidlo, Daniel J. Wesche, Lance Martin, et al. 2015. “Intrinsic Retroviral Reactivation in Human Preimplantation Embryos and Pluripotent Cells.” Nature 522 (7555): 221–25. https://doi.org/10.1038/nature14308.

Hassan, Nozhat T., and David L. Adelson. 2023. “Fake IDs? Widespread Misannotation of DNA Transposons as a General Transcription Factor.” *Genome Biology* 24 (1): 260. https://doi.org/10.1186/s13059-023-03102-9.

Hermant, Clara, and Maria-Elena Torres-Padilla. 2021. “TFs for TEs: The Transcription Factor Repertoire of Mammalian Transposable Elements.” Genes & Development 35 (1–2): 22–39. https://doi.org/10.1101/gad.344473.120.

Hoen, Douglas R., Glenn Hickey, Guillaume Bourque, Josep Casacuberta, Richard Cordaux, Cédric Feschotte, Anna-Sophie Fiston-Lavier, et al. 2015. “A Call for Benchmarking Transposable Element Annotation Methods.” Mobile DNA 6 (1): 13. https://doi.org/10.1186/s13100-015-0044-6.

Hoyt, Savannah J., Jessica M. Storer, Gabrielle A. Hartley, Patrick G. S. Grady, Ariel Gershman,

Leonardo G. de Lima, Charles Limouse, et al. 2022. “From Telomere to Telomere: The Transcriptional and Epigenetic State of Human Repeat Elements.” Science 376 (6588): eabk3112. https://doi.org/10.1126/science.abk3112.

Imbeault, Michaël, Pierre-Yves Helleboid, and Didier Trono. 2017. “KRAB Zinc-Finger Proteins Contribute to the Evolution of Gene Regulatory Networks.” Nature 543 (7646): 550–54. https://doi.org/10.1038/nature21683.

Inoue, Fumitaka, Anat Kreimer, Tal Ashuach, Nadav Ahituv, and Nir Yosef. 2019. “Identification and Massively Parallel Characterization of Regulatory Elements Driving Neural Induction.” Cell Stem Cell 25 (5): 713-727.e10. https://doi.org/10.1016/j.stem.2019.09.010.

Ito, Jumpei, Ryota Sugimoto, Hirofumi Nakaoka, Shiro Yamada, Tetsuaki Kimura, Takahide Hayano, and Ituro Inoue. 2017. “Systematic Identification and Characterization of Regulatory Elements Derived from Human Endogenous Retroviruses.” PLOS Genetics 13 (7): e1006883. https://doi.org/10.1371/journal.pgen.1006883.

Jacques, Pierre-Étienne, Justin Jeyakani, and Guillaume Bourque. 2013. “The Majority of Primate-Specific Regulatory Sequences Are Derived from Transposable Elements.” PLOS Genetics 9 (5): e1003504. https://doi.org/10.1371/journal.pgen.1003504.

Jurka, J., V.V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. 2005. “Repbase Update, a Database of Eukaryotic Repetitive Elements.” Cytogenetic and Genome Research 110 (1–4): 462–67. https://doi.org/10.1159/000084979.

Jurka, Jerzy. 1998. “Repeats in Genomic DNA: Mining and Meaning.” Current Opinion in Structural Biology 8 (3): 333–37. https://doi.org/10.1016/S0959-440X(98)80067-5.

Kheradpour, Pouya, Jason Ernst, Alexandre Melnikov, Peter Rogov, Li Wang, Xiaolan Zhang, Jessica Alston, Tarjei S. Mikkelsen, and Manolis Kellis. 2013. “Systematic Dissection of Regulatory Motifs in 2000 Predicted Human Enhancers Using a Massively Parallel Reporter Assay.” Genome Research 23 (5): 800–811. https://doi.org/10.1101/gr.144899.112.

Kunarso, Galih, Na-Yu Chia, Justin Jeyakani, Catalina Hwang, Xinyi Lu, Yun-Shen Chan, Huck-Hui Ng, and Guillaume Bourque. 2010. “Transposable Elements Have Rewired the Core Regulatory Network of Human Embryonic Stem Cells.” Nature Genetics 42 (7): 631–34. https://doi.org/10.1038/ng.600.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. “Initial Sequencing and Analysis of the Human Genome.” Nature 409 (6822): 860–921. https://doi.org/10.1038/35057062.

Le Rouzic, Arnaud, Thibaud S. Boutin, and Pierre Capy. 2007. “Long-Term Evolution of Transposable Elements.” Proceedings of the National Academy of Sciences 104 (49): 19375–80. https://doi.org/10.1073/pnas.0705238104.

Leigh, Jessica W., and David Bryant. 2015. “Popart: Full-Feature Software for Haplotype Network Construction.” Methods in Ecology and Evolution 6 (9): 1110–16. https://doi.org/10.1111/2041-210X.12410.

Löytynoja, Ari. 2014. “Phylogeny-Aware Alignment with PRANK.” Methods in Molecular Biology (Clifton, N.J.) 1079: 155–70. https://doi.org/10.1007/978-1-62703-646-7_10.

Ma, Gang, Isaac A. Babarinde, Xuemeng Zhou, and Andrew P. Hutchins. 2022. “Transposable Elements in Pluripotent Stem Cells and Human Disease.” Frontiers in Genetics 13: 902541. https://doi.org/10.3389/fgene.2022.902541.

Mayer, J., and E. Meese. 2005. “Human Endogenous Retroviruses in the Primate Lineage and Their Influence on Host Genomes.” Cytogenetic and Genome Research 110 (1–4): 448–56. https://doi.org/10.1159/000084977.

Minh, Bui Quang, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear. 2020. “IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.” Molecular Biology and Evolution 37 (5): 1530–34. https://doi.org/10.1093/molbev/msaa015.

Molaro, Antoine, and Harmit S. Malik. 2016. “Hide and Seek: How Chromatin-Based Pathways Silence Retroelements in the Mammalian Germline.” Current Opinion in Genetics & Development 37 (April): 51–58. https://doi.org/10.1016/j.gde.2015.12.001.

Nakamura, Tsukasa, Kazunori D. Yamada, Kentaro Tomii, and Kazutaka Katoh. 2018. “Parallelization of MAFFT for Large-Scale Multiple Sequence Alignments.” Bioinformatics (Oxford, England) 34 (14): 2490–92. https://doi.org/10.1093/bioinformatics/bty121.

Naser-Khdour, Suha, Bui Quang Minh, and Robert Lanfear. 2022. “Assessing Confidence in Root Placement on Phylogenies: An Empirical Study Using Nonreversible Models for Mammals.” Systematic Biology 71 (4): 959–72. https://doi.org/10.1093/sysbio/syab067.

Patoori, Sruti, Samantha M. Barnada, Christopher Large, John I. Murray, and Marco Trizzino. 2022. “Young Transposable Elements Rewired Gene Regulatory Networks in Human and Chimpanzee Hippocampal Intermediate Progenitors.” Development 149 (19): dev200413. https://doi.org/10.1242/dev.200413.

Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” Bioinformatics 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.

Rosspopoff, Olga, and Didier Trono. 2023. “Take a Walk on the KRAB Side.” Trends in Genetics 39 (11): 844–57. https://doi.org/10.1016/j.tig.2023.08.003.

Scognamiglio, Sante, Nicole Grandi, Eleonora Pessiu, and Enzo Tramontano. 2023. “Identification, Comprehensive Characterization, and Comparative Genomics of the HERV-K(HML8) Integrations in the Human Genome.” Virus Research 323 (January): 198976. https://doi.org/10.1016/j.virusres.2022.198976.

Senft, Anna D., and Todd S. Macfarlan. 2021. “Transposable Elements Shape the Evolution of Mammalian Development.” Nature Reviews Genetics 22 (11): 691–711. https://doi.org/10.1038/s41576-021-00385-1.

Sexton, Corinne E., Richard L. Tillett, and Mira V. Han. 2021. “The Essential but Enigmatic Regulatory Role of HERVH in Pluripotency.” Trends in Genetics 38 (1): 12–21. https://doi.org/10.1016/j.tig.2021.07.007.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. “Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.” Genome Research 13 (11): 2498–2504. https://doi.org/10.1101/gr.1239303.

Sotero-Caio, Cibele G., Roy N. Platt II, Alexander Suh, and David A. Ray. 2017. “Evolution and Diversity of Transposable Elements in Vertebrate Genomes.” Genome Biology and Evolution 9 (1): 161–77. https://doi.org/10.1093/gbe/evw264.

Sundaram, Vasavi, Yong Cheng, Zhihai Ma, Daofeng Li, Xiaoyun Xing, Peter Edge, Michael P. Snyder, and Ting Wang. 2014. “Widespread Contribution of Transposable Elements to the Innovation of Gene Regulatory Networks.” Genome Research 24 (12): 1963–76. https://doi.org/10.1101/gr.168872.113.

Trizzino, Marco, YoSon Park, Marcia Holsbach-Beltrame, Katherine Aracena, Katelyn Mika, Minal Caliskan, George H Perry, Vincent J Lynch, and Christopher D Brown. 2017. “Transposable Elements Are the Primary Source of Novelty in Primate Gene Regulation.” Genome Research 27 (10): 1623–33. https://doi.org/10.1101/gr.218149.116.

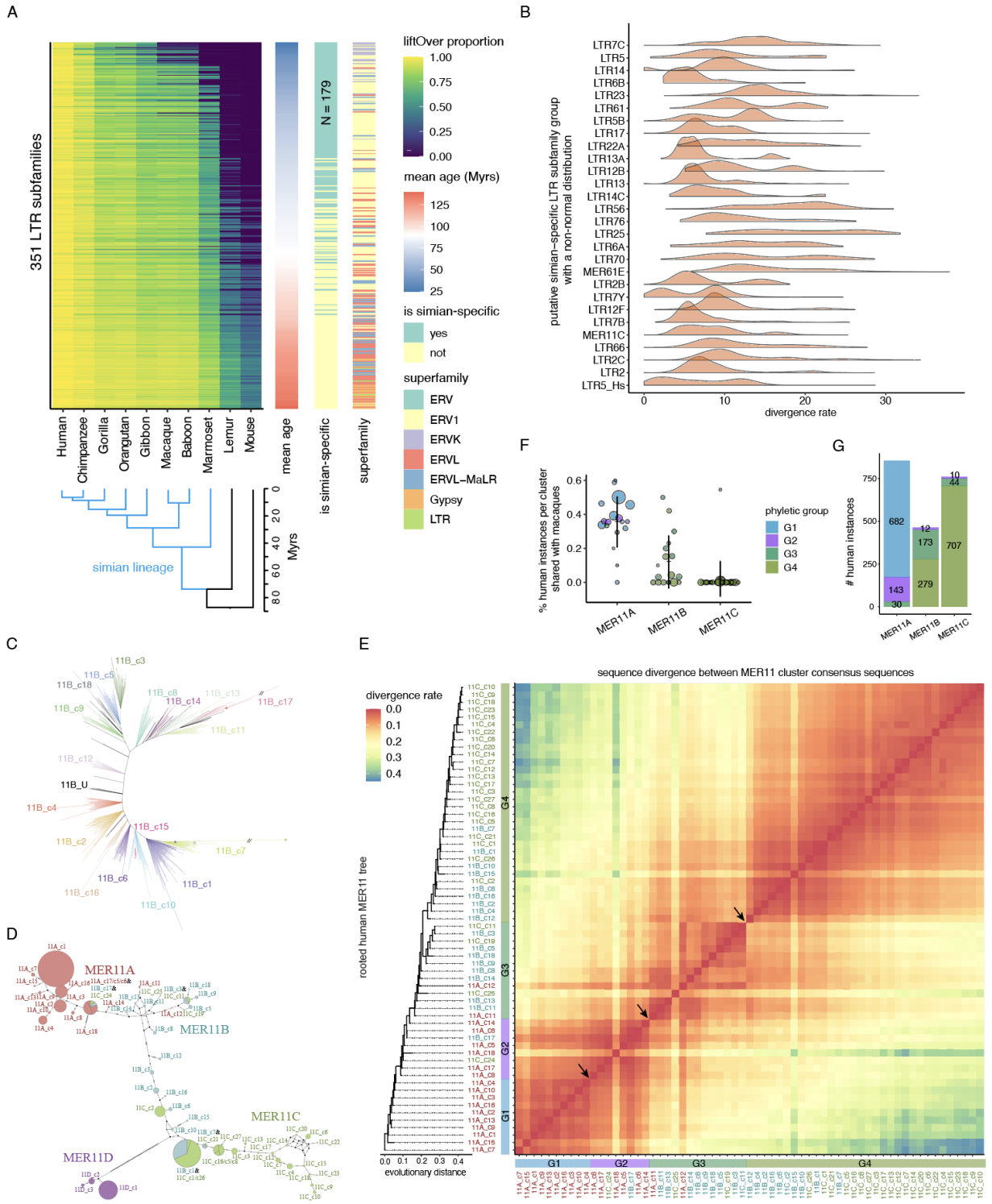
Wang, Ting, Jue Zeng, Craig B. Lowe, Robert G. Sellers, Sofie R. Salama, Min Yang, Shawn M. Burgess, Rainer K. Brachmann, and David Haussler. 2007. “Species-Specific Endogenous Retroviruses Shape the Transcriptional Network of the Human Tumor Suppressor Protein P53.” Proceedings of the National Academy of Sciences 104 (47): 18613–18. https://doi.org/10.1073/pnas.0703637104.

Wicker, Thomas, François Sabot, Aurélie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, et al. 2007. “A Unified Classification System for Eukaryotic Transposable Elements.” *Nature Reviews Genetics* 8 (12): 973–82. https://doi.org/10.1038/nrg2165.

Xiang, Xinyu, Yu Tao, Jonathan DiRusso, Fei-Man Hsu, Jinchun Zhang, Ziwei Xue, Julien Pontis, Didier Trono, Wanlu Liu, and Amander T. Clark. 2022. “Human Reproduction Is Regulated by Retrotransposons Derived from Ancient Hominidae-Specific Viral Infections.” Nature Communications 13 (1): 463. https://doi.org/10.1038/s41467-022-28105-1.

Xie, Wei, Matthew D. Schultz, Ryan Lister, Zhonggang Hou, Nisha Rajagopal, Pradipta Ray, John W. Whitaker, et al. 2013. “Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells.” Cell 153 (5): 1134–48. https://doi.org/10.1016/j.cell.2013.04.022.

1 **FIGURES**



2

3 **Figure 1. The sequence heterogeneity of simian-specific LTR subfamilies.**

4 (A) LiftOver analysis of LTR subfamilies from human (hg19) to other primate and mouse genomes (see

5 Methods). Evolutionary ages were estimated based on the divergence rates. LTR subfamilies that have a

minimum of 100 instances (≥ 200 bp) and a maximum of 20% of instances shared with the lemur genome were selected.

(B) Divergence rate (% substitutions) distribution of instances relative to the subfamily consensus sequence. Subfamilies that have unexpected distributions (Chi-square test, Bonferroni adjusted *p* values < 0.001) are shown.

(C) Unrooted tree of MER11B instances showing 18 clusters (labeled 11B_c1 to c18).

(D) Median-joining network of the 66 MER11 clusters. Circle sizes refer to the relative number of instances between clusters. Ticks indicate the number of nucleotide mutations between cluster consensus sequences. Clusters derived from different MER11 subfamilies are in different colors. Pie charts indicate clusters that have the same consensus sequences after the removal of gaps.

(E) Rooted tree containing 63 MER11A/B/C clusters. Heatmap displays divergence rates between each pair of cluster consensus sequences. Four phyletic groups are identified using different colors (see Methods). Arrows indicate the high divergence rates between clusters from adjacent phyletic groups. Clusters derived from each subfamily are in different colors.

(F) Proportion of human instances shared with macaques. Dot size refers to the number of instances per cluster.

(G) Number of instances per MER11A/B/C subfamily assigned to each phyletic group.

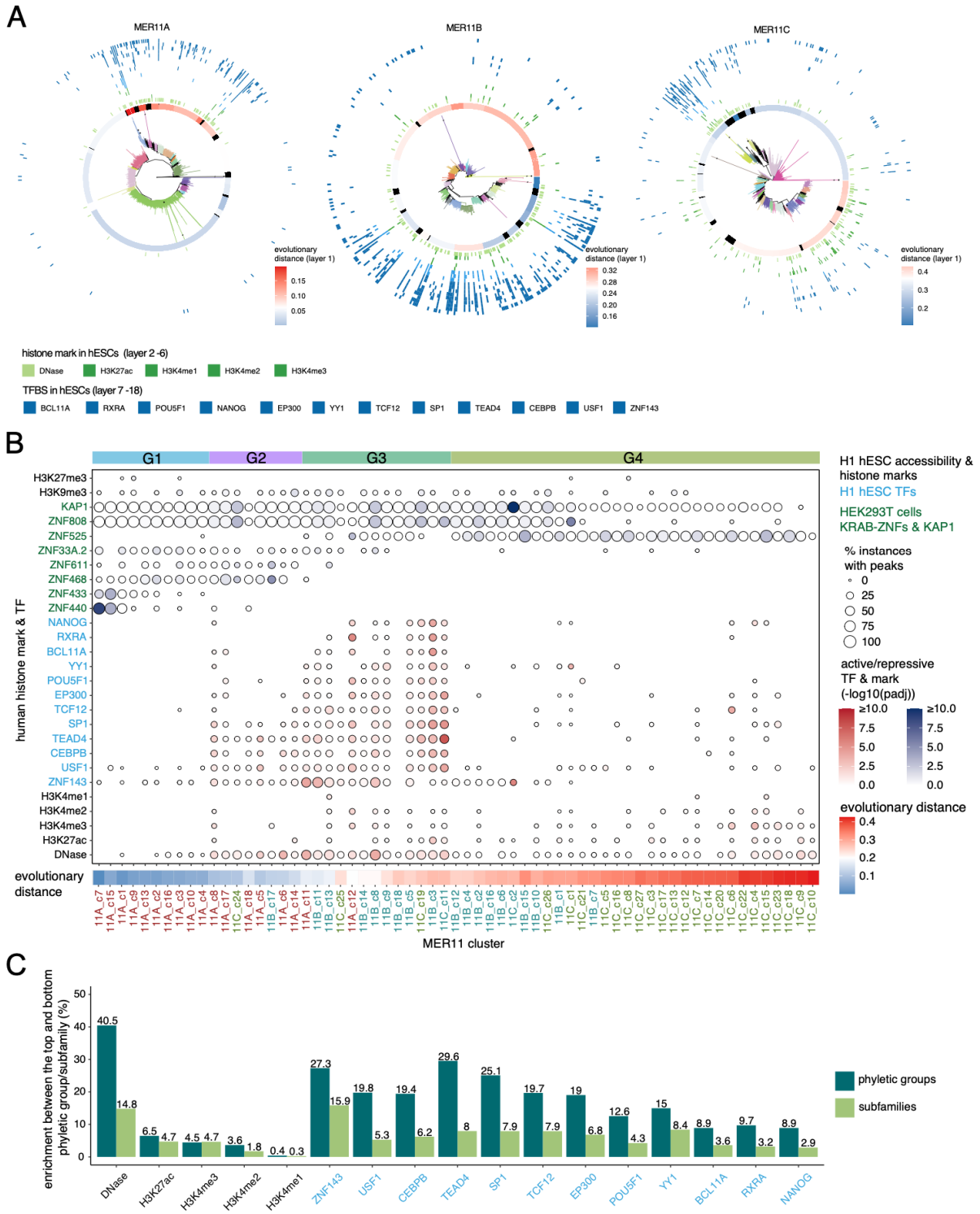


Figure 2. MER11 phyletic groups display more consistent epigenetic profiles as compared to MER11 subfamilies.

(A) Circular plots of MER11A/B and C. Unrooted trees are used to order instances (center). Evolutionary distance per cluster was calculated using the root from the tree of **Figure 1E**. Active histone marks and significantly enriched TFBSs in any MER11 subfamilies are shown (see Methods).

(B) Enrichment of active histone marks and TF peaks for every MER11 cluster ordered based on the evolutionary distance from the root. Active and repressive histone marks and significantly enriched TFs and KRAB-ZNFs in any MER11 subfamily are shown. Phyletic groups are highlighted on the top and separated by dotted lines. Clusters derived from each subfamily are colored differently.

(C) Enrichment of active histone marks and TF peaks in phyletic groups versus subfamilies. Enrichment was computed as the proportion of peaks-associated instances in the top phyletic group, the one with the highest proportion of peak-associated instances, minus the proportion in the bottom phyletic group, the one with the lowest proportion, for each histone mark and TF (in blue). The same was calculated between the top subfamily and the bottom subfamily.

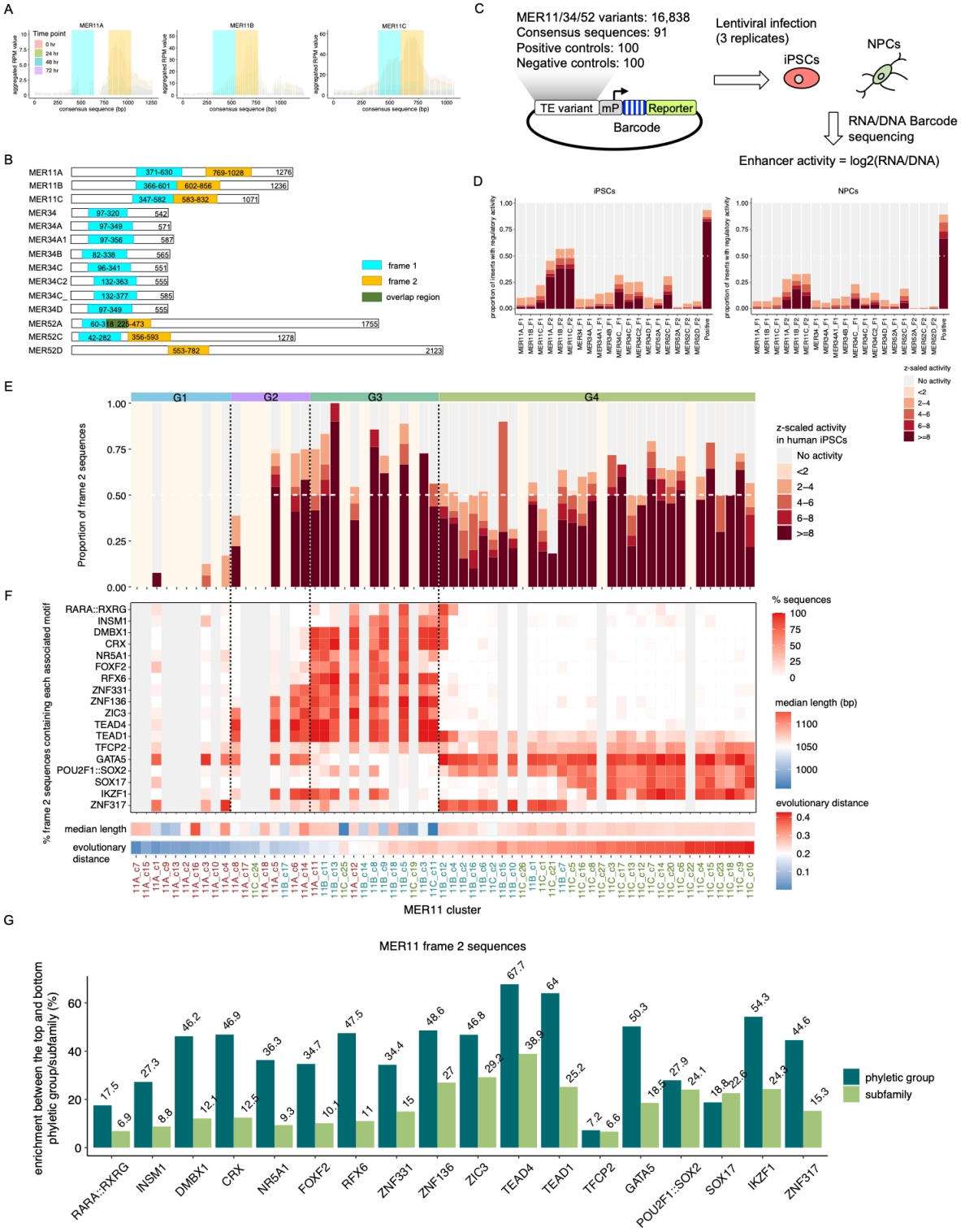


Figure 3. MPRA and MER11 phyletic groups help resolve the functional heterogeneity of MER11 subfamilies.

(A) Determination of accessible regions along the consensus sequence per subfamily. ATAC-seq RPM (reads per million) values are aggregated across instances.

(B) MER11, MER34, and MER52 consensus sequence frames designed for measuring enhancer activity using MPRA.

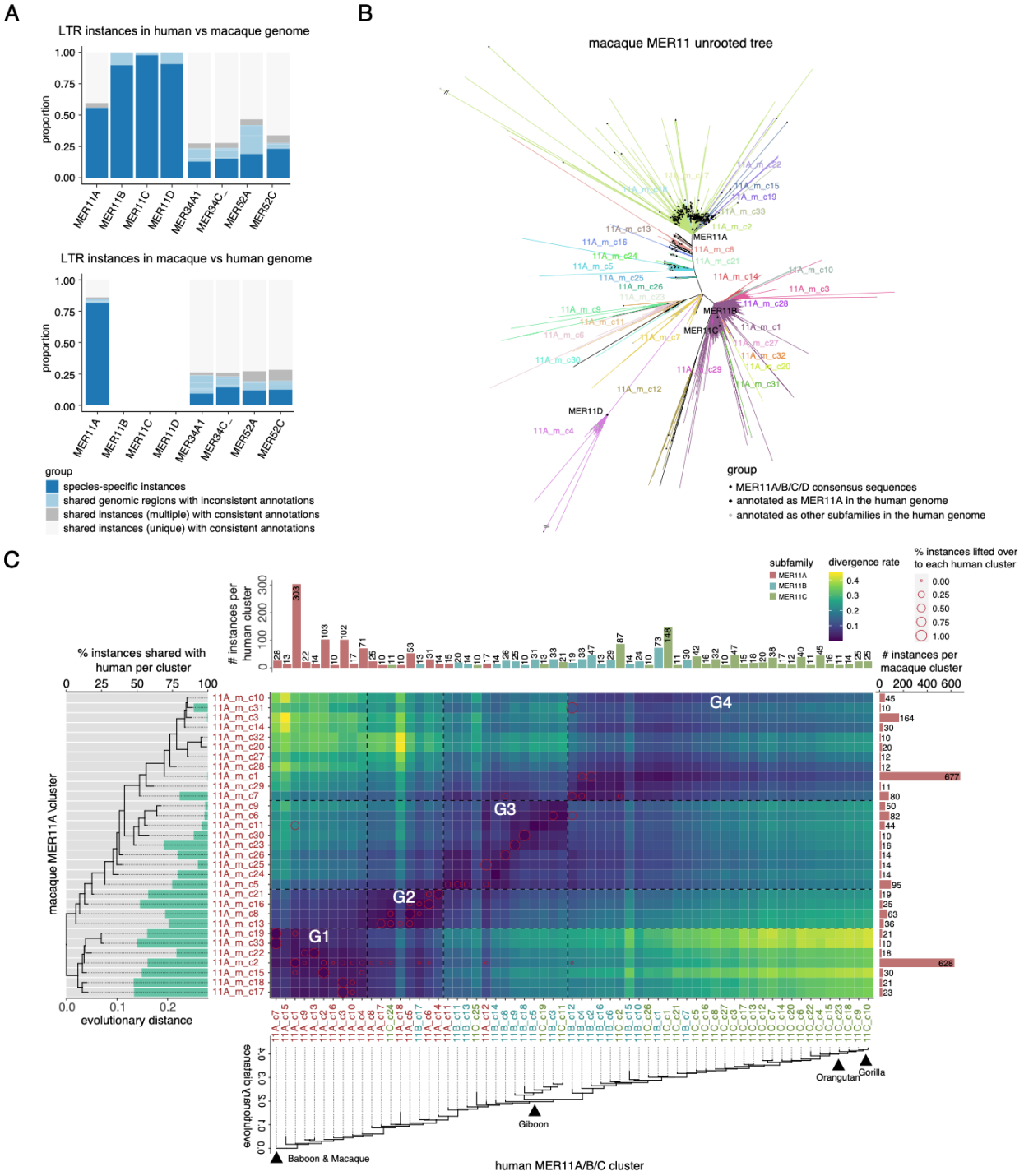
(C) MPRA experiment workflow. MER11/34/52 variants, consensus, and control sequences were inserted into the MPRA vector with random barcodes. The MPRA library was infected into iPSCs or NPCs using lentivirus with three replicates. RNA and DNA barcodes were measured to quantify their enhancer activity.

(D) Proportion of active sequences per sequence frame. Chimpanzee and macaque sequences orthologous to each cluster are also included. The normalization of MPRA activity was described in Methods.

(E) Proportion of active MER11 frame 2 sequences per cluster. Clusters with less than 10 instances measured by MPRA are not shown. Background is in light pink color. Chimpanzee and macaque frame 2 sequences orthologous to each cluster are also included.

(F) Percentage of associated motifs in the frame 2 sequences across MER11 clusters and phyletic groups. Chimpanzee and macaque sequences are also included.

(G) Motif enrichment in phyletic groups versus subfamilies. Enrichment was computed as the proportion of instances containing each motif in the top phyletic group, the one with the highest proportion, minus the bottom phyletic group, the one with the lowest proportion. Motif enrichment was computed similarly between the top and the bottom subfamily.


A
LTR instances in human vs macaque genome
LTR instances in macaque vs human genome
proportion
MER11A
MER11B
MER11C
MER11D
MER34A1
MER34C_
MER52A
MER52C
group
species-specific instances
shared genomic regions with inconsistent annotations
shared instances (multiple) with consistent annotations
shared instances (unique) with consistent annotations
B
macaque MER11 unrooted tree
group
MER11A/B/C/D consensus sequences
annotated as MER11A in the human genome
annotated as other subfamilies in the human genome
C
subfamily
divergence rate
% instances lifted over to each human cluster
# instances per human cluster
% instances shared with human per cluster
# instances per macaque cluster
macaque MER11A \cluster
evolutionary distance
G1
G2
G3
G4
Baboon & Macaque
Gibbon
Orangutan
Gorilla
human MER11A/B/C cluster
D
liftOver proportion
G1
G2
G3
G4
E
human
macaque
No. instances
F
MER11 frame 1
MER11 frame 2
total N:
proportion
z-scaled activity in human iPSCs
No activity
<2
2–4
4–6
6–8
>=8

1

Figure 4. The presence of MER11 phyletic groups in both human and macaque lineages.

(A) Conservation and annotation of MER11 instances in human versus macaque based on liftOver and RepeatMasker. Instances intersected with regions annotated as the same or different subfamilies in another species are shown separately. Instances intersected with unique or multiple regions are also shown separately.

(B) Unrooted tree of macaque MER11A instances. MER11A/B/C/D consensus sequences are included as references. Clusters are colored differently. MER11A_m_c4/c12 clusters are clustered with MER11D consensus sequence (**Figure S9B**) and removed from further analyses.

(C) The comparison of MER11 clusters and phyletic groups between humans and macaques. Clusters are arranged by the macaque and human cluster consensus sequence rooted trees. Clusters from every subfamily are colored differently. The primate lineage when they first integrated is highlighted. Human and macaque phyletic groups are separated by dotted lines.

(D) Comparison of the liftOver proportion per phyletic group between human and macaque.

(E) Number of instances per phyletic group between human and macaque.

(F) Frame1 and frame2 MPRA activity per phyletic group between human and macaque.

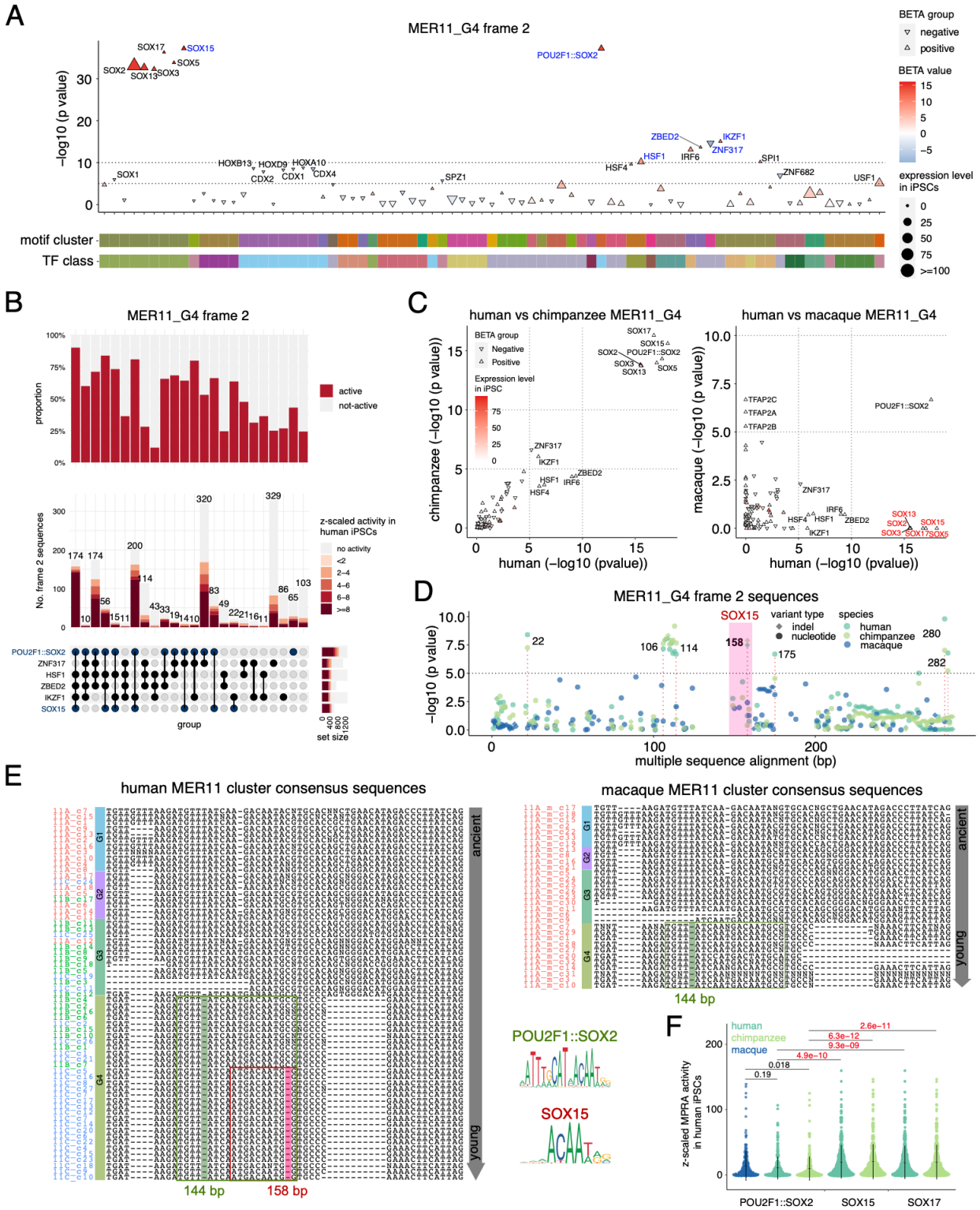


Figure 5. Nucleotide changes and gain of functional motifs during separate expansions of MER11_G4 in primate lineages.

(A) Motifs contributing to the activity of MER11_G4 frame 2 sequences. *P* values and effect sizes (BETA values) were computed by the linear regression model using Plink2.

(B) Upset plots of different sets of motifs and the proportion of frame 2 sequences with MPRA activity.
(C) Association between motifs and the MPRA activity in human, chimpanzee, and macaque MER11_G4 frame 2 sequences. *P* values were computed by the linear regression model using Plink2.
(D) Nucleotides associated with the MPRA activity in humans, chimpanzees, and macaques. *P* values were computed by the linear regression model using Plink2.
(E) Multiple sequence alignment of human and macaque cluster consensus sequences. Clusters derived from each MER11 subfamily are colored differently. Phyletic groups are highlighted. POU2F1::SOX2 and SOX15 motifs are highlighted in the boxes. Single-nucleotide deletions associated with the gain of motifs are also highlighted.
(F) Comparison of MPRA activity between MER11_G4 frame 2 sequences containing POU2F1::SOX2 and SOX15/17. Sequences containing both motifs are grouped as the sequences with SOX15/17. *P* values were computed using the student's *t*-test.

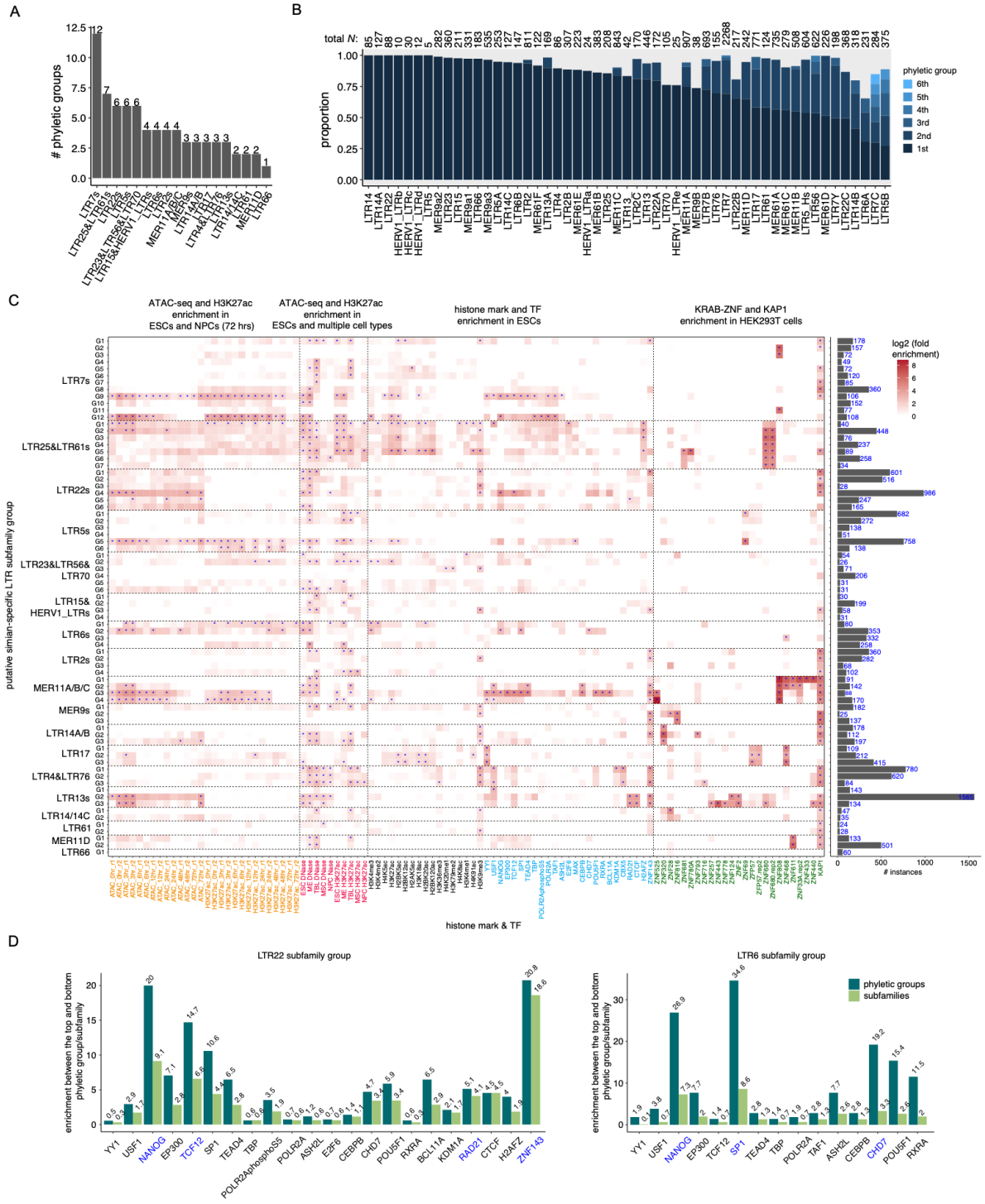


Figure 6. Cryptic endogenous retrovirus subfamilies in the primate lineage with distinct epigenetic profiles.

(A) Number of phyletic groups per subfamily group.

(B) Proportion of instances per subfamily classified into the top (representative) phyletic group. The top phyletic group contains the most instances per subfamily, second phyletic group contains the second most instances, and so on.

(C) Epigenetic profiles across phyletic groups of 18 simian-specific LTR subfamily groups. Permutation was used to compute the *p* values. Significantly enriched (log2((actual counts+1)/(mean shuffled counts + 1)) ≥ 1 and *p* value ≤ 0.05) phyletic groups relative to 100 random genomic controls are highlighted. (D) TF specificity of peaks-associated instances between phyletic groups and subfamilies in LTR22 and LTR6 subfamily groups. Enrichment was computed as the proportion of peaks-associated instances in the top phyletic group, the one with the highest proportion of peak-associated instances, minus the proportion in the bottom phyletic group, the one with the lowest proportion, for each TF. The same was calculated between the top subfamily and the bottom subfamily. Blue color indicates TFs significantly enriched in any phyletic group.

Supplementary Materials for:

Cryptic endogenous retrovirus subfamilies in the primate lineage

Xun Chen[1,#], Zicong Zhang[1], Yizhi Yan[1], Clement Goubert[2], Guillaume Bourque[1,2,3,4,#], Fumitaka Inoue[1,#]

[1] Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto 606-8501, Japan
[2] Department of Human Genetics, McGill University, Montréal, QC H3A 0C7, Canada
[3] Victor Phillip Dahdaleh Institute of Genomic Medicine at McGill University, Montréal, QC H3A 0G1, Canada
[4] Canadian Center for Computational Genomics, McGill University, Montréal, QC H3A 0G1, Canada

[#] Correspondence: Xun Chen (chen.xun.3r@kyoto-u.ac.jp), Guillaume Bourque (guil.bourque@mcgill.ca) and Fumitaka Inoue (inoue.fumitaka.7a@kyoto-u.ac.jp)

Supplementary Figures

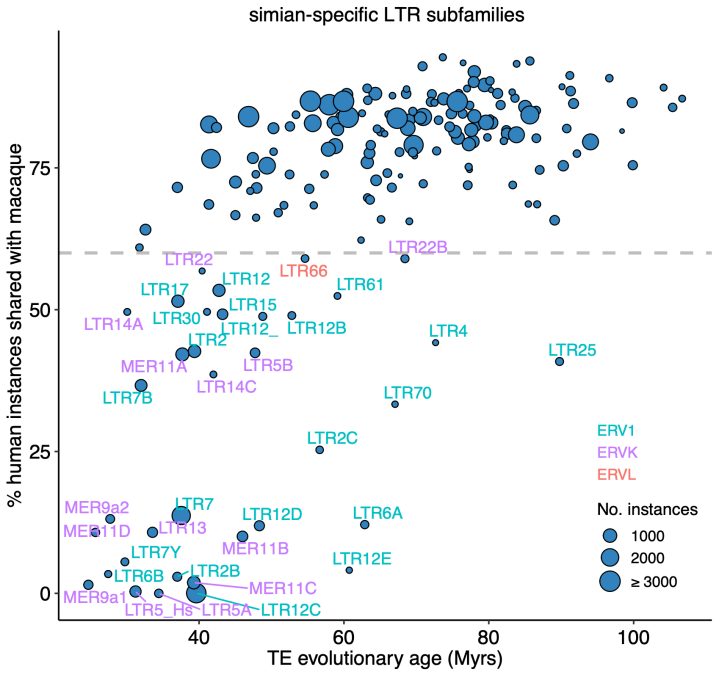


Figure S1. Proportion of instances in the human genome from the 179 simian-specific LTR subfamilies that are shared with macaques.

LTR subfamilies with less than 60% of instances shared with macaques are highlighted. Subfamilies derived from each ERV superfamily are colored differently.

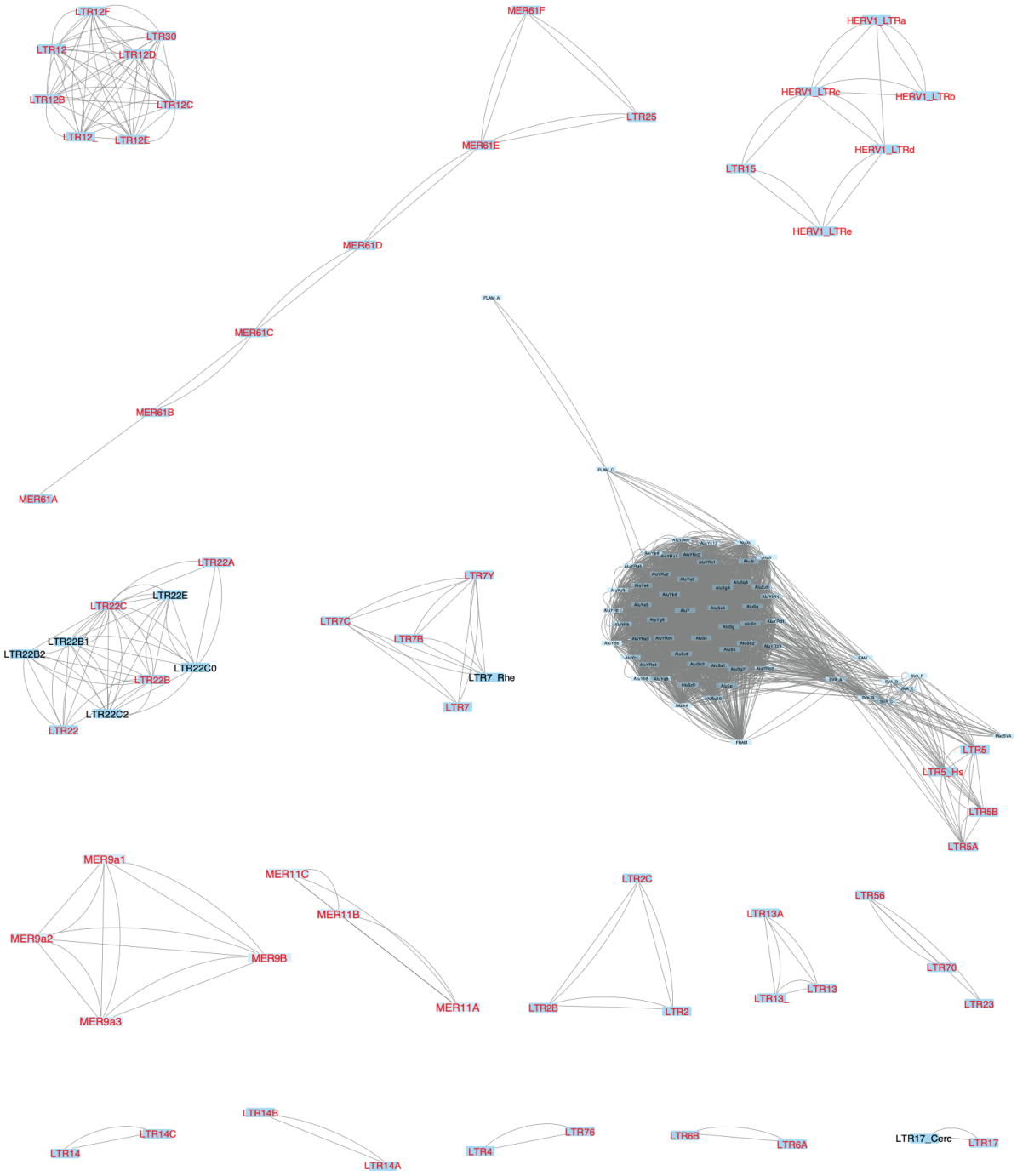
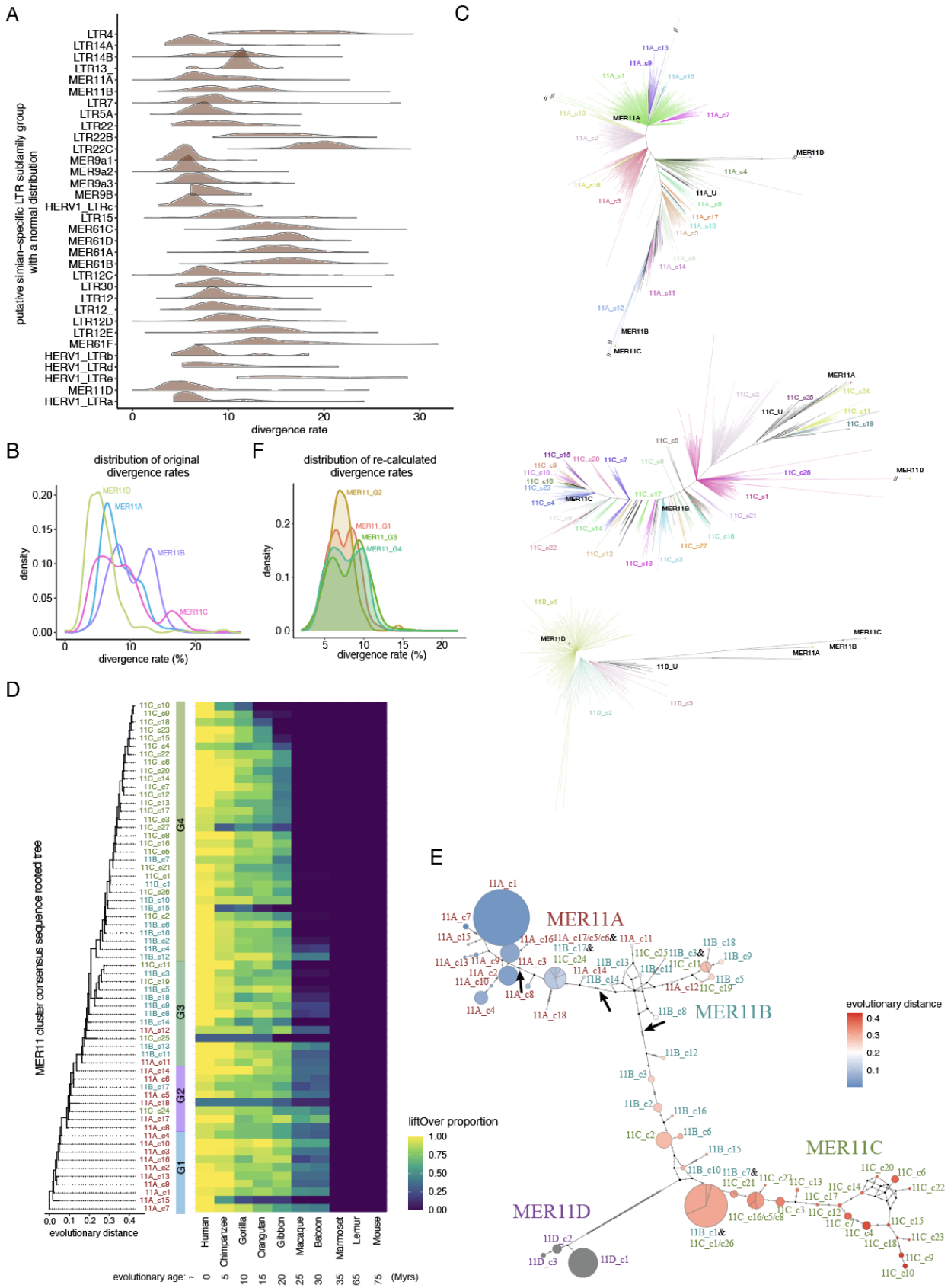


Figure S2. Putative simian-specific subfamily groups.

Visualization of 16 LTR subfamily groups except three groups with a single subfamily (LTR61, LTR66, and MER11D). Network analysis based on subfamily consensus sequence similarity was used to create clusters (see Methods). Subfamilies that are present in the human genome are highlighted in red. For the subfamily group containing LTR5 subfamilies, Alu and SVA non-LTR subfamilies were excluded from downstream analyses.


A
putative simian-specific LTR subfamily group with a normal distribution
LTR4
LTR14A
LTR14B
LTR13_
MER11A
MER11B
LTR7
LTR5A
LTR22
LTR22B
LTR22C
MER9a1
MER9a2
MER9a3
MER9B
HERV1_LTRc
LTR15
MER61C
MER61D
MER61A
MER61B
LTR12C
LTR30
LTR12
LTR12_
LTR12D
LTR12E
MER61F
HERV1_LTRb
HERV1_LTRd
HERV1_LTRe
MER11D
HERV1_LTRa
divergence rate
B
distribution of original divergence rates
density
MER11D
MER11A
MER11B
MER11C
divergence rate (%)
F
distribution of re-calculated divergence rates
MER11_G2
MER11_G1
MER11_G3
MER11_G4
C
MER11A
MER11B
MER11C
MER11D
D
MER11 cluster consensus sequence rooted tree
G1
G2
G3
G4
evolutionary distance
Human
Chimpanzee
Gorilla
Orangutan
Gibbon
Macaque
Baboon
Marmoset
Lemur
Mouse
evolutionary age: ~ 0 5 10 15 20 25 30 35 65 75 (Myrs)
liftOver proportion
E
MER11A
MER11B
MER11C
MER11D
evolutionary distance

1

Figure S3. Evolution of human MER11 subfamilies.

(A) Divergence rate distribution of instances relative to the subfamily consensus sequence. Subfamilies that have expected distributions (Chi-square test, Bonferroni adjusted *p* values ≥ 0.001) are shown.

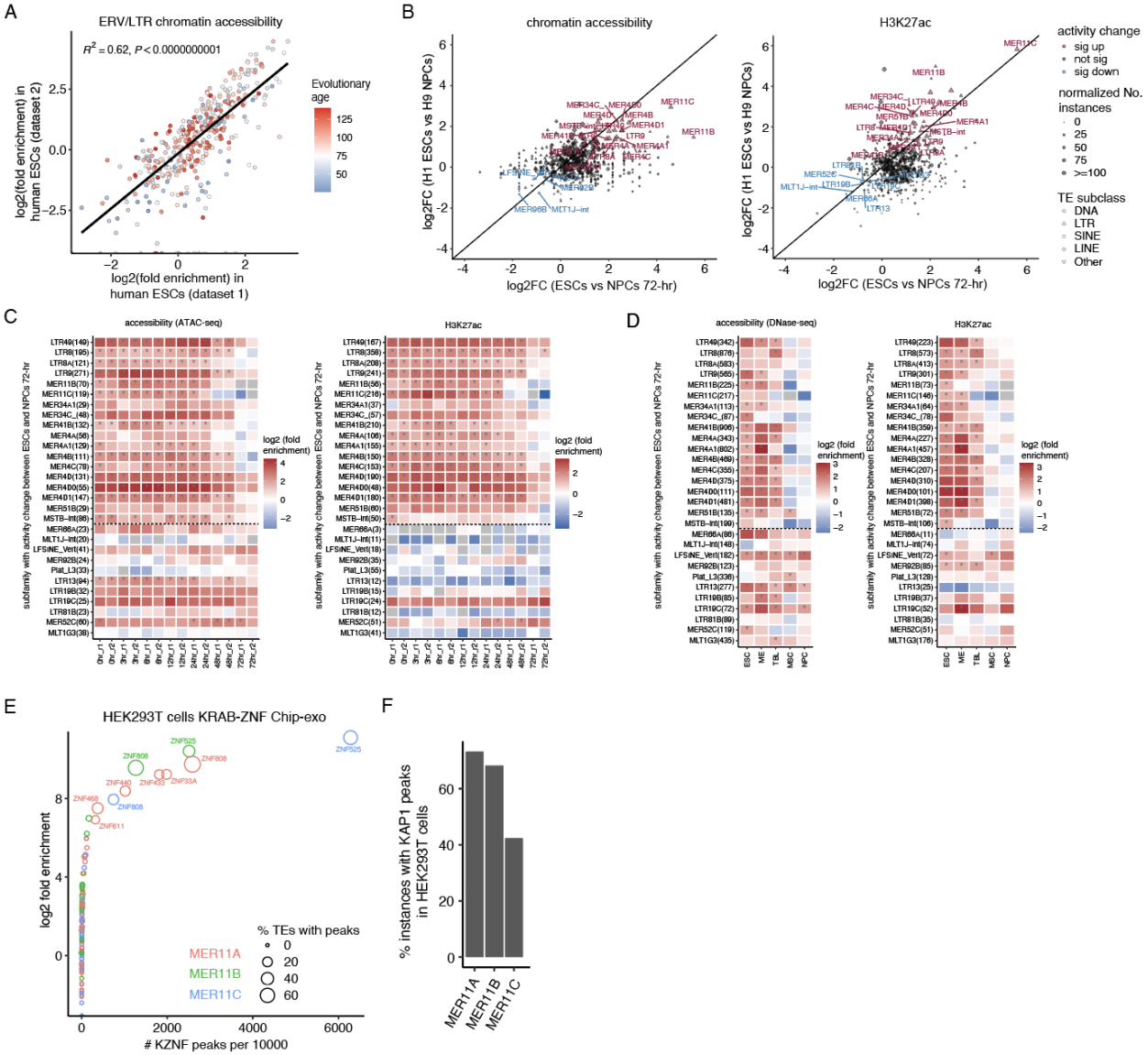
(B) Divergence rate distribution of MER11 subfamilies.

(C) Unrooted trees of MER11A, MER11C, and MER11D subfamilies. Human instances ≥ 200 bp were analyzed. Consensus sequences are in black and newly identified subfamilies are in different colors.

(D) Selected rooted tree containing 63 MER11A/B/C cluster consensus sequences and the proportion of instances shared with mouse and other primate lineages.

(E) Median-joining network of 66 MER11 cluster consensus sequences. Color refers to the evolutionary age. Size refers to the relative number of instances amongst clusters. The arrow indicates the edges have the most mutations between phyletic groups.

(F) Distribution of re-computed divergence rates of instances relative to each phyletic group consensus sequence (see Methods).


A
ERV/LTR chromatin accessibility
$R^2 = 0.62$, $P < 0.0000000001$
Evolutionary age
log2(fold enrichment) in human ESCs (dataset 2)
log2(fold enrichment) in human ESCs (dataset 1)
B
chromatin accessibility
H3K27ac
log2FC (H1 ESCs vs H9 NPCs)
log2FC (ESCs vs NPCs 72-hr)
activity change
sig up
not sig
sig down
normalized No. instances
TE subclass
DNA
LTR
SINE
LINE
Other
C
accessibility (ATAC-seq)
H3K27ac
D
accessibility (DNase-seq)
H3K27ac
log2 (fold enrichment)
subfamily with activity change between ESCs and NPCs 72-hr
E
HEK293T cells KRAB-ZNF Chip-exo
log2 fold enrichment
KZNF peaks per 10000
% TEs with peaks
MER11A
MER11B
MER11C
F
% instances with KAP1 peaks in HEK293T cells
MER11A
MER11B
MER11C
G
highest percentage between phyletic group/subfamily (%)
lowest percentage between phyletic group/subfamily (%)
phyletic groups
subfamilies
DNase
H3K27ac
H3K4me3
H3K4me2
H3K4me1
ZNF143
USF1
CEBPB
TEAD4
SP1
TCF12
EP300
POU5F1
YY1
BCL11A
RXRA
NANOG

1

Figure S4. Epigenetic profiles of human MER11A/B/C subfamilies.

(A) Correlation of log2 fold enrichment relative to the random genomic background between two ESCs from different sources. ERV/LTR subfamilies are included. Log2 fold enrichment was computed as the number of instances overlapped with the actual ATAC-seq or DNase peaks versus 1000 randomized peaks (see Methods). R^2 and *p* values were computed by the linear regression model.

(B) Log2 fold enrichment of the chromatin accessibility and H3K27ac mark per TE subfamily in ESCs relative to NPCs.

(C) Log2 fold enrichment of candidate subfamilies (**Figure S2B**) with increased or decreased accessibility and H3K27ac activity during the differentiation from ESCs to NPCs (72 hours).

(D) Log2 fold enrichment of candidate subfamilies (**Figure S2B**) across five cell types. Chromatin accessibility and H3K27ac were analyzed per cell type.

(E) Log2 fold enrichment of each KRAB-ZNF overlapped with each MER11 subfamily in HEK293T cells. Available KRAB-ZNF binding sites in HEK293T cells (Imbeault, Helleboid, and Trono 2017) overlapped with a minimum of 20 instances and more than two-fold enrichment related to the random genomic background with *p* value ≤ 0.05 are highlighted. *P* values were computed using the same permutation test. We consistently found the enrichment of ZNF525, ZNF808, ZNF440, ZNF433, and ZNF468 in these MER11 subfamilies. Moreover, we newly identified the enrichment of ZNF33A and ZNF611 in MER11A subfamily.

(F) Proportion of instances per MER11 subfamily overlapped with KAP1 in HEK293T cells. 73.2% of MER11A, 68.2% of MER11B, and 42.4% of MER11C are also bound by KAP1, confirming their potential repression in HEK293T cells.

(G) Highest and lowest proportion of peaks-associated instances among MER11 phyletic groups or subfamilies per epigenetic mark.

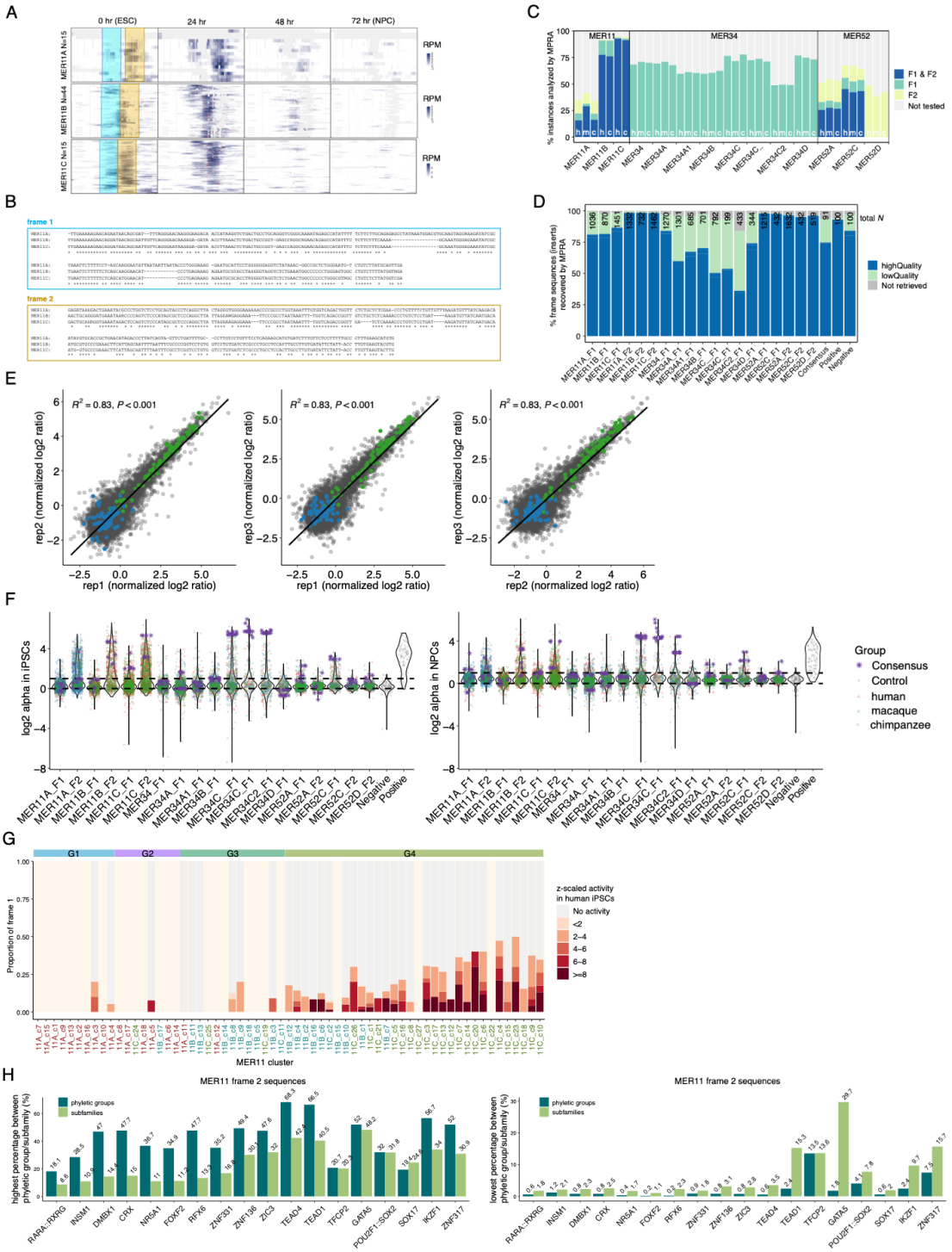
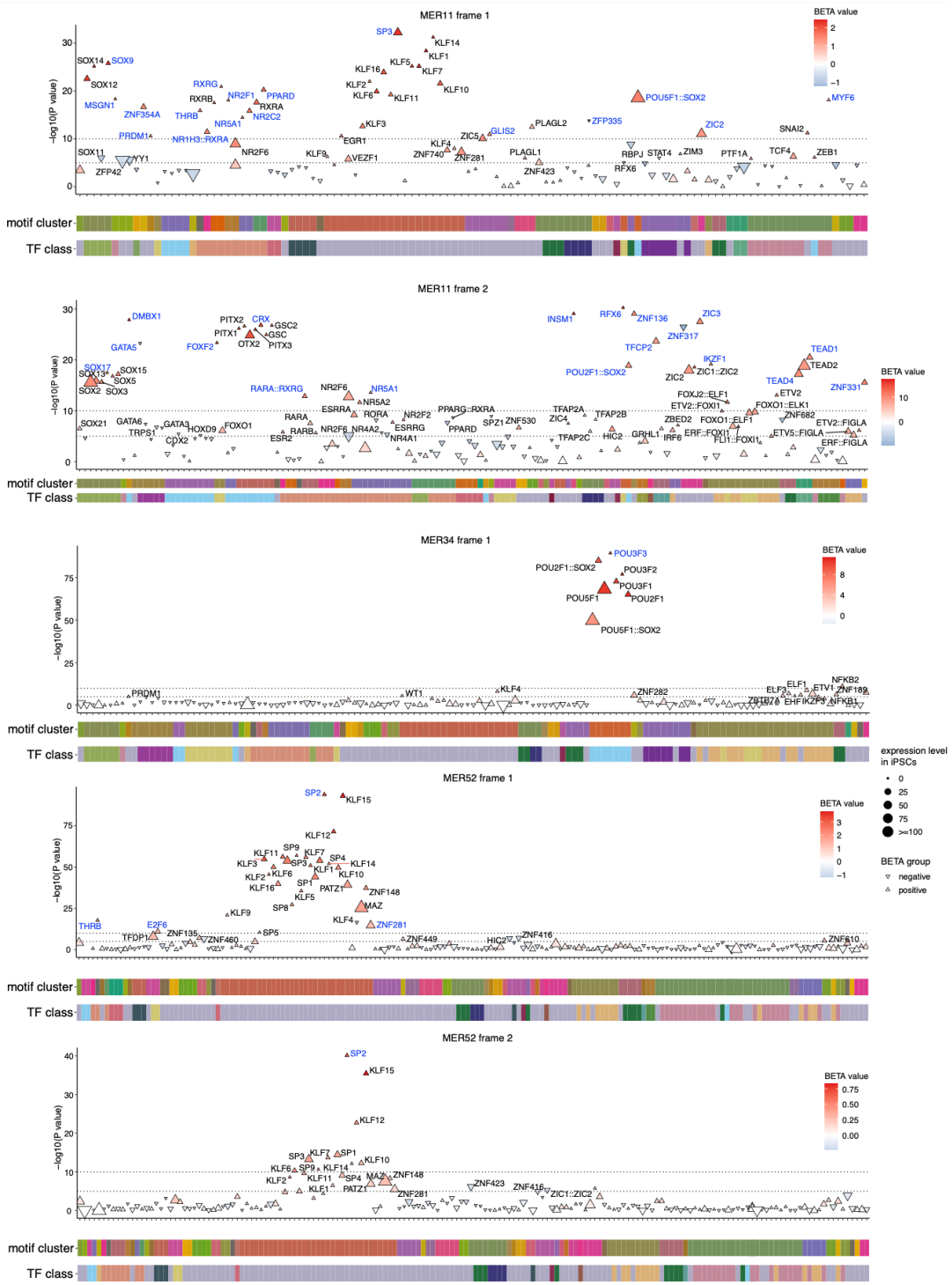


Figure S5. MPRA activity of MER11, MER34, and MER52 frame sequences in human iPSCs and differentiated NPCs.

(A) ATAC-seq read distribution (read per million) on each accessible MER11A/B/C instance along each consensus sequence. Frame 1 and frame 2 regions are highlighted in light blue and brown.

(B) Multiple sequence alignment of extracted MER11A/B/C frame 1/2 consensus sequences. ClustalW2 with default parameters (https://www.ebi.ac.uk/Tools/msa/clustalw2/) was used for the analysis.
(C) Proportion of genomic instances (≥ 200 bp) from each of MER11, MER34, and MER52 subfamilies analyzed by MPRA. Human (hg19), chimpanzee (panTro4) and macaque (macFas5) genomes were analyzed.
(D) Proportion of MER11, MER34, and MER52 1/2 frame sequences successfully retrieved by MPRA experiment. Number of consensus sequences refers to the sequences different from other frame sequences. High-quality inserts refer to examined frame sequences associated with ≥ 10 barcodes in two or more DNA libraries. Low-quality inserts refer to frame sequences associated with < 10 barcodes in less than two DNA libraries. Not retrieved inserts refer to frame sequences that fail to be associated with any barcodes in three replicates.
(E) Correlation of normalized log2 RNA/DNA ratio in human iPSCs between three replicates. Normalized RNA/DNA ratio was achieved by MPRAflow (see Methods). Positive and negative controls are in green and blue color.
(F) Violin plots of log2 alpha values per subfamily in human iPSCs and NPCs. Alpha value was computed by MPRAnalyze (see Methods). Dotted lines indicate the 0 and 1 log2 alpha values. Consensus sequences with different nucleotides for each ambiguous nucleotide position are included.
(G) Proportion of active MER11 frame 1 sequences per cluster. Clusters with less than 10 instances measured by MPRA are excluded and shown as the background color (light pink).
(H) Highest and lowest proportion of instances containing each motif amongst phyletic groups or subfamilies.



1

2 **Figure S6. Motifs associated with the MPRA activity.**

3 Motifs significantly associated with the MPRA activity are highlighted (-log10 *p* value ≥ 5) and the top

4 selected motifs amongst each motif group are highlighted in blue. *P* value refers to the significance and

was computed using plink2 (see Methods). BETA value refers to the effect size. Motif cluster and TF class information could be found at https://jaspar.genereg.net/matrix-clusters/vertebrates/?detail=false.

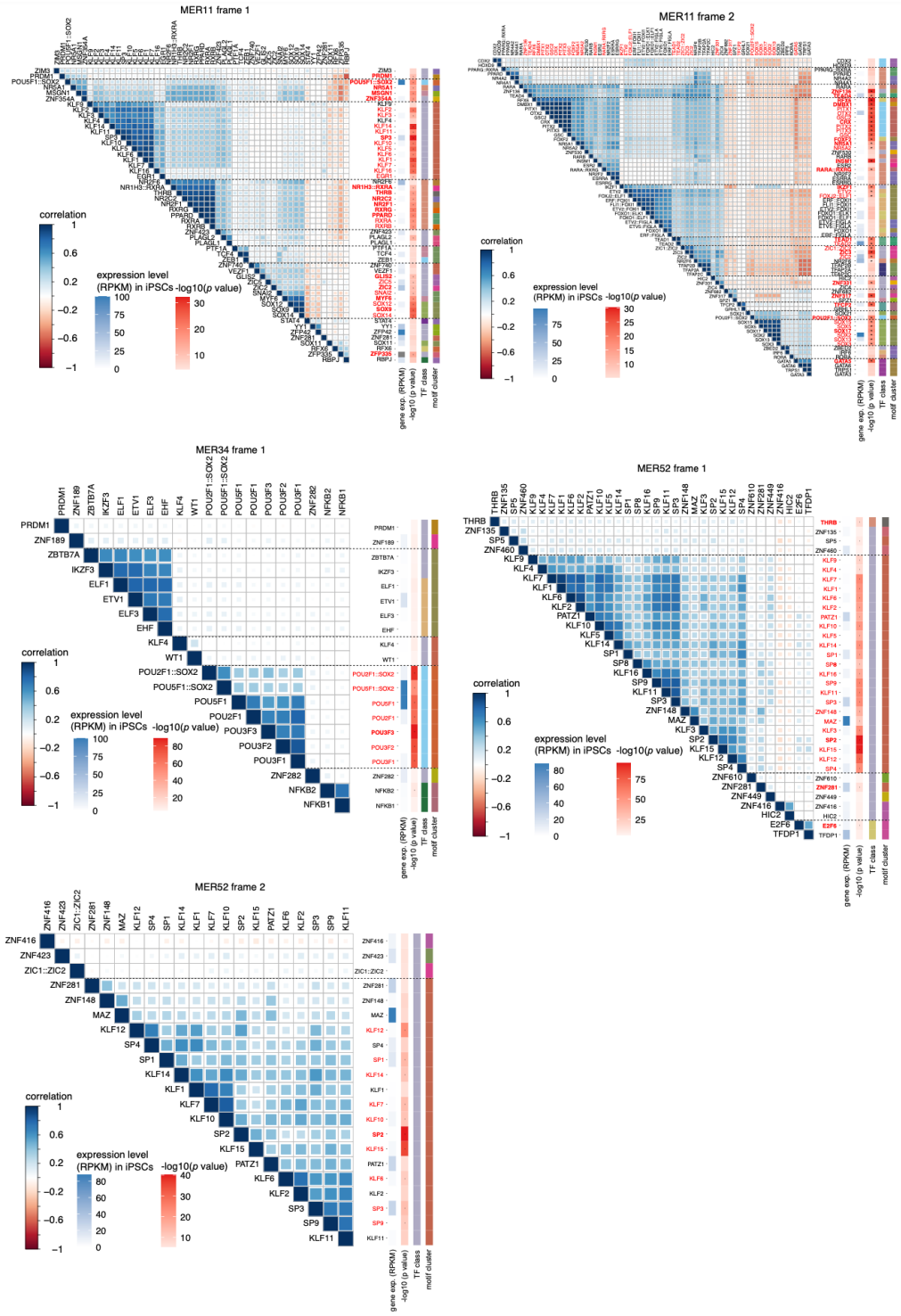


Figure S7. Correlation amongst candidate motifs per frame region.

Motifs with p value ≤ 1×10^{-5} are selected as the candidates. Motifs with p value ≤ 1×10^{-10} are highlighted in red. Dotted lines separate the different groups of strongly associated motifs. Top candidates within each group are highlighted in bold.

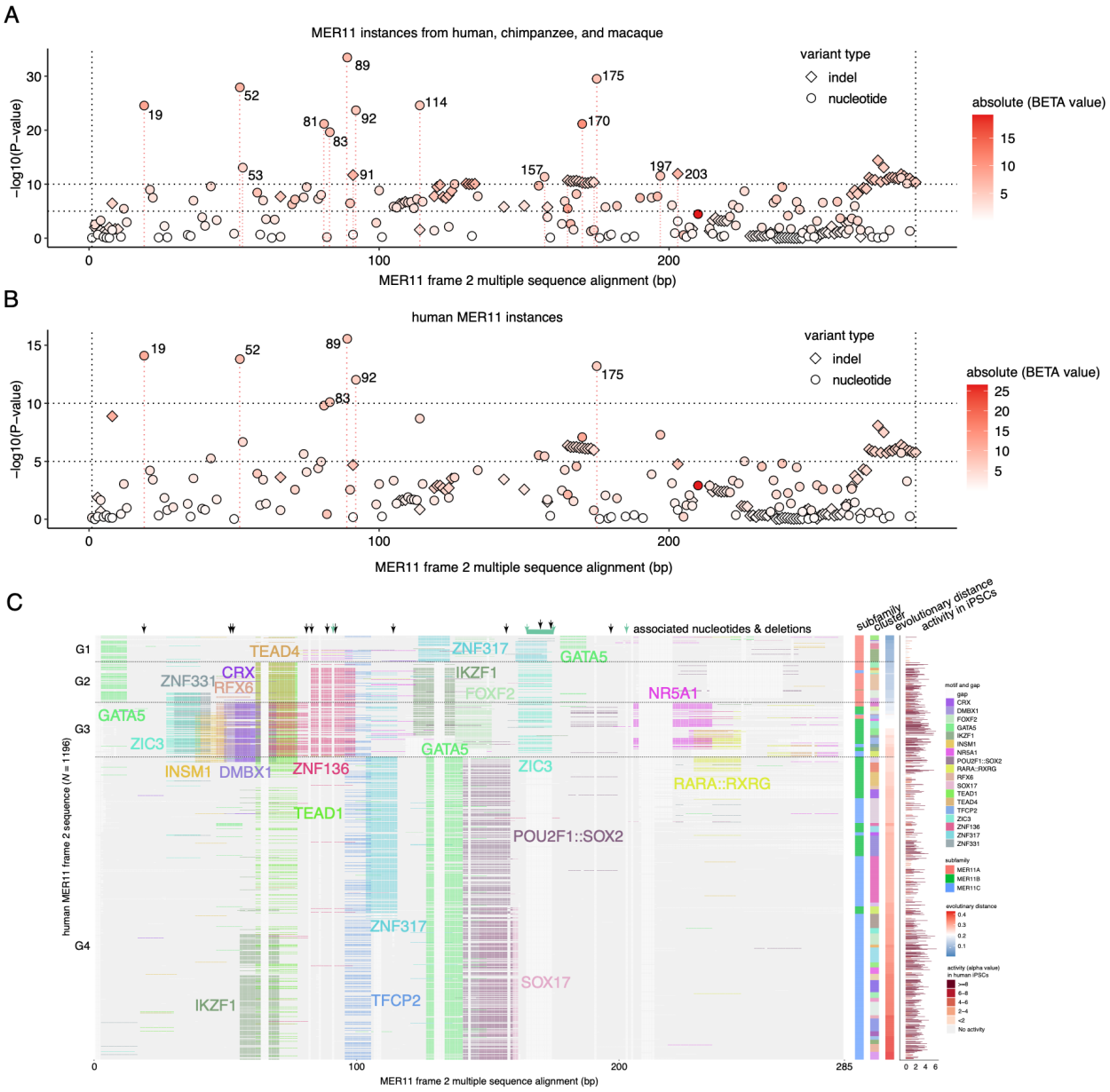


Figure S8. Nucleotides and motifs associated with the MPRA activity.

(A) Nucleotides significantly associated with the MPRA activity amongst all MER11 frame 2 genomic sequences. Nucleotide positions refer to the locations along the multiple sequence alignment. Alleles of each indel and nucleotide mutation are shown separately.

(B) Nucleotides significantly associated with the MPRA activity amongst human MER11 frame 2 sequences. Nucleotide positions refer to the locations along the multiple sequence alignment. Alleles of each indel and nucleotide change are shown separately.

(C) Positions of associated nucleotides and motifs on the human MER11 frame 2 multiple sequence alignment. Associated nucleotides detected using the human, chimpanzee, and macaque sequences are

shown. Detected motifs and nucleotides on the multiple sequence alignment of human sequences is shown as an example. Black and green arrows indicate the positions of strongly associated nucleotide mutations and indels.

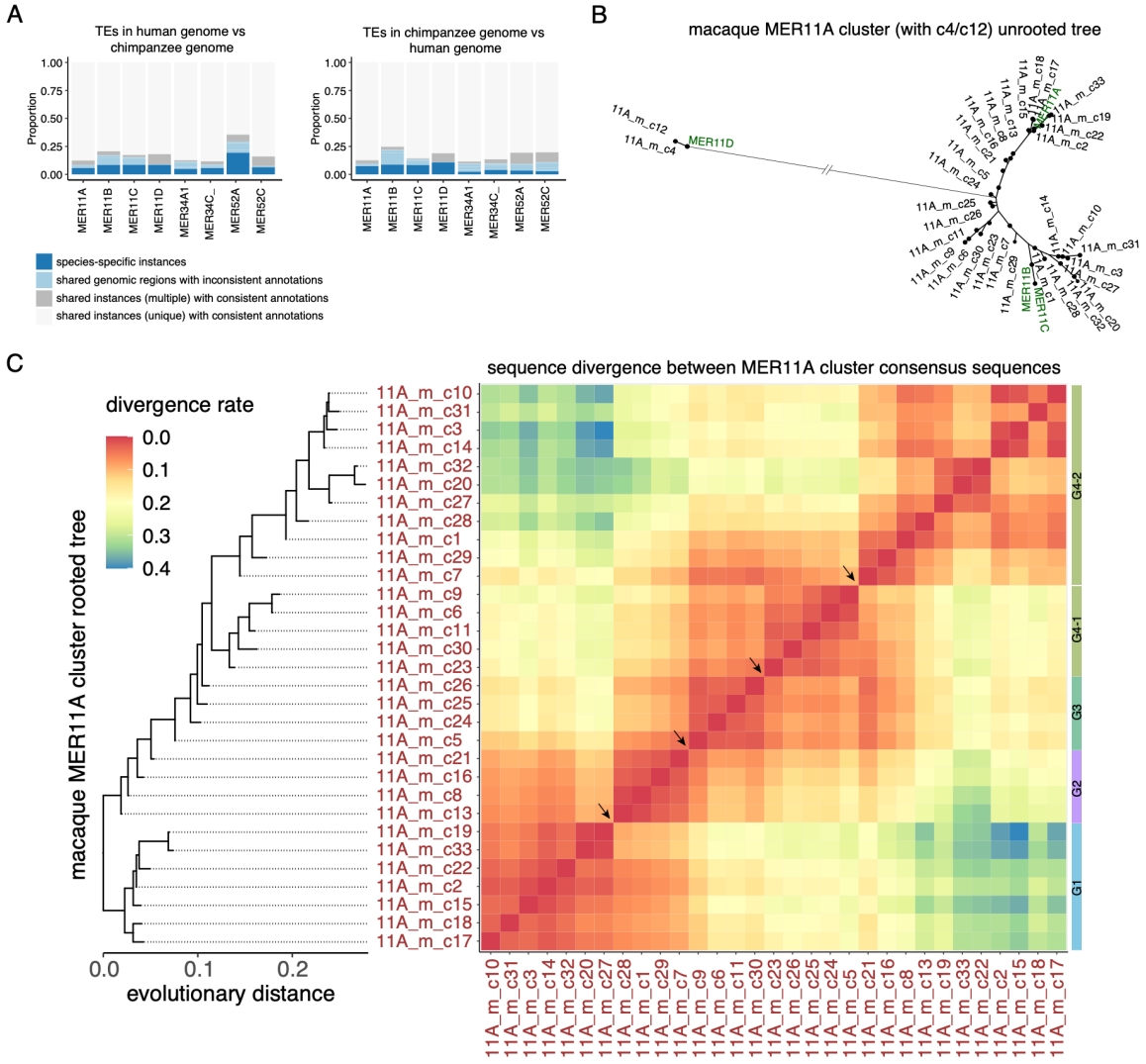


Figure S9. Evolution and MPRA activity of macaque MER11A subfamily.

(A) Conservation and annotation of MER11 instances in human versus chimpanzee based on liftOver and RepeatMasker. Instances intersected with regions annotated as the same or different subfamilies in another species are shown separately. Instances intersected with unique or multiple regions are also shown separately.

(B) Unrooted tree of all MER11A cluster consensus sequences. MER11A/B/C/D subfamily consensus sequences are highlighted in green.

(C) Rooted tree and divergent rates of macaque MER11A cluster consensus sequences. 11A_m_c4/c12 were excluded. Phyletic groups are determined as we previously described (see Methods). Due to the large number of macaque MER11A instances, phyletic groups with a minimum of 50 were kept. Arrows indicate the boundaries between them. Macaque MER11_G4-1/G4-2 are defined according to their sequence similarity with human MER11_G4 cluster consensus sequences.

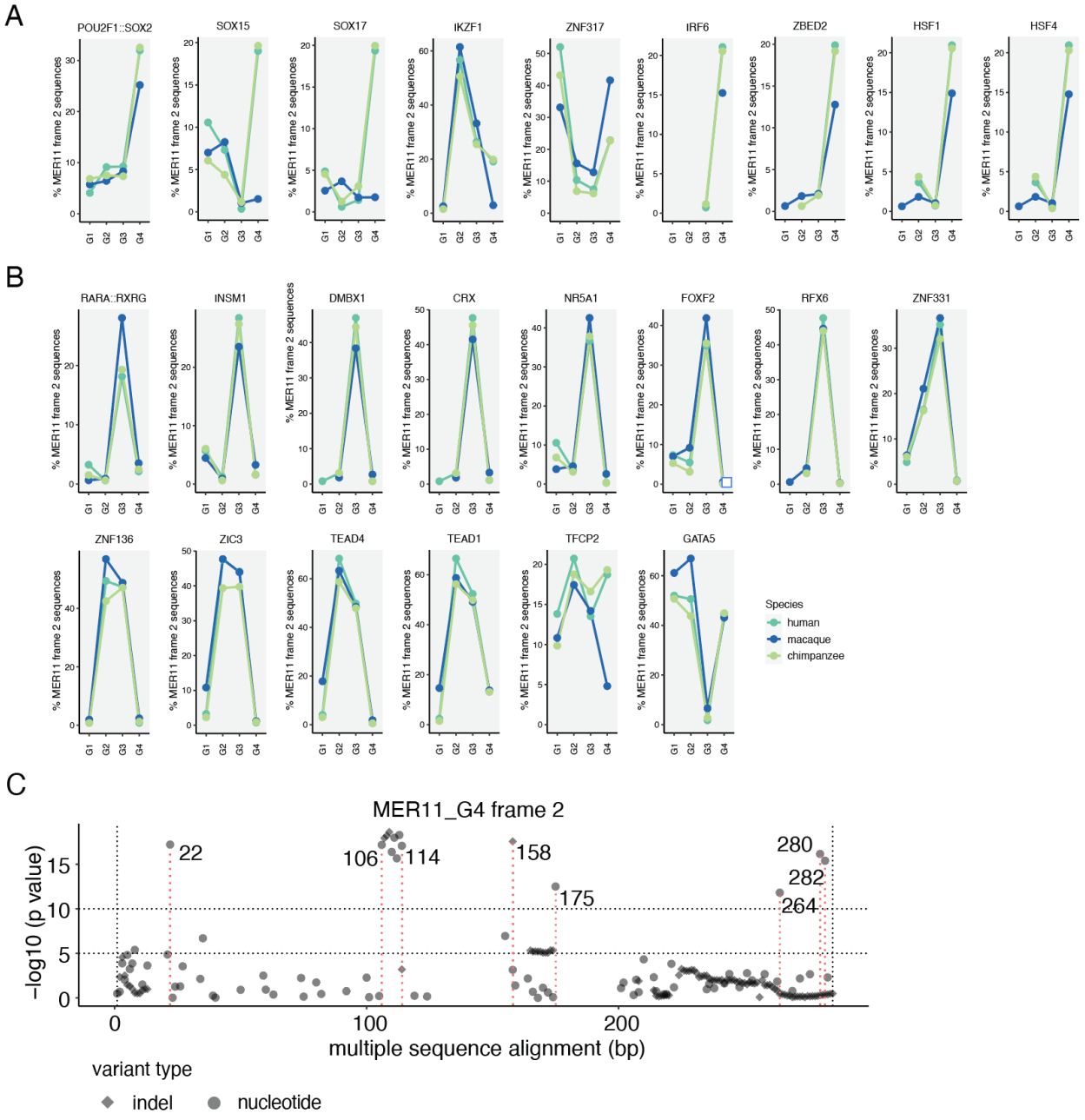
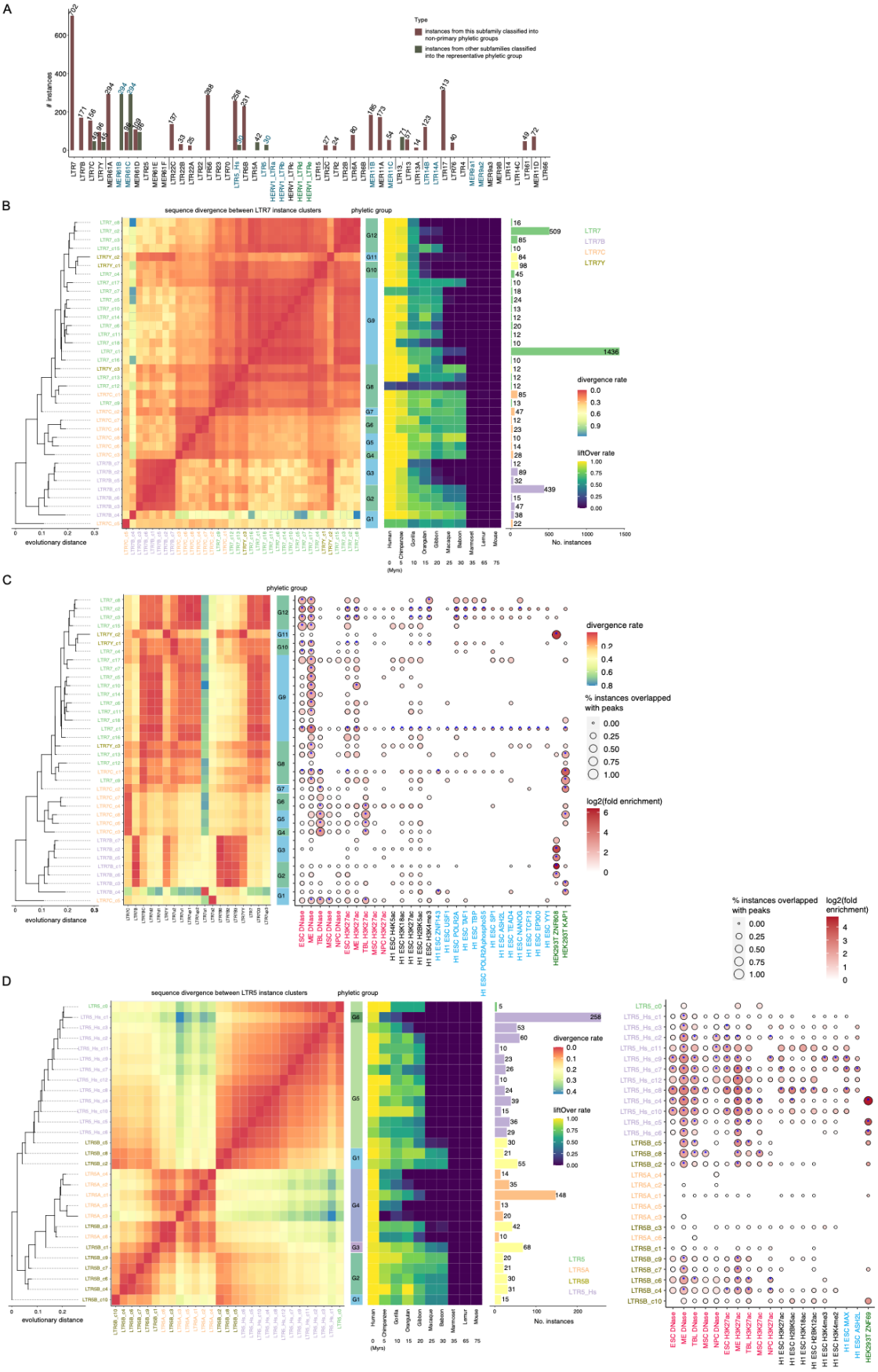


Figure S10. Nucleotides and motifs associated with the MPRA activity between species.

(A) Proportion of frame 2 sequences containing each candidate species-specific motif (e.g., SOX15) across human, chimpanzee, and macaque phyletic groups.

(B) Proportion of frame 2 sequences containing each associated and none species-specific motif across human, chimpanzee, and macaque phyletic groups.

(C) Nucleotides associated with the MPRA activity across all MER11_G4 frame 2 sequences. Human, chimpanzee, and macaque sequences were included to increase the detection power.


A
B
C
D
Type
instances from this subfamily classified into non-primary phyletic groups
instances from other subfamilies classified into the representative phyletic group
instances
sequence divergence between LTR7 instance clusters
phyletic group
evolutionary distance
No. instances
divergence rate
liftOver rate
% instances overlapped with peaks
log2(fold enrichment)
sequence divergence between LTR5 instance clusters
LTR7
LTR7B
LTR7C
LTR7Y
LTR5
LTR5A
LTR5B
LTR5_Hs

1

Figure S11. Phylogenetic and epigenetic analysis of LTR5 and LTR7 subfamily groups.

(A) Number of misannotated instances corrected by defined phyletic groups. Subfamilies are ordered by the number of phyletic groups per subfamily group. The phyletic group containing the most instances (top) is used to represent each subfamily. Subfamilies with the same representative phyletic groups are highlighted in blue or green. Number of misannotated instances counted in two subfamilies is highlighted in blue.

(B) LTR7 subfamily group rooted tree and the sequence similarity between clusters.

(C) Epigenetic profile of every LTR7 cluster and phyletic group and the sequence similarity with publicly available subfamily consensus sequences. Subfamily consensus sequences were downloaded from the DFAM database. Epigenetic marks overlapped with a minimum of five instances for small clusters (< 100 instances) and 20 instances for large clusters (≥ 100 instances) were kept. Permutation was used to compute the *p* values. Significantly enriched (log2((actual counts+1)/(mean shuffled counts + 1)) ≥ 1 and *p* value ≤ 0.05) clusters relative to 100 random genomic controls are highlighted.

(D) LTR5 subfamily group rooted tree and the sequence similarity and epigenetic profile of every cluster.

Supplementary Tables

Table S1. List of refined annotations of 53 simian-specific LTR subfamilies. Instances shorter than 200 bp were also annotated by *Blastn-short* against every cluster consensus sequence from each subfamily group. The top target cluster consensus sequence with a minimum of 50% alignment length versus the instance length was kept. Instances that failed to be annotated were classified as the unannotated group. We also re-annotated 2,926 instances from analyzed 53 subfamilies that were shorter than 200 bp through Blastn against each cluster consensus sequence and 2,055 of them were classified into different clusters and phyletic groups.

Table S2. Summary of MER11, MER34, and MER52 sequences submitted to MPRA.

Table S3. Characteristics of examined MER11, MER34, and MER52 frame sequences by MPRA. Subfamily consensus sequences, and positive and negative control sequences were also included. Computed alpha values, *z*-scaled alpha values and corresponding *p* values in iPSCs and NPCs were also included.

Table S4. List of phyletic groups and clusters amongst 53 simian-specific LTR subfamilies. The number of instances, branch length (bootstrap value) to the selected root, and the liftOver rate to the macaque genome per cluster were included.

Table S5. Epigenetic profile of every phyletic group amongst 53 simian-specific LTR subfamilies.

Table S6. List of primers used for lentiMPRA.