HUMAN GENETICS

A phylogenetic approach uncovers cryptic endogenous retrovirus subfamilies in the primate lineage

Xun Chen[1,2]*, Zicong Zhang[1], Yizhi Yan[1,3], Clement Goubert[3,4], Guillaume Bourque[1,3,5,6]*, Fumitaka Inoue[1]*

Current approaches for classifying and annotating endogenous retroviruses (ERVs) and their long terminal repeats (LTRs) have limited resolution and are inaccurate. Here, we developed an annotation approach based on phylogenetic analysis and cross-species conservation. Focusing on the evolutionarily young LTR subfamilies known as MER11A/B/C, we revealed the presence of four "new subfamilies," suggesting a new annotation for 412 (19.8%) of these repeat elements. We then validated their regulatory potential using a massively parallel reporter assay. We further identified motifs associated with their differential activities including an ape-specific gain of SOX-related motifs through a single-nucleotide deletion. By applying our approach across 53 simian-enriched LTR subfamilies, we defined 75 new subfamilies and found a novel annotation for a total of 3807 (30.0%) instances from 26 subfamilies. With this refined annotation of simian-enriched LTRs, it will be possible to better understand the evolution in primate genomes and potentially identify critical roles for ERVs and their LTRs in the 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 (*1*–*6*). In particular, at least 8% of the human genome finds its origin in endogenous retroviruses (ERVs) (*7*, *8*), which include subfamilies of ERV internal and long terminal repeat (LTR) sequences separately, while LTR sequences are often found in open chromatin and have the potential to act as genomic regulatory elements (*1*, *2*, *4*, *9*). ERVs originate from retrovirus infections, where the viral fragment interacts with transcription factors (TFs) in the host cell via its LTR sequences to express viral RNA, and then spreads throughout the genome by a copy-and-paste mechanism (*10*). ERVs including their internal and LTR sequences could spread in the host genome mainly through the vertical and horizontal transmission during evolution (*11*–*13*). 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 repressor genes, and H3K9me3 modification (*14*–*16*). Along with this active silencing, most TEs, including LTR sequences, accumulate mutations, which eventually result in their inactivation (*17*). Nevertheless, some LTR sequences may retain their regulatory activity or acquire beneficial mutations within TF-binding motifs (*6*). This can affect the regulatory activity of nearby gene expression and contribute to the adaptation of the LTR copies in the host genome (*6*, *18*, *19*). In addition, ERVs have a higher chance of survival in the next generation when they are expressed in germline or pluripotent cells during early embryonic development (*20*–*22*). Several LTR subfamilies contain binding sequences for pluripotency TFs such as POU5F1 and SOX2 (*6*, *23*).

Many ERVs integrated into the primate genomes after the divergence from other mammals (*24*). In the process, these relatively young ERV elements have contributed a substantial number of regulatory sequences to the human genome (*9*) and are associated with the evolution of TF-binding sites during primate evolution (*25*). Some subfamilies retain regulatory or transcriptional activity and have influenced human transcriptional networks (*1*, *2*, *4*, *26*–*30*). 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 (*30*). LTR5_Hs has also been shown, using a chimeric array of guide RNA oligos and CRISPR, to act as enhancers that regulate hundreds of human genes (*29*). Furthermore, the divergent expansion of LTR subfamilies within individual branches of the primate lineage, which provides a wealth of species-specific enhancers, substantially influences the regulatory network of these genomes during speciation (*31*).

The proper classification and annotation of LTR instances is critical to understanding their evolution, co-option and potential impact on the host (*32*). The standard approach, i.e., Repeatmasker, used to characterize a genomic region as a TE instance 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 (*33*). It is commonly used to identify and classify TEs in genomic sequences. However, it may not be an optimal approach to study TE co-option and accurately annotate TE sequences matched to multiple consensus sequences. 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 (*34*–*37*). Deep analysis of ERVs/LTRs across primates further emphasizes the complexity of their evolution and potential annotation problems (*38*–*40*). 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

[1]Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto 606-8501, Japan. [2]Shanghai Institute of Immunity and Infection, Chinese Academy of Sciences, Shanghai 200031, China. [3]Department of Human Genetics, McGill University, Montréal, QC H3A 0C7, Canada. [4]R. Ken Coit College of Pharmacy, University of Arizona, Tucson, AZ 85721, USA. [5]Victor Phillip Dahdaleh Institute of Genomic Medicine at McGill University, Montréal, QC H3A 0G1, Canada. [6]Canadian Center for Computational Genomics, McGill University, Montréal, QC H3A 0G1, Canada.
*Corresponding author. Email: xchen@siii.cas.cn (X.C.); guil.bourque@mcgill.ca (G.B.); inoue.fumitaka.7a@kyoto-u.ac.jp (F.I.)

viruses has led to various misclassifications (*41*). Furthermore, instances of an established ERV internal and LTR subfamily continuously evolve further into divergent sub-lineages, introducing another layer of complexity (*42*, *43*). Phylogenetic analyses have been used to study ERV/LTR sequence evolution with either a small set of full-length ERV or solo LTR sequences (*44*, *45*) or to interpret their regulatory heterogeneity (*46*). Unfortunately, most of these studies relied on the current ERV internal and LTR classification. More recently, phylo-regulatory approaches, combining phylogenetic reconstruction of LTR subfamilies and layering of epigenetic data, have helped overcome some of these issues for specific LTR subfamilies, e.g., LTR7 (*42*).

Considering the nature of LTRs, we hypothesized that the classification and annotation of LTR copies would require investigating their phylogenetic relationships to infer their evolution, function, and co-option. In this work, we present an improved annotation of simian-enriched LTR subfamilies 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 heterogeneous

We wanted to investigate the evolution of LTR subfamilies (i.e., subfamilies of LTR sequences but not ERV internal sequences) and observed that the distribution of the liftOver proportion of Lemur against human was closer to mouse than marmoset and other primate species (fig. S1A). Thus, we aimed to focus on the simian-enriched subfamilies and identified 179 subfamilies enriched in the human genome with copies in the marmoset or more closely related primate species but mostly absent from lemur (hence, putatively integrated or extensively expanded since the simian ancestor ~40 million years ago) (Fig. 1A). Evolutionary ages were computed on the basis of the substitution rates; thus, some of the simian-enriched subfamilies that were identified using the liftOver approach may still have high substitution rates. The selected subfamilies were consistently observed to be enriched in simian genomes by further examining a total of 47 primate species (see Materials and Methods, fig. S1B, and table S1). We focused on the LTR subfamilies 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 (fig. S1C). As expected, the use of recent genome builds may improve their locations in larger contigs but not notably change the presence/absence calls (fig. S1D and table S2). On the basis of a network analysis of repeat consensus sequence similarity (*47*), we also identified 26 closely related LTR subfamilies for a total of 61 putative simian-enriched subfamilies organized in 19 groups, which refer to the subfamilies that were either first integrated or widely spread in simian genomes (figs. S2 and S3). 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 (Fig. 1B and fig. S4A), 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 (*42*) and suggests that many simian-enriched LTR subfamilies are heterogeneous.

From this list, the MER11 subfamilies are of particular interest because they were among the youngest and had the lowest proportion of shared instances in macaque and displayed non-normal divergence rate distributions (Fig. 1B and figs. S1C and S4, A and B). To 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 (*42*), 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) (fig. S4C). 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 (Fig. 1C) (*48*). 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 nonreversible model without the use of an outgroup (see Materials and Methods) (*49*). After lifting over the instances from each cluster to other primate genomes, we selected the root that was most consistent with the proportions of shared instances across species (fig. S4D). This rooted tree was also found to be consistent with the network we constructed, further confirming the expansion progress of these repeat instances (fig. S4E). On the basis of the rooted tree and internal branch lengths between cluster consensus sequences, we defined four new subfamilies (see Materials and Methods and Fig. 1D), i.e., MER11_G1, MER11_G2, MER11_G3, and MER11_G4. As expected, instances in these new subfamilies displayed more homogenous divergence rates compared to the original subfamilies (fig. S3, B and F). Notably, some clusters of instances from the evolutionarily old MER11A were put in MER11_G1, while others were grouped with clusters from MER11B/C to form MER11_G2. Moreover, half of the 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 relatively 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 new subfamilies (Fig. 1E). Notably, if we were to select the top new subfamily to represent each original annotation, we found that a total of 412 (19.8%) MER11 instances would be annotated differently (Fig. 1F and table S3). On the basis of the new annotation, we further profiled their distribution on the human genome. As expected, instances from different new subfamilies were overall randomly distributed across each genome with some enrichment in specific chromosomes, e.g., MER11_G1 in chr4 and MER11_G3 in chrY (fig. S4F). Together, detailed sequence analysis revealed four new subfamilies with many MER11 instances that contrasted with the original subfamily classification.

MER11 new subfamilies display more consistent epigenetic profiles as compared to original MER11 subfamilies

Given that the new subfamilies of MER11 were more homogeneous on the basis of divergence rates (fig. S4G), we hypothesized that they

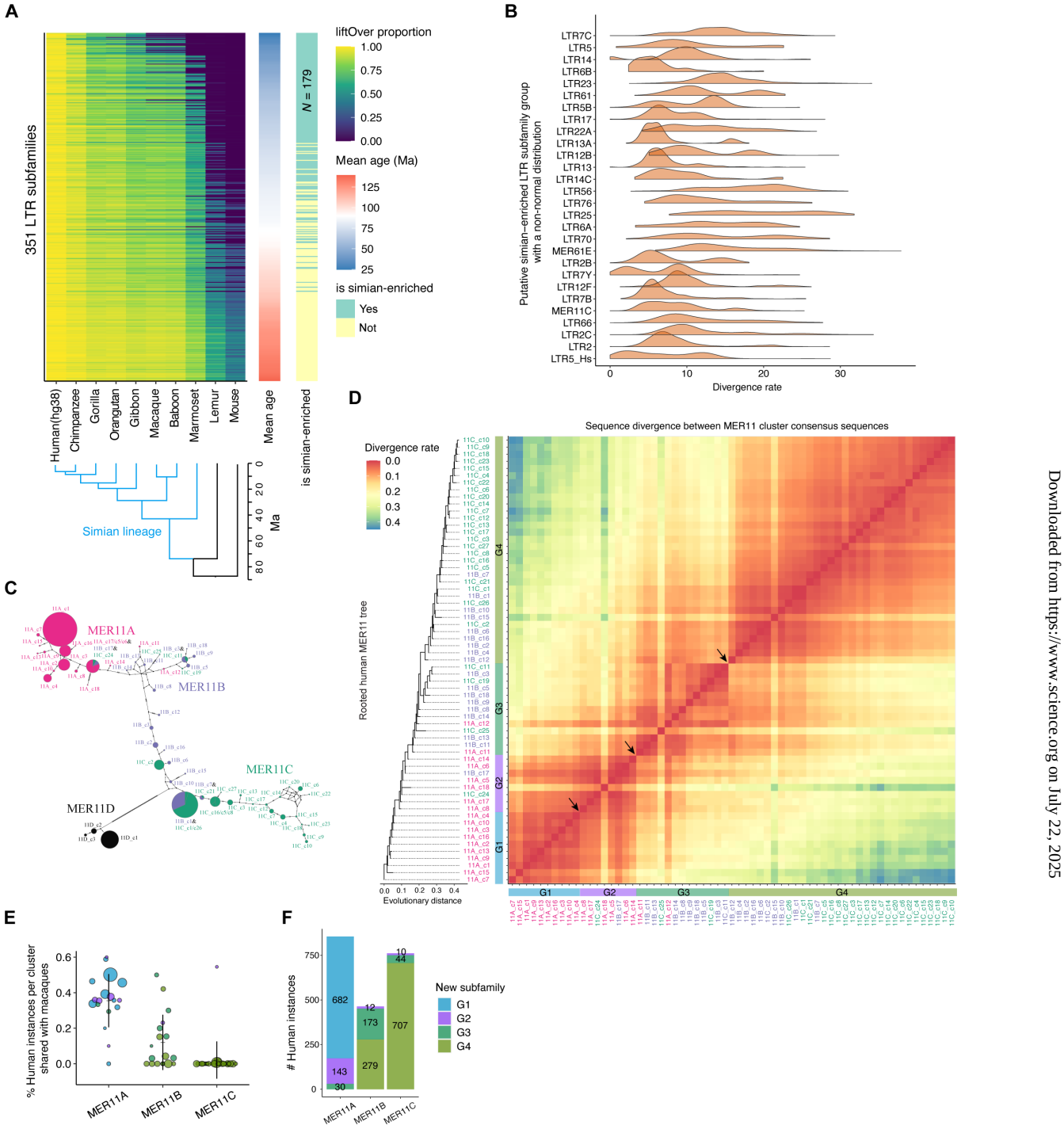

Fig. 1. The sequence heterogeneity of simian-enriched LTR subfamilies. (**A**) LiftOver analysis of LTR subfamilies (not including ERV internal sequences) from human (hg19) to other primate and mouse genomes (see Materials and Methods). Evolutionary ages were estimated on the basis of the divergence 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**) 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 original subfamilies are in different colors. Pie charts indicate clusters that have the same consensus sequences after the removal of gaps. (**D**) Rooted tree containing 63 MER11A/B/C clusters. Heatmap displays divergence rates between each pair of cluster consensus sequences. Four new subfamilies are identified using different colors (see Materials and Methods). Arrows indicate the high divergence rates between clusters from adjacent new subfamilies. Clusters derived from each original subfamily are in different colors. (**E**) Proportion of human instances shared with macaques. Dot size refers to the number of instances per cluster. (**F**) Number of instances per MER11A/B/C original subfamily assigned to each new subfamily.

would also have more consistent epigenetic profiles in human cells. As ERV expression and endogenization often occur in early developmental stages (*18*, *50*), we first compared chromatin accessibility and H3K27ac active histone mark in human embryonic stem cells (hESCs) and hESC-derived neural progenitor cells (NPCs) using two published datasets (*51*, *52*). From this, we identified 18 LTR subfamilies that were significantly enriched in hESCs (fig. S5, A to C). In particular, we found that MER11B and MER11C subfamilies showed relatively high cell-type specificity in hESCs and mesendoderm (ME) cells (fig. S5D). 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 TF chromatin immunoprecipitation sequencing (ChIP-seq) datasets from the ENCODE project (Fig. 2A). Notably, we observed that specific clusters within each subfamily tree were enriched for specific histone marks and TF peaks. Specifically, the younger MER11A clusters, with a long evolutionary distance to the root, were enriched for active histone marks and TF peaks; the more ancient clusters were enriched for active histone marks and TF peaks in MER11B and MER11C; in contrast, the youngest MER11C clusters were enriched for active histone marks but less enriched for TF peaks. Thus, each original subfamily had a high epigenetic heterogeneity and a distinct enrichment pattern.

To inspect whether our four new subfamilies of MER11 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 new subfamilies (Fig. 2B). For instance, MER11_G1 lacked TF peaks and active histone marks; MER11_G3 was 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. It has been hypothesized that the canonical function of KRAB-ZNFs and KAP1 is for the transcriptional repression of TEs (*53*). Because of the balance between chromatin accessibility and repression over TEs (*54*), we also looked at the enrichment of KRAB-ZNFs and KAP1 (fig. S5, E and F) binding in human embryonic kidney (HEK) 293T cells. We further observed the sequential loss of ZNF440, ZNF433, ZNF468, ZNF611, ZNF33A, and ZNF808 binding and the gain of ZNF525 binding along the evolution of the MER11 clusters (Fig. 2B). KAP1 binding was mostly enriched in MER11_G3 and relatively old MER11_G4 clusters, which was consistent with the enrichment of ZNF808 binding. Last, compared to the original annotation of the MER11 subfamilies, we found that the four new subfamilies achieved a higher specificity across these active marks (Fig. 2C). For example, TEAD4 enrichment was 29.6% using new subfamilies (it overlapped 29.6% of instances in MER11_G3 but 0% of instances in MER11_G1), which was much higher than original subfamily annotations (the enrichment was 8% since it overlapped 11% of MER11B instances but 3% instances in MER11A). We further compared the new and original subfamilies with the highest percentage of peak-associated instances and found that the proportion of new subfamilies was notably higher than that of the original ones (fig. S5, G and H). Together, the four new subfamilies of MER11 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 new subfamilies and reveals associated TF-binding motifs

To further assess the biological relevance of the reconstituted MER11 subfamily annotations, we leveraged a massively parallel reporter assay (MPRA). MPRA is a technique that allows us to measure the regulatory activity of thousands of DNA elements in parallel. Briefly, a regulatory element library is first cloned into a reporter vector with a unique barcode and transduced into cells via lentivirus. The regulatory activity of each element in the cells is quantified by measuring the level of transcribed barcodes using high-throughput sequencing (*51*). In this study, we first identified two peaks of accessible regions within the MER11A/B/C instances and extracted their sequences [~250 base pair (bp)] as the frames—putative functional sequences of a suitable length—to be analyzed by MPRA (Fig. 3A and fig. S6, A and B). As controls, we analyzed two older LTR subfamilies, MER34 and MER52 (Fig. 3B); MER34A1/C_ subfamilies were enriched for both Assay for transposase-accessible chromatin using sequencing (ATAC-seq) and H3K27ac peaks in hESCs compared to NPCs, while the enrichment of MER52C subfamily was slightly increased during the NPC differentiation (fig. S5, B and C). 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 (fig. S6C). Some sequences had to be excluded because of the high number of mutations and truncations relative to the core frames used for the MPRA experiment (Materials and Methods). In the end, we analyzed 16,929 unique LTR sequences, including 6912 MER11s, 5751 MER34s, and 4266 MER52s sequences, together with 100 positive control sequences that were previously reported to play a regulatory role in hESCs and 100 negative control sequences with randomized nucleotides as we also described in Materials and Methods (Fig. 3C and table S3 and S4) (*51*), 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 to 75% of MER34; fig. S6D). The 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 (fig. S6E).

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 (Fig. 3D, fig. S6F, and table S5). 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 new subfamilies. Notably, although the overall activity of frame 1 remained low, the activity levels were increased for the more recent MER11_G4 clusters (fig. S6G). For frame 2, the activity varied between new subfamilies, specifically some clusters from MER11_G2/3 and the young clusters from MER11_G4, had the highest activity levels (Fig. 3E and tables S3 and S5). Most instances from the oldest MER11_G1 new subfamily could not be analyzed by MPRA because of 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 (*55*). Among all

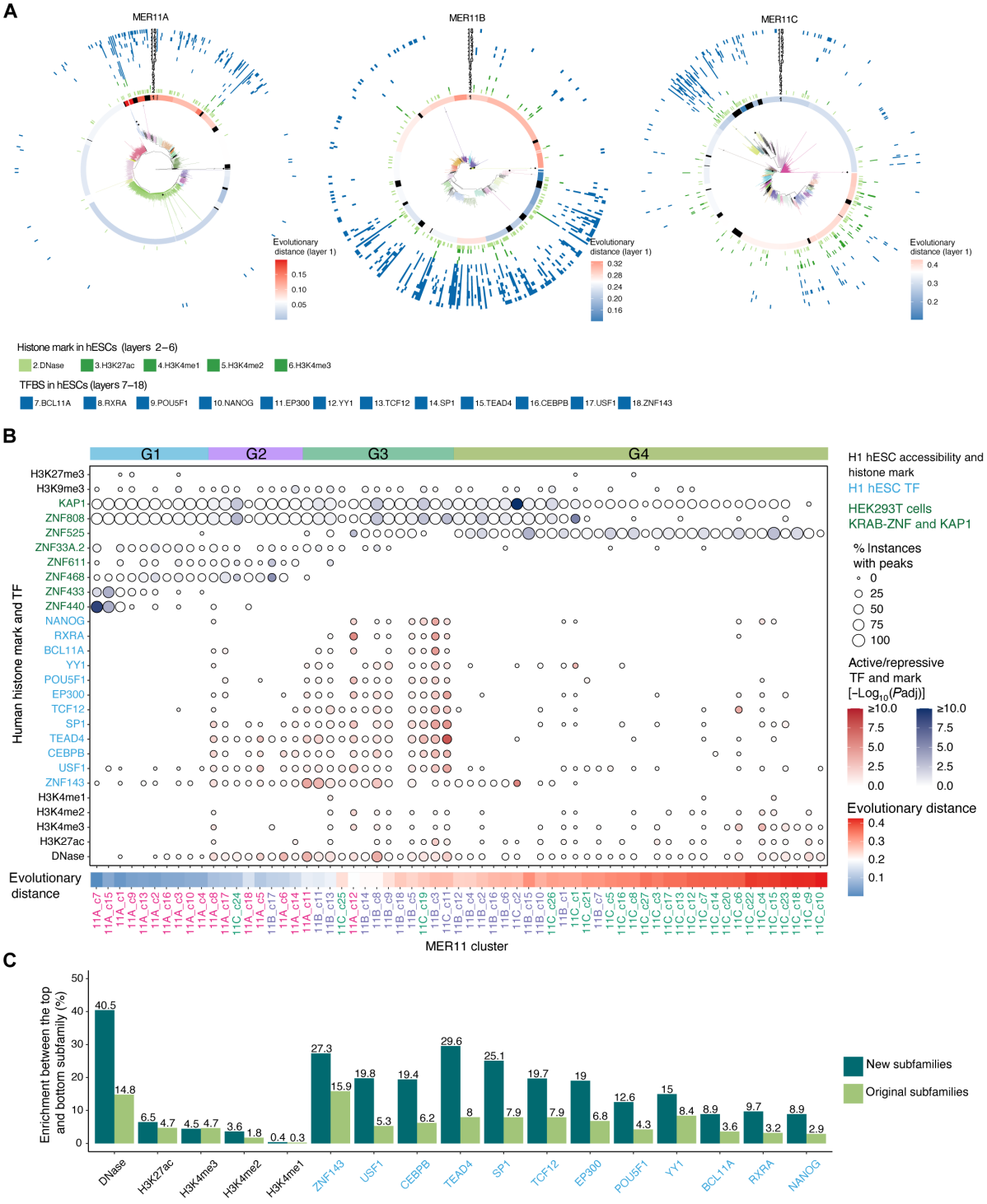

Fig. 2. MER11 new subfamilies display more consistent epigenetic profiles as compared to original annotations. (**A**) Circular plots of MER11A/B and C. Unrooted trees are used to order instances (center). Evolutionary distance per cluster was calculated using the root from the tree of Fig. 1D. Active histone marks and significantly enriched TFBSs in any MER11 original subfamilies are shown (see Materials and Methods). (**B**) Enrichment of active histone marks and TF peaks for every MER11 cluster ordered on the basis of the evolutionary distance from the root. Active and repressive histone marks and significantly enriched TFs and KRAB-ZNFs in any MER11 original subfamily are shown. New subfamilies are highlighted on the top and separated by dotted lines. Clusters derived from each original subfamily are colored differently. (**C**) Enrichment of active histone marks and TF peaks in new versus original subfamilies. Enrichment was computed as the proportion of peaks-associated instances in the top new subfamily, the one with the highest proportion of peak-associated instances, minus the proportion in the bottom new subfamily, the one with the lowest proportion, for each histone mark and TF (in blue). The same was calculated between the top original subfamily and the bottom original subfamily.

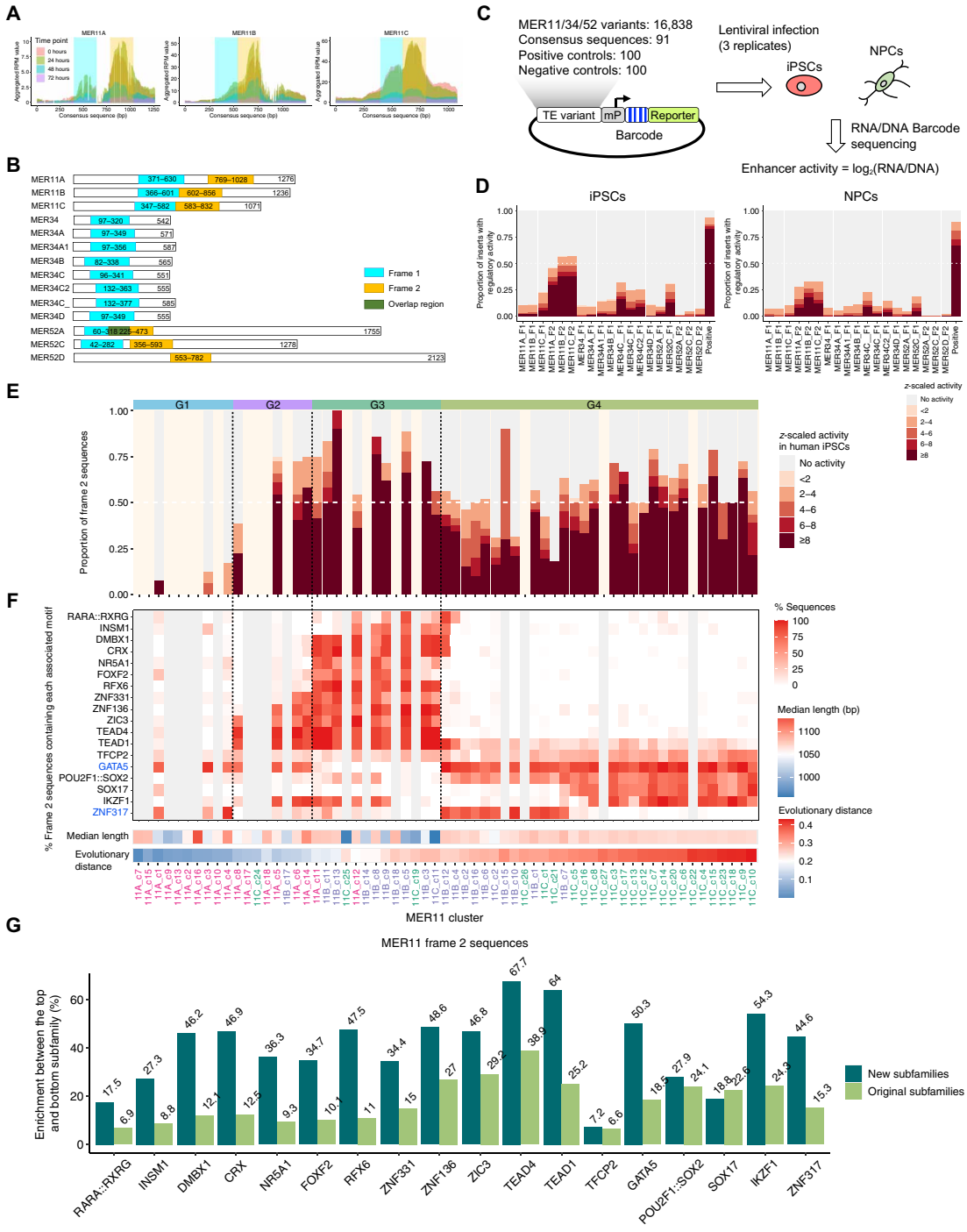

Fig. 3. MPRA and MER11 new annotations help resolve the functional heterogeneity of MER11 subfamilies. (**A**) Determination of accessible regions along the consensus sequence per subfamily. ATAC-seq RPM (reads per million) values are aggregated across instances. (**B**) MER11, MER34, and MER52 consensus 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 Materials and 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 new subfamilies. Motifs that are negatively associated with the MPRA activity (fig. S7) are highlighted in blue. Chimpanzee and macaque sequences are also included. Clusters derived from each original subfamily are colored differently. (**G**) Motif enrichment in new versus original subfamilies. Enrichment was computed as the proportion of instances containing each motif in the top new subfamily, the one with the highest proportion, minus the bottom new subfamily, the one with the lowest proportion. Motif enrichment was computed similarly between the top and the bottom original subfamily.

frame 1 genomic sequences, we identified SP3 and related motifs (e.g., KLF12) followed by ZICs, POU::SOXs, and SOXs motifs (fig. S7). We then looked at the MER11 frame 2 sequences and identified many motifs that were significantly positively or negatively associated with the MPRA activity including SOXs, POU2F1::SOX2, PITXs, ZICs, and TEADs, suggesting their enhancer or repressor roles in hESCs. The motifs found to be enriched in MER34 (e.g., SPs and POU::SOXs) and MER52 (e.g., SPs and 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 (fig. S8). For each motif group, we 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 (Fig. 3F). Notably, we observed a clear clustering of most motifs among the different new subfamilies. 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.

Last, we investigated whether these TF motifs overlapped with nucleotides associated with the MPRA activity based on TE-wide nucleotide association analysis (*56*). Focusing on MER11 frame 2, we identified 15 single-nucleotide variants and 32 indels that were significantly associated with the MPRA activity ($P \leq 1 \times 10^{-10}$) (fig. S9, A and B). 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 (fig. S9C). 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 among MER11 new subfamilies relative to the originally assigned subfamilies (Fig. 3G and fig. S6H). For example, the highest proportion of frame 2 sequences containing TEAD1 was 66.5% among new subfamilies and 40.5% among original subfamily annotations, while the lowest proportion was 2.4% among new subfamilies and 15.3% among original annotations. Together, we concluded that the four new subfamilies of MER11 were distinguishable on the basis of distinct sets of motifs associated with MPRA activity.

The human MER11 new subfamilies are conserved in the primate lineage but spread independently

To further characterize the evolution of the new subfamilies of MER11, 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 (Fig. 4A). We observed a similar pattern by the alignments of human MER11/34/52 sequences to macaque and vice versa (fig. S10A). In contrast, only 7.4% of the chimpanzee MER11A instances are absent in the human genome (fig. S10B). As expected, the LTR sequences in panTro4 and macFas5 genome builds are conserved among other genome builds (fig. S1D; tables S2 and S6; and fig. S10, C and D). These species-specific MER11A instances may regulate different sets of genes among primate species (table S7 and fig. S10E). Next, using the approach described above (fig. S3), we built the unrooted tree among MER11A macaque instances and identified 33 clusters (11A_m_c1 to c33; Fig. 4B and fig. S10F). 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 (Fig. 4C, left) and validated that evolutionarily old clusters had among the highest proportions of instances shared with humans. We found that the MER11 rooted trees had a high consistency between two lineages (Fig. 4C, right). Notably, on the basis of the sequence similarity patterns and the rooted trees, we could also mostly recapture the four new subfamilies we had defined on the basis of the human instances. We further compared the features of each new subfamily 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 (Fig. 4D and fig. S10, G and H). Although MER11_G4 was the least conserved, we also observed that most instances were in the oldest and youngest groups in both species (Fig. 4E). Last, we compared the MPRA activities of the frame 1/2 sequences obtained from the two species (Fig. 4F). Overall, the activities of each new subfamily were comparable between two species, with macaque having relatively lower activities (4.4 to 15.0% for highly active MER11_G2/G3/G4). Together, we observed a high conservation in both lineages of the new subfamilies with respect to sequence and MPRA activity. This is especially notable in the youngest new subfamily (G4) despite an independent spread following the species divergence.

The gain of SOX-related motifs within a new subfamily 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 first focused on all MER11_G4 frame 2 sequences in humans, chimpanzees, and macaques. Applying the approach described previously, we identified SOXs, POU2F1::SOX2, HSF1, ZBED2, ZNF317, and IKZF1 to be the most associated motifs across three species (Fig. 5A). We then examined the association between the motif combination and MPRA activities among MER11_G4 sequences (Fig. 5B). We found that MER11_G4 frame sequences in three species containing either or both POU2F1::SOX2 and SOX-related motifs (SOX15 as a representative) had the highest activity levels compared to others.

Next, we inspected whether the gain of POU::SOX and SOX-related 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 (Fig. 5C); 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 (Fig. 5C). We further analyzed the prevalence of the POU2F1::SOX2 and SOX15/17 motifs among the four new subfamilies' frame 2 sequences across species. The proportion of sequences containing POU2F1::SOX2 remained below 10% in MER11_G1/2/3 and was substantially increased in MER11_G4 sequences, which was consistent across the three species. In contrast, the proportions of SOX15/17 were substantially increased in humans and chimpanzees but remained low in macaque MER11_G4 sequences (fig. S11A). Only a few other

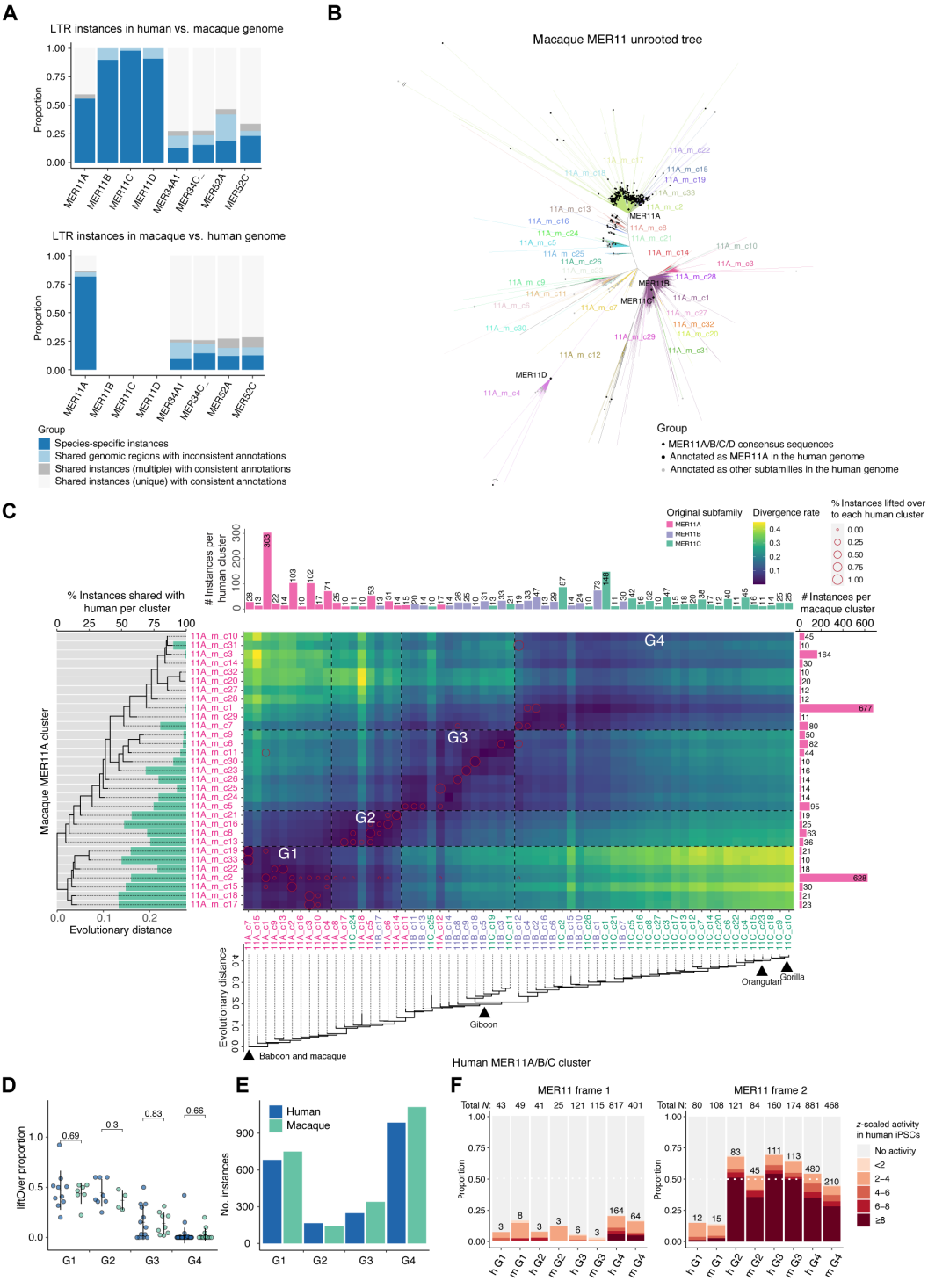

Fig. 4. The presence of MER11 new subfamilies in both human and macaque lineages. (**A**) Conservation and original annotation of MER11 instances in human versus macaque based on liftOver and RepeatMasker. Instances intersected with regions annotated as the same or different subfamilies in another species 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 (fig. S10F) and removed from further analyses. (**C**) The comparison of MER11 clusters and new subfamilies between humans and macaques. Clusters are arranged by the macaque and human cluster consensus sequence rooted trees. Clusters from every original subfamily are colored differently. The primate lineage when they first integrated is highlighted. Human and macaque new subfamilies are separated by dotted lines. (**D**) Comparison of the liftOver proportion per new subfamily between human and macaque. (**E**) Number of instances per new subfamily between human and macaque. (**F**) Frame1 and frame 2 MPRA activity per new subfamily between human (h) and macaque (m).

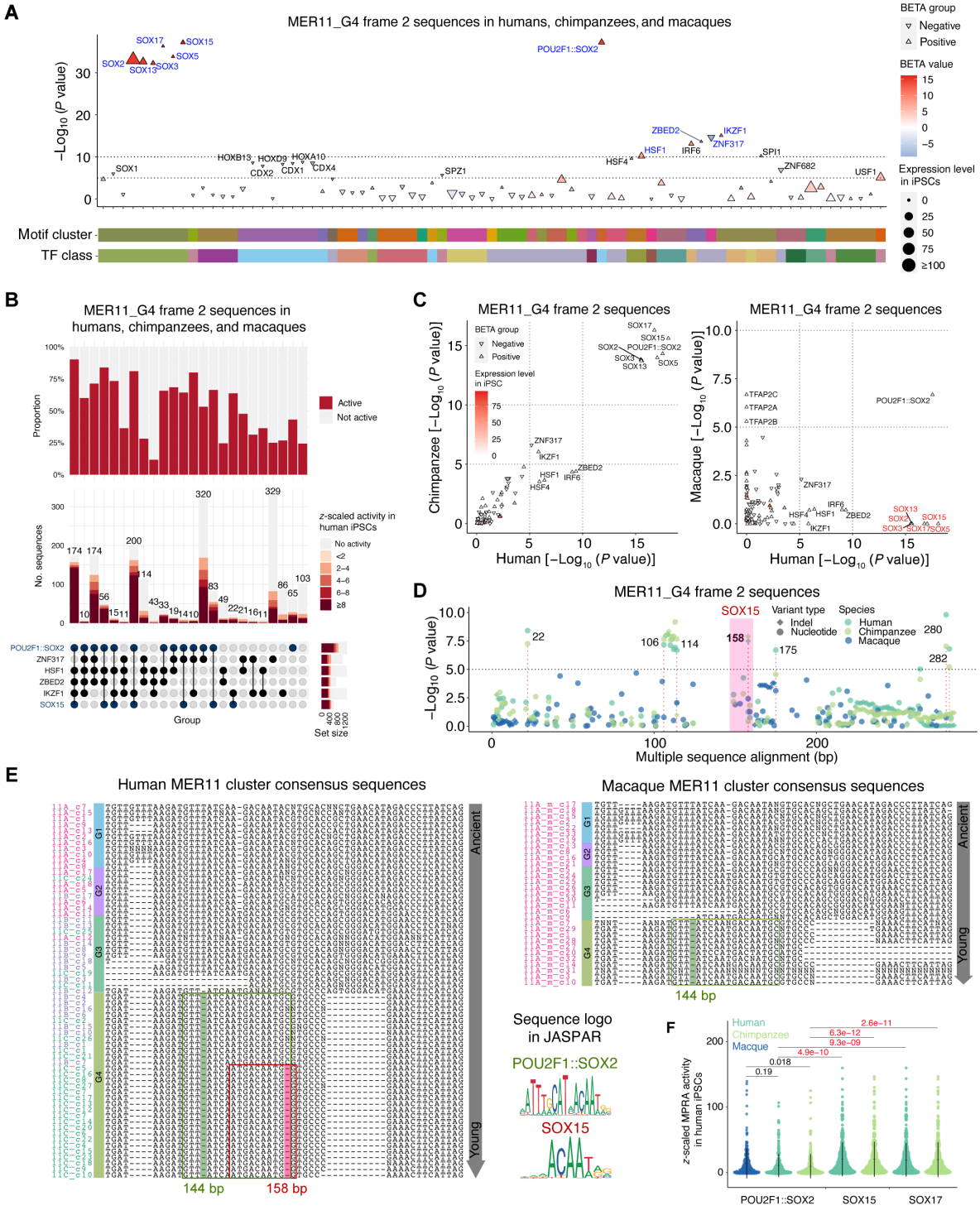

Fig. 5. Nucleotide changes and gain of functional motifs during separate expansions of MER11_G4 in primate lineages. (**A**) Motifs contributing to the activity of MER11_G4 frame 2 sequences from all the three species. *P* values and effect sizes (BETA values) were computed by the linear regression model 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 separately. *P* values were computed by the linear regression model using Plink2 and then compared between human and chimpanzee and between human and macaque sequences. (**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 originally annotated MER11 subfamily are colored differently. New subfamilies 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.

enriched motifs showed a divergence between species with most (e.g., INSM1, DMBX1, and CRX) being consistently enriched (fig. S11, A and B). MER11_G2–specific nucleotide association further revealed that many human- and chimpanzee-specific variations in the multiple sequence alignment were significantly associated with the frame 2 MPRA activity (Fig. 5D and fig. S11C). For instance, nucleotide variations at positions 106 and 114 overlapped with ZNF317, which exhibited a negative correlation with MPRA activity. The nucleotide variation at position 175 overlapped with ZBED2, HSF1, and IRF6, showing a weak association with the MPRA activity. Only the single-nucleotide deletion at position 158 was located within the uniquely enriched region of SOX-related motifs in the human frame 2 sequences, which exhibited a strong association with MPRA activity (fig. S9C).

Last, we inspected the POU::SOX and SOX-related motifs across the reconstructed frame 2 cluster consensus sequences at the target region from all new subfamilies (Fig. 5E). Human and macaque cluster consensus sequences were compared. We observed that a single-nucleotide deletion at position 144 occurred between MER11_G3 and MER11_G4 leading to the gain of a POU2F1::SOX2-like motif; 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 (Fig. 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. 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 increased regulatory potential of the instances in this evolutionarily young new subfamily.

Cryptic ERV subfamilies in the primate lineage with distinct epigenetic profiles

Having shown the usefulness of defining new subfamilies of MER11-type LTRs to understand the evolution of this family in primate lineages, we wanted to apply the same approach to examine other simian-enriched LTR subfamilies (Fig. 1B and figs. S3 and S4A). 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 (Materials and Methods). In this way, we identified 75 new subfamilies from 18 subfamily groups (containing 53 original LTR subfamilies) (Fig. 6A and table S8). Among them, the LTR7 subfamily group had the most ($N = 12$) new subfamilies and LTR66 remained as a single subfamily. In total, 26 of the individual LTR subfamilies could be subdivided into multiple new subfamilies with a maximum of 7 for LTR7C, supporting their high sequence heterogeneity (Fig. 6B). For each LTR subfamily, we selected the new subfamily with the most instances to be the representative for that subfamily. With this, a total of 3807 (30.0%) instances from these 26 LTR subfamilies were classified into a different new subfamily (fig. S12A). For instance, a total of 258 LTR5_Hs instances (42.7%) were now classified in a non-primary new subfamily.

To validate the approach, we reanalyzed the LTR7 subfamily, which was carefully studied through phylo-regulatory analysis (*42*). Here, we also included other related subfamilies from the LTR7 subfamily group (i.e., LTR7B, LTR7A, and LTR7Y) to depict their full evolutionary history (fig. S12B). As expected, we observed a high consistency between the 12 new subfamilies we identified and the sequence clusters previously reported (fig. S12C) (*42*). Moreover, by looking at the epigenetic profiles across different cell types, we found that the evolutionarily young LTR7_G12 new subfamily, similar to previously reported LTR7up1/up2/up3 (*42*), were more active and enriched for several TFBSs as compared to other LTR7 new subfamilies in ESCs (fig. S12C). In addition, LTR7_G4/G5 were found to be enriched for accessibility and H3K27ac peaks in trophoblast (TBL) cells compared to other cell types, while LTR7_G9 was enriched for accessibility for ME 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 new subfamilies have distinct cell–specific epigenetic profiles.

Last, we explored the epigenetic properties of the 75 newly annotated LTR subfamilies (Fig. 6C and table S9). As expected, we found that new subfamilies of many subfamily groups were significantly active during the differentiation from hESC to NPCs, such as MER11s, LTR5s, LTR6s, LTR7s, LTR13s, LTR22s, LTR25s, and LTR61s. LTR new subfamilies were also active in a cell-specific manner, such as LTR6_G4 in mesendoderm (ME) cells and LTR14_G1 in TBL cells. 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 new subfamilies. We identified that some TFBSs were enriched in specific new subfamilies per family group. For instance, MER9_G2/G3 were enriched for ZNF143 and LTR13_G9/G12 were enriched for RAD21, CTCF, and ZNF143. RAD21 and CTCF are key factors involved in chromatin looping and architecture (*57*), and ZNF143 also plays a crucial role in chromatin looping and gene regulation (*58*). Thus, they may have a notable impact on the chromatin organization and regulatory networks. We also observed a distinct enrichment of KRAB-ZNF– and KAP1-binding sites in HEK293T cells between new subfamilies within certain subfamily groups, such as ZNF808 in LTR7_G2/G3/G11, ZNF69 in LTR5_G1/G5, ZNF28 in MER9_G2, ZNF816 in MER9_G3, and multiple ZNFs in LTR13 and MER11 new subfamilies. Focusing on the six LTR5 new subfamilies revealed that LTR5_G5 had among the highest epigenetic enrichment (fig. S12D). 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 among new subfamilies relative to the subfamilies, such as LTR6 and LTR22 subfamily groups (Fig. 6D). Together, the newly annotated simian-enriched LTR subfamilies were enriched for different sets of active histone marks, TFs, and KRAB-ZNFs, suggesting differential evolutionary trajectories.

DISCUSSION

The MER11 family of ERVs was previously analyzed for its phylogeny and TF-binding sites and motifs (*45*, *46*, *59*, *60*). Consistently, MER11A was found to be more ancient compared to MER11B based on the LTR sequences of 78 human HML8 proviral sequences, and MER11C is the youngest (*45*). They also observed that MER11A/B/C may not be monophyletic groups. Recently, Kosuge *et al.* (*61*) also performed the phylogenetic analysis of MER11A/B/C/D family and obtained a total of six new subfamilies, probably due to the use of different thresholds in this study. Overall, their results were mostly consistent with ours, including the enrichment of KRAB-ZNFs in these new subfamilies. However, these studies lacked the assays

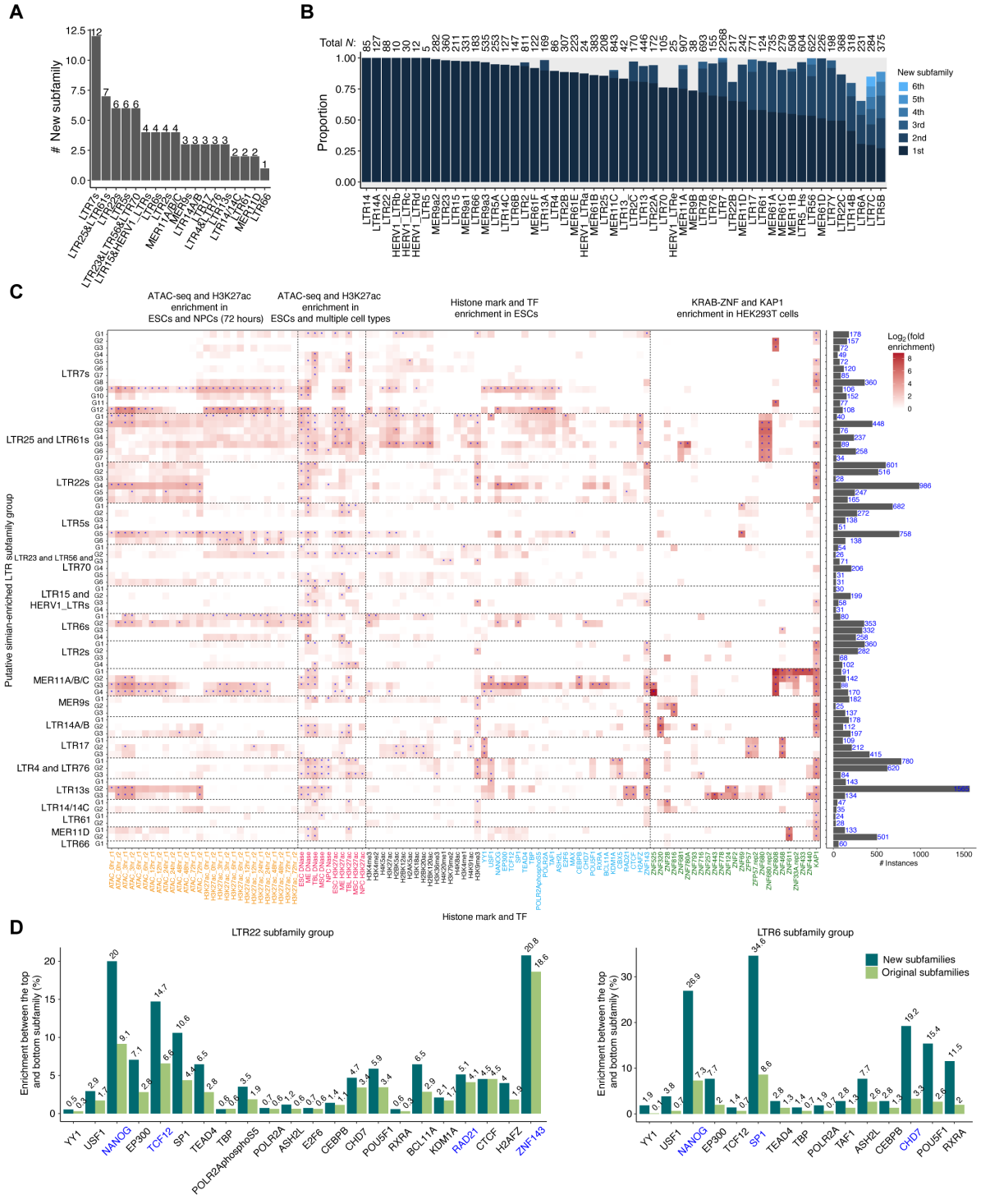

Fig. 6. Cryptic ERV subfamilies in the primate lineage with distinct epigenetic profiles. (**A**) Number of newly annotated subfamilies per subfamily group. (**B**) Proportion of instances per original subfamily classified into the top (representative) new subfamily. The top new subfamily contains the most instances per original subfamily, second new annotation contains the second most instances, and so on. (**C**) Epigenetic profiles across new subfamilies of 18 simian-enriched 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} new subfamilies relative to 100 random genomic controls are highlighted. (**D**) TF specificity of peaks-associated instances between new and original subfamilies in LTR22 and LTR6 subfamily groups. Enrichment was computed as the proportion of peaks-associated instances in the top new subfamily, the one with the highest proportion of peak-associated instances, minus the proportion in the bottom new subfamily, the one with the lowest proportion, for each TF. The same was calculated between the top original subfamily and the bottom original subfamily. Blue color indicates TFs significantly enriched in any new subfamily.

demonstrating the regulatory activity of distinct genomic variants, nor elucidate the evolutionary process of their regulatory function. Here, on the basis of 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 subfamilies. We found such an annotation critical to understanding the expansion process and diverse co-option strategies (e.g., coding and noncoding transcripts, cis-regulatory elements, and 3D chromatin organization) of MER11 across diverse primate genomes (*5*, *40*). Moreover, such a phylo-regulatory approach allowed us to detect stronger and missed associations between MER11 subfamilies 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 (*56*); however, none of these studies focused on refining the annotation of the instances themselves.

Among the four new subfamilies of MER11, we found that the intermediate-aged MER11_G2/G3 but not MER11_G1 (oldest) contained multiple TF motifs, such as ZIC and TEAD (fig. S9C). These TF motifs were conserved between human and macaque (fig. S11B), 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 (Fig. 5E). This suggests an increase in binding affinity for the SOX proteins, supported by the observed increase in the MPRA activity (Fig. 5F), although further investigation is necessary to validate this observation. This deletion emerged after the divergence between humans and macaques, potentially contributing to the human-specific expansion of the younger MER11_G4. SOX15 and SOX17 are known to play crucial roles in human primordial germ cell differentiation and tissue-specific gene regulation (*62*, *63*). Thus, these ape-specific SOX motifs in MER11 subfamilies may influence the gene regulatory network during development in a lineage-specific manner. These results demonstrate that using the reconstituted MER11 subfamilies combined with MPRA can shed light into the process of LTR expansion and divergence at the single-nucleotide level. The regulatory function of the MER11 enhancer in relation to the host genome awaits validation in future studies.

When we examined the presence of ZIC, TEAD, CTCF, and SOX motifs in the LTR sequences of the 64 retroviral genomes from the ICTV database, we found a total of 6, 13, 26, and 23 viral genomes with these motifs, separately. On the other hand, Ito *et al.* (*46*) previously reported the presence of SOX- and CTCF-binding sites in a certain subfamily. There is no clear enrichment pattern relevant to the evolutionary age. However, they found the enrichment of POU5F1, SOX2, and some other TF bindings in the young LTR7 and LTR5 families, supporting the gain of many critical TF-binding sites during evolution.

From another perspective, sequence variations can arise either before or after the loss of transposition capability. Variations that occur before the loss of transposition capability may result from selection acting on their transposition potential, e.g., more efficient transcription or escape from the host silencing machinery. In contrast, variations that occur after the loss of transposition capability reflect selection acting on the host genome, e.g., enhancer potential of TE-derived elements. Such variations can be species-specific, depending on TF variations and TE landing sites within a genome. It is important to acknowledge that even with our classification, it remains challenging to distinguish between these two scenarios.

The enhancer function of LTRs is influenced not only by transcription activators but also by the derepression of KRAB zinc-finger transcription repressors that play a role in the defense mechanism of the host genome (*64*, *65*). In our study, we observed that various KRAB repressor–binding sequences were differently enriched in the new subfamilies of MER11 (Fig. 2B). This suggests that classification and annotation with phylogenetic relationships could also be useful to understanding the arms race between transcriptional activation and repressive mechanisms. Previous research reported ZNF808 binding to MER11 during pancreatic development (*60*). 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, 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 derepression. 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.

The LiftOver method and chain files are commonly used in analyzing TEs in primate genomes (*42*, *56*, *66*, *67*), the latest assemblies and whole-genome alignments could be further used to improve the accuracy in near future. Although our proposed method does not allow distinguishing between the radiation and co-option of ERVs, our approach with MER11 highlights the importance of the phylogenetic classification and annotation of LTRs to better understand their sequence and functional evolution, co-option, and arms race with the host. Unfortunately, the current annotation of LTR subfamilies groups together heterogeneous sets of sequences thus is not ideal for these purposes (Fig. 1B). Moreover, a subset of LTR instances from a subfamily with emerged TF-binding sites through mutations may be grouped into a different subfamily. We also observed that macaque MER11B/C/D instances were all mis-annotated as MER11A (Fig. 4, A and B), which suggests that nonhuman 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-enriched LTR subfamilies and observed that nearly half (26 of 53) encompasses new annotations. This analysis suggests a new annotation for 30% of the genomic instances from these 26 simian-enriched LTR subfamilies (table S3). 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.

Last, the classification and annotation of TEs across species has been a challenging problem because of the lack of ground truth (*34*). 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.

MATERIALS AND 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-enriched 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. 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.

We then explored the orthologous of LTR sequences across more primate genomes. Briefly, 77 assemblies of a total of 47 primate species were downloaded from various databases (table S1). After extracting the human (hg19) LTR sequences together with ±1-kb adjacent sequences using Bedtools slop and getfasta functions (*68*), we aligned them against each of the primate genomes using minimap2 with parameters "-ax map-pb" (*69*). We then kept top alignments with MAPQ ≥5 and mappable length of ≥30% LTR sequences and mappable adjacent sequences (either upstream or downstream) ≥300 bp.

Evolutionary ages were computed on the basis of the divergence rates as we previously described (*70*, *71*). Briefly, the divergence rate of each LTR instance relative the corresponding consensus sequence was computed by RepeatMasker. 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 its evolutionary age in million years.

LTR consensus sequence similarity and network analysis

We first retrieved the TE consensus sequences in FASTA format from the Repbase database (*72*), 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 among themselves using Blastn (BLAST 2.13.0+) (*73*) 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 (*47*, *74*). 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-enriched 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) (*75*) with the parameters "--localpair --maxiterate 1000." The sequence alignment was further refined using PRANK (v170427) (*76*) with the parameters "-showanc -njtree -uselogs -prunetree -F –showevents." We then used trimAL (v1.4.1) (*77*) 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) (*78*) 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) (*79*) 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 recalculated the divergence rates of every instance versus their corresponding original and new subfamily 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 new subfamily. 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 new subfamilies 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 on the basis of

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 among 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 new subfamilies 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 new subfamily, which has among the top branch lengths to others. We also confirmed the new subfamilies by examining the pair-wide divergence rates to look at extreme values between every adjacent clusters. We lastly kept the new annotations 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 new subfamilies for the originally annotated macaque MER11A instances, while two new subfamilies 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-enriched LTR subfamily groups phylogenetic analysis

We performed a similar analysis for other simian-enriched LTR subfamily groups. Briefly, we first constructed the unrooted trees of instances (≥200 bp) for every subfamily from a group. After identifying the clusters per subfamily, we constructed the rooted trees and selected the best one to determine new subfamilies for each subfamily group. Using the same approach (fig. S3), we kept new subfamilies with a minimum of 25 instances except the 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 database (GSE115046) (*51*). 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 (*71*). 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 (*68*) intersect function with the parameters "intersect -wa -u –a." After that, we counted the number of instances that were associated with peaks per new subfamily. 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 (*52*).

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 reanalyzed 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 (*71*). We then kept subfamilies with a minimum of twofold enrichment of chromatin accessibility in ESCs compared to NPCs. We further validated the results in other differentiated cells including ME cells (E004), TBL-like cells (E005), mesenchymal stem cells (E006), and NPCs (E007) (*52*). 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://ncbi.nlm.nih.gov/geo/roadmap/epigenomics/), and HEK293T KRAB-ZNF and KAP1 ChIP-seq peaks (*59*). 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 that 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 and Hochberg method.

Permutation test of epigenetic marks overlapped with new subfamilies of each subfamily group

We examined the enrichment of the above epigenetic marks overlapped with each determined MER11 new subfamily. Specifically, we first shuffled the peak regions 100 times relative to the distribution of peaks using our previous approach (*71*). We then intersected the actual and shuffled peaks with instances from each new subfamily 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 new subfamily. 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. New subfamilies with a minimum of twofold enrichment $\{\log_2[(\text{actual counts} + 1)/(\text{mean shuffled counts} + 1)] \geq 1\}$ with *P* value ≤0.05 were kept. We also filtered out epigenetic marks overlapped with less than five instances from new annotations containing a maximum of 100 instances and less than 5% for new subfamilies containing more than 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 (*71*). 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.genome.jp/tools-bin/clustalw). 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 (panTro4) and macaque (macFas5) TE annotation files (Repeat library 20140131) were downloaded from http://www.repeatmasker.org/species/macFas.html and http://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 we previously used (*51*) were included 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 among 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 (*80*). In brief, the synthesized oligo pool was amplified by seven-cycle polymerase chain reaction (PCR) using forward primer (5BC-AG-f01; table S10) and reverse primer (5BC-AG-r01; table S10) that added mP and spacer sequences downstream of the sequence. The amplified fragments were purified with 1.8× AMPure XP (Beckman Coulter) and proceeded to the second round nine-cycle PCR using forward primer (5BC-AG-f02) and reverse primer (5BC-AG-r02; table S10) 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 S10). 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 S10).

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 the manufacturer's instruction. WTC11 cells were used for the MPRA experiments at passage 43. Lentivirus packaging was performed as previously described with modifications (*80*). Briefly, 50,000 cells/cm^2 293T cells (ATCC, CRL-3216) were seeded in four T175 flasks and cultured for 48 hours. The cells were cotransfected with 7.5 μg per flask of plasmid libraries, 2.5 μg per flask of pMD2.G (Addgene, 12259), and 5 μg per flask of psPAX2 (Addgene, 12260) using EndoFectin Lenti transfection reagent (GeneCopoeia, EF002) according to the manufacturer's instruction. After 8 hours, the cell culture medium was refreshed and ViralBoost reagent (Alstem, VB100) was added. The transfected cells were cultured for 2 days, and the lentivirus was filtered through a 0.45-μm PES filter system (Thermo Fisher Scientific, 165-0045) and concentrated by Lenti-X concentrator (Takara Bio, 631232) according to the manufacturer's protocol. We obtained in total 1.2 ml of lentivirus solution from the four T175 flasks (100× 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 per dish of the concentrated lentivirus library, 150 μl per dish ViroMag R/L reagent (OZ Biosciences, RL41000), and 750 μl per dish mTeSR plus medium were mixed and incubated at room temperature for 20 min. WTC11 cells in a 10-cm dish were added with the virus-ViroMag mixture and placed on a magnetic plate for 30 min. 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 (*51*). 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 (*80*). Briefly, genomic DNA and

total RNA were purified from the infected cells using an AllPrep DNA/RNA mini kit (Qiagen, 80204). RNA (120 μg) 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 S10), which has a 16-bp unique molecular identifier (UMI). The cDNA and 48 μg of genomic DNA from each sample were used for three-cycle PCR with specific primers (P7-pLSmp-assUMI-gfp and P5-pLSmP-5bc-i#; table S10) 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 and P7; table S10). 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, and pLSmP-5bc-seqR2; table S10).

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) (*80*). 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 (*51*). 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 on the basis of 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 (*81*). 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 new subfamily. Human, chimpanzee, and macaque sequences were analyzed separately.

Motif association analysis

We implemented a TE motif association approach to identify motifs that contributed to the activity. The nonredundant 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 seventh column) referred to the two pseudo alleles of a motif. The z-scaled MPRA activity was used as the quantitative phenotypes (sixth 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://cog-genomics.org/plink/2.0/) (*82*). 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 among 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.

Supplementary Materials

The PDF file includes:

Figs. S1 to S12

Legends for tables S1 to S10

Other Supplementary Material for this manuscript includes the following:

Tables S1 to S10

REFERENCES AND NOTES

1. T. Wang, J. Zeng, C. B. Lowe, R. G. Sellers, S. R. Salama, M. Yang, S. M. Burgess, R. K. Brachmann, D. Haussler, Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 18613–18618 (2007).
2. G. Bourque, B. Leong, V. B. Vega, X. Chen, Y. L. Lee, K. G. Srinivasan, J.-L. Chew, Y. Ruan, C.-L. Wei, H. H. Ng, E. T. Liu, Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.* **18**, 1752–1762 (2008).
3. V. Sundaram, Y. Cheng, Z. Ma, D. Li, X. Xing, P. Edge, M. P. Snyder, T. Wang, Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res.* **24**, 1963–1976 (2014).

4. E. B. Chuong, N. C. Elde, C. Feschotte, Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083–1087 (2016).
5. J. A. Frank, C. Feschotte, Co-option of endogenous viral sequences for host cell function. *Curr. Opin. Virol.* **25**, 81–89 (2017).
6. R. Fueyo, J. Judd, C. Feschotte, J. Wysocka, Roles of transposable elements in the regulation of mammalian transcription. *Nat. Rev. Mol. Cell Biol.* **23**, 481–497 (2022).
7. E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
8. S. J. Hoyt, J. M. Storer, G. A. Hartley, P. G. S. Grady, A. Gershman, L. G. de Lima, C. Limouse, R. Halabian, L. Wojenski, M. Rodriguez, N. Altemose, A. Rhie, L. J. Core, J. L. Gerton, W. Makalowski, D. Olson, J. Rosen, A. F. A. Smit, A. F. Straight, M. R. Vollger, T. J. Wheeler, M. C. Schatz, E. E. Eichler, A. M. Phillippy, W. Timp, K. H. Miga, R. J. O'Neill, From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science* **376**, eabk3112 (2022).
9. P.-É. Jacques, J. Jeyakani, G. Bourque, The majority of primate-specific regulatory sequences are derived from transposable elements. *PLOS Genet.* **9**, e1003504 (2013).
10. C. Feschotte, C. Gilbert, Endogenous viruses: Insights into viral evolution and impact on host biology. *Nat. Rev. Genet.* **13**, 283–296 (2012).
11. W. E. Johnson, Origins and evolutionary consequences of ancient endogenous retroviruses. *Nat. Rev. Microbiol.* **17**, 355–370 (2019).
12. H.-H. Zhang, J. Peccoud, M.-R.-X. Xu, X.-G. Zhang, C. Gilbert, Horizontal transfer and evolution of transposable elements in vertebrates. *Nat. Commun.* **11**, 1362 (2020).
13. C. Gilbert, C. Feschotte, Horizontal acquisition of transposable elements and viral sequences: Patterns and consequences. *Curr. Opin. Genet. Dev.* **49**, 15–24 (2018).
14. A. Molaro, H. S. Malik, Hide and seek: How chromatin-based pathways silence retroelements in the mammalian germline. *Curr. Opin. Genet. Dev.* **37**, 51–58 (2016).
15. M. V. Almeida, G. Vernaz, A. L. K. Putman, E. A. Miska, Taming transposable elements in vertebrates: From epigenetic silencing to domestication. *Trends Genet.* **38**, 529–553 (2022).
16. T. Chelmicki, E. Roger, A. Teissandier, M. Dura, L. Bonneville, S. Rucli, F. Dossin, C. Fouassier, S. Lameiras, D. Bourc'his, m^{6}A RNA methylation regulates the fate of endogenous retroviruses. *Nature* **591**, 312–316 (2021).
17. J. Jurka, Repeats in genomic DNA: Mining and meaning. *Curr. Opin. Struct. Biol.* **8**, 333–337 (1998).
18. E. J. Grow, R. A. Flynn, S. L. Chavez, N. L. Bayless, M. Wossidlo, D. J. Wesche, L. Martin, C. B. Ware, C. A. Blish, H. Y. Chang, R. A. Reijo Pera, J. Wysocka, Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522**, 221–225 (2015).
19. E. B. Chuong, N. C. Elde, C. Feschotte, Regulatory activities of transposable elements: From conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).
20. J. Göke, X. Lu, Y.-S. Chan, H.-H. Ng, L.-H. Ly, F. Sachs, I. Szczerbinska, Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell Stem Cell* **16**, 135–141 (2015).
21. P. Gerdes, S. R. Richardson, D. L. Mager, G. J. Faulkner, Transposable elements in the mammalian embryo: Pioneers surviving through stealth and service. *Genome Biol.* **17**, 100 (2016).
22. G. Ma, I. A. Babarinde, X. Zhou, A. P. Hutchins, Transposable elements in pluripotent stem cells and human disease. *Front. Genet.* **13**, (2022).
23. G. Kunarso, N.-Y. Chia, J. Jeyakani, C. Hwang, X. Lu, Y.-S. Chan, H.-H. Ng, G. Bourque, Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* **42**, 631–634 (2010).
24. J. Mayer, E. Meese, Human endogenous retroviruses in the primate lineage and their influence on host genomes. *Cytogenet. Genome Res.* **110**, 448–456 (2005).
25. G. Andrews, K. Fan, H. E. Pratt, N. Phalke, Zoonomia Consortium§, E. K. Karlsson, K. Lindblad-Toh, S. Gazal, J. E. Moore, Z. Weng, Mammalian evolution of human cis-regulatory elements and transcription factor binding sites. *Science* **380**, eabn7930 (2023).
26. S. Patoori, S. M. Barnada, C. Large, J. I. Murray, M. Trizzino, Young transposable elements rewired gene regulatory networks in human and chimpanzee hippocampal intermediate progenitors. *Development* **149**, dev200413 (2022).
27. X. Xiang, Y. Tao, J. DiRusso, F.-M. Hsu, J. Zhang, Z. Xue, J. Pontis, D. Trono, W. Liu, A. T. Clark, Human reproduction is regulated by retrotransposons derived from ancient Hominidae-specific viral infections. *Nat. Commun.* **13**, 463 (2022).
28. M. Trizzino, Y. Park, M. Holsbach-Beltrame, K. Aracena, K. Mika, M. Caliskan, G. H. Perry, V. J. Lynch, C. D. Brown, Transposable elements are the primary source of novelty in primate gene regulation. *Genome Res.* **27**, 1623–1633 (2017).
29. D. R. Fuentes, T. Swigut, J. Wysocka, Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. *eLife* **7**, e35989 (2018).
30. C. E. Sexton, R. L. Tillett, M. V. Han, The essential but enigmatic regulatory role of HERVH in pluripotency. *Trends Genet.* **38**, 12–21 (2022).
31. A. D. Senft, T. S. Macfarlan, Transposable elements shape the evolution of mammalian development. *Nat. Rev. Genet.* **22**, 691–711 (2021).
32. C. G. Sotero-Caio, R. N. Platt II, A. Suh, D. A. Ray, Evolution and diversity of transposable elements in vertebrate genomes. *Genome Biol. Evol.* **9**, 161–177 (2017).
33. T. Wicker, F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel, A. H. Schulman, A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).
34. D. R. Hoen, G. Hickey, G. Bourque, J. Casacuberta, R. Cordaux, C. Feschotte, A.-S. Fiston-Lavier, A. Hua-Van, R. Hubley, A. Kapusta, E. Lerat, F. Maumus, D. D. Pollock, H. Quesneville, A. Smit, T. J. Wheeler, T. E. Bureau, M. Blanchette, A call for benchmarking transposable element annotation methods. *Mob. DNA* **6**, 13 (2015).
35. I. R. Arkhipova, Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. *Mob. DNA* **8**, 19 (2017).
36. K. M. Carey, G. Patterson, T. J. Wheeler, Transposable element subfamily annotation has a reproducibility problem. *Mob. DNA* **12**, 4 (2021).
37. N. T. Hassan, D. L. Adelson, Fake IDs? Widespread misannotation of DNA transposons as a general transcription factor. *Genome Biol.* **24**, 260 (2023).
38. S. Mun, J. Lee, Y.-J. Kim, H.-S. Kim, K. Han, Chimpanzee-specific endogenous retrovirus generates genomic variations in the chimpanzee genome. *PLOS ONE* **9**, e101195 (2014).
39. M. Escalera-Zamudio, A. D. Greenwood, On the classification and evolution of endogenous retrovirus: Human endogenous retroviruses may not be 'human' after all. *APMIS* **124**, 44–51 (2016).
40. Y. Li, G. Zhang, J. Cui, Origin and deep evolution of human endogenous retroviruses in pan-primates. *Viruses* **14**, 1370 (2022).
41. J. Blomberg, F. Benachenhou, V. Blikstad, G. Sperber, J. Mayer, Classification and nomenclature of endogenous retroviral sequences (ERVs): Problems and recommendations. *Gene* **448**, 115–123 (2009).
42. T. A. Carter, M. Singh, G. Dumbović, J. D. Chobirko, J. L. Rinn, C. Feschotte, Mosaic cis-regulatory evolution drives transcriptional partitioning of HERVH endogenous retrovirus in the human embryo. *eLife* **11**, e76257 (2022).
43. A. Le Rouzic, T. S. Boutin, P. Capy, Long-term evolution of transposable elements. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 19375–19380 (2007).
44. N. Grandi, M. P. Pisano, E. Pessiu, S. Scognamiglio, E. Tramontano, HERV-K(HML7) integrations in the human genome: Comprehensive characterization and comparative analysis in non-human primates. *Biology* **10**, 439 (2021).
45. S. Scognamiglio, N. Grandi, E. Pessiu, E. Tramontano, Identification, comprehensive characterization, and comparative genomics of the HERV-K(HML8) integrations in the human genome. *Virus Res.* **323**, 198976 (2023).

46. J. Ito, R. Sugimoto, H. Nakaoka, S. Yamada, T. Kimura, T. Hayano, I. Inoue, Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. *PLOS Genet.* **13**, e1006883 (2017).
47. H. J. Atkinson, J. H. Morris, T. E. Ferrin, P. C. Babbitt, Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLOS ONE* **4**, e4345 (2009).
48. H. J. Bandelt, P. Forster, A. Röhl, Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16**, 37–48 (1999).
49. S. Naser-Khdour, B. Q. Minh, R. Lanfear, Assessing confidence in root placement on phylogenies: An empirical study using nonreversible models for mammals. *Syst. Biol.* **71**, 959–972 (2022).
50. C. Hermant, M.-E. Torres-Padilla, TFs for TEs: The transcription factor repertoire of mammalian transposable elements. *Genes Dev.* **35**, 22–39 (2021).
51. F. Inoue, A. Kreimer, T. Ashuach, N. Ahituv, N. Yosef, Identification and massively parallel characterization of regulatory elements driving neural induction. *Cell Stem Cell* **25**, 713–727.e10 (2019).
52. W. Xie, M. D. Schultz, R. Lister, Z. Hou, N. Rajagopal, P. Ray, J. W. Whitaker, S. Tian, R. D. Hawkins, D. Leung, H. Yang, T. Wang, A. Y. Lee, S. A. Swanson, J. Zhang, Y. Zhu, A. Kim, J. R. Nery, M. A. Urich, S. Kuan, C. Yen, S. Klugman, P. Yu, K. Suknuntha, N. E. Propson, H. Chen, L. E. Edsall, U. Wagner, Y. Li, Z. Ye, A. Kulkarni, Z. Xuan, W.-Y. Chung, N. C. Chi, J. E. Antosiewicz-Bourget, I. Slukvin, R. Stewart, M. Q. Zhang, W. Wang, J. A. Thomson, J. R. Ecker, B. Ren, Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–1148 (2013).
53. M. Bruno, M. Mahgoub, T. S. Macfarlan, The arms race between KRAB–Zinc finger proteins and endogenous retroelements and its impact on mammals. *Annu. Rev. Genet.* **53**, 393–416 (2019).
54. G. Bourque, K. H. Burns, M. Gehring, V. Gorbunova, A. Seluanov, M. Hammell, M. Imbeault, Z. Izsvák, H. L. Levin, T. S. Macfarlan, D. L. Mager, C. Feschotte, Ten things you should know about transposable elements. *Genome Biol.* **19**, 199 (2018).
55. P. Kheradpour, J. Ernst, A. Melnikov, P. Rogov, L. Wang, X. Zhang, J. Alston, T. S. Mikkelsen, M. Kellis, Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* **23**, 800–811 (2013).
56. A. Y. Du, X. Zhuo, V. Sundaram, N. O. Jensen, H. G. Chaudhari, N. L. Saccone, B. A. Cohen, T. Wang, Functional characterization of enhancer activity during a long terminal repeat's evolution. *Genome Res.* **32**, 1840–1851 (2022).
57. R. Yu, S. Roseman, A. P. Siegenfeld, Z. Gardner, S. C. Nguyen, K. A. Tran, E. F. Joyce, R. Jain, B. B. Liau, I. D. Krantz, K. A. Alexander, S. L. Berger, CTCF/RAD21 organize the ground state of chromatin–nuclear speckle association. *Nat. Struct. Mol. Biol.* **32**, 1069–1080 (2025).
58. M. D. Magnitov, M. Maresca, N. A. Saiz, H. Teunissen, J. Dong, K. M. Sathyan, L. Braccioli, M. J. Guertin, E. de Wit, ZNF143 is a transcriptional regulator of nuclear-encoded mitochondrial genes that acts independently of looping and CTCF. *Mol. Cell* **85**, 24–41. e11 (2025).
59. M. Imbeault, P.-Y. Helleboid, D. Trono, KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550–554 (2017).
60. E. De Franco, N. D. L. Owens, H. Montaser, M. N. Wakeling, J. Saarimäki-Vire, A. Triantou, H. Ibrahim, D. Balboa, R. C. Caswell, R. E. Jennings, J. A. Kvist, M. B. Johnson, S. Muralidharan, S. Ellard, C. F. Wright, S. Maddirevula, F. S. Alkuraya, N. A. Hanley, S. E. Flanagan, T. Otonkoski, A. T. Hattersley, M. Imbeault, Primate-specific ZNF808 is essential for pancreatic development in humans. *Nat. Genet.* **55**, 2075–2081 (2023).
61. M. Kosuge, J. Ito, M. Hamada, Landscape of evolutionary arms races between transposable elements and KRAB-ZFP family. *Sci. Rep.* **14**, 23358 (2024).
62. M. Pierson Smela, A. Sybirna, F. C. K. Wong, M. A. Surani, Testing the role of SOX15 in human primordial germ cell fate. *Wellcome Open Res.* **4**, 122 (2019).
63. N. Irie, L. Weinberger, W. W. C. Tang, T. Kobayashi, S. Viukov, Y. S. Manor, S. Dietmann, J. H. Hanna, M. A. Surani, SOX17 is a critical specifier of human primordial germ cell fate. *Cell* **160**, 253–268 (2015).
64. M. Friedli, D. Trono, The developmental control of transposable elements and the evolution of higher species. *Annu. Rev. Cell Dev. Biol.* **31**, 429–451 (2015).
65. O. Rosspopoff, D. Trono, Take a walk on the KRAB side. *Trends Genet.* **39**, 844–857 (2023).
66. M. N. Choudhary, R. Z. Friedman, J. T. Wang, H. S. Jang, X. Zhuo, T. Wang, Co-opted transposons help perpetuate conserved higher-order chromosomal structures. *Genome Biol.* **21**, 16 (2020).
67. A. G. Diehl, N. Ouyang, A. P. Boyle, Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes. *Nat. Commun.* **11**, 1796 (2020).
68. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
69. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
70. L. Bogdan, L. Barreiro, G. Bourque, Transposable elements have contributed human regulatory regions that are activated upon bacterial infection. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **375**, 20190332 (2020).
71. X. Chen, A. Pacis, K. A. Aracena, S. Gona, T. Kwan, C. Groza, Y. L. Lin, R. Sindeaux, V. Yotova, A. Pramatarova, M.-M. Simon, T. Pastinen, L. B. Barreiro, G. Bourque, Transposable elements are associated with the variable response to influenza infection. *Cell Genom.* **3**, 100292 (2023).
72. J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, J. Walichiewicz, Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
73. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden, BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
74. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
75. T. Nakamura, K. D. Yamada, K. Tomii, K. Katoh, Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* **34**, 2490–2492 (2018).
76. A. Löytynoja, Phylogeny-aware alignment with PRANK. *Methods Mol. Biol.* **1079**, 155–170 (2014).
77. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
78. B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, R. Lanfear, IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
79. J. W. Leigh, D. Bryant, popart: Full-feature software for haplotype network construction. *Methods Ecol. Evol.* **6**, 1110–1116 (2015).
80. M. G. Gordon, F. Inoue, B. Martin, M. Schubach, V. Agarwal, S. Whalen, S. Feng, J. Zhao, T. Ashuach, R. Ziffra, A. Kreimer, I. Georgakopoulos-Soares, N. Yosef, C. J. Ye, K. S. Pollard, J. Shendure, M. Kircher, N. Ahituv, lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. *Nat. Protoc.* **15**, 2387–2412 (2020).
81. T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, W. S. Noble, MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
82. C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, J. J. Lee, Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).

Acknowledgments: We would like to thank G. Gordon and T. Ashuach for help with the use of MPRAflow and MPRAnalyze tools. We also acknowledge the Institute for the Advanced Study of Human Biology (ASHBi), Institutional Center for Shared Technologies and Facilities of Shanghai Institute of Immunity and Infection, CAS, 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 support. We thank S. Goulas for critical reading and suggestions on the manuscript. **Funding:** This work was supported by World Premier International Research Center Initiative (WPI), MEXT, Japan (G.B.); JSPS KAKENHI (JP21K06119) (F.I.); JSPS KAKENHI (JP21K15066) (X.C.); the Takeda Science Foundation, Bioscience Research Grants (F.I.); the Mitsubishi Foundation, research grants in the Natural Sciences (F.I.); Canadian Institute of Health Research (CIHR) program grant (CEE-151618) (G.B.); Canada Research Chair Tier 1 award (G.B.), an FRQ-S, Distinguished Research Scholar award (G.B.); and Calcul Quebec and the Digital Research Alliance of Canada (G.B.). **Author contributions:** Conceptualization: F.I., X.C., and G.B. Methodology: F.I., X.C., G.B., Z.Z., and C.G. Software: X.C. Validation: X.C. Formal analysis: X.C. and Z.Z. Investigation: F.I., X.C., Z.Z., and Y.Y. Resources: F.I. and G.B. Data curation: X.C. Writing—original draft: F.I. and X.C. Writing—review and editing: F.I., X.C., G.B., Z.Z., and C.G. Visualization: X.C. Supervision: F.I. and G.B. Project administration: F.I. and G.B. Funding acquisition: F.I., X.C., and G.B. **Competing interests:** F.I. receives funding from Relation Therapeutics. All other authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The datasets generated in this study are available at the NCBI Gene Expression Omnibus (GEO) as accession number GEO: GSE245662, https://ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE245662. Scripts for main analyses are available on Zenodo at DOI:10.5281/zenodo.10016500, https://zenodo.org/records/10016500. Scripts are also available on Github at https://github.com/xunchen85/TE-MPRA-and-phylogenetic-analysis as an additional resource. The following cell line was obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research: GM25256. The WTC11 cells can be provided by Corriell pending scientific review and a completed material transfer agreement. Requests for the WTC11 should be submitted to: CORIELL CELL REPOSITORIES (ccr@coriell.org).

Submitted 4 September 2024
Accepted 13 June 2025
Published 18 July 2025
10.1126/sciadv.ads9164