scientific reports

Check for updates

OPEN

Landscape of evolutionary arms races between transposable elements and KRAB-ZFP family

Masato Kosuge[1,2], Jumpei Ito[3] & Michiaki Hamada[1,2,4]✉

Transposable elements (TEs) are mobile parasitic sequences that have expanded within the host genome. It has been hypothesized that host organisms have expanded the Krüppel-associated box-containing zinc finger proteins (KRAB-ZFPs), which epigenetically suppress TEs, to counteract disorderly TE transpositions. This process is referred to as the evolutionary arms race. However, the extent to which this evolutionary arms race occurred across various TE families remains unclear. In the present study, we systematically explored the evolutionary arms race between TE families and human KRAB-ZFPs using public ChIP-seq data. We discovered and characterized new instances of evolutionary arms races with KRAB-ZFPs in endogenous retroviruses. Furthermore, we found that the regulatory landscape shaped by this arms race contributed to the gene regulatory networks. In summary, our results provide insight into the impact of the evolutionary arms race on TE families, the KRAB-ZFP family, and host gene regulatory networks.

Transposable elements (TEs) are mobile parasitic genetic sequences that comprise approximately 46% of the human genome[1]. Although most insertions of TEs near genes are likely either harmful or neutral to their host organisms, TEs have significantly influenced the evolution of their host organisms through transpositions[1]. TEs possess numerous binding sites for transcription factors (TFs) and their insertion generates new binding sites for TFs near genes[2,3]. Some new TE insertions can function as cis-regulatory elements, such as enhancers or alternative promoters of genes near their insertion sites, thereby altering the expression patterns of these genes[4,5]. Furthermore, TEs gain or lose the binding sites of specific TFs during evolution, resulting in each TE subfamily having distinct expression profiles and effects on nearby genes[2,6]. Therefore, uncovering the evolution of TEs is crucial for understanding the evolution of host organisms.

Krüppel-associated box-containing zinc finger proteins (KRAB-ZFPs) are transcriptional repressors that epigenetically suppress the transcription of TEs[7]. The human genome contains approximately 400 genes encoding KRAB-ZFPs [8–10]. Each KRAB-ZFP comprises one or two KRAB domains that interact with TRIM28 and several zinc finger (ZF) domains that recognize sequences in TEs[7]. KRAB-ZFPs recruit SETDB1, HP1, and the nucleosome and remodeling deacetylase complex via TRIM28, which induces H3K9me3 and DNA methylation, thereby epigenetically suppressing TEs[11–13]. In humans, depletion of KRAB-ZFPs and TRIM28 leads to de-repression of TEs in developmental stages[14,15]. In addition, in mice, deletion of the KRAB-ZFP cluster slightly induced transpositions of ERVs[16]. Thus, KRAB-ZFPs play a crucial role in the suppression of TEs.

It has been hypothesized that the significant expansion of KRAB-ZFPs is the result of an evolutionary arms race with TEs[7]. This arms race is a co-evolutionary process among competitors. In the context of KRAB-ZFPs and TEs, the proposed evolutionary scenario is as follows. As new TEs emerge and proliferate within the host genome, KRAB-ZFPs that specifically suppress these TEs emerge through gene duplication. Subsequently, TEs acquire mutations in the binding sites of KRAB-ZFPs, enabling them to avoid suppression by KRAB-ZFPs. This evolutionary scenario is supported by reports indicating that TEs are often targeted by KRAB-ZFPs that emerged concurrently[8,9,17]. Additionally, the long interspersed nuclear element (LINE) 1 family evades suppression by a specific KRAB-ZFP, ZNF93, through deletions at its binding sites[18]. However, the extent to which this evolutionary arms race has occurred in other TE families remains unclear.

Such a relationship between TEs and KRAB-ZFPs has been proposed to be co-opted as part of the gene regulatory networks[8,19]. KRAB-ZFPs regulate gene expression by suppressing TEs that function as cis-regulatory

[1]Graduate School of Advanced Science and Engineering, Waseda University, Tokyo, Japan. [2]Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan. [3]Division of Systems Virology, Department of Microbiology and Immunology, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [4]Graduate School of Medicine, Nippon Medical School, Tokyo, Japan. ✉email: mhamada@waseda.jp

elements[8,20–22]. Perturbation of KRAB-ZFP or TRIM28 expression led to aberrant expression of nearby genes[23–25]. However, the effect of the evolutionary arms race between TE families and KRAB-ZFPs on the gene regulatory networks is still not well understood.

In this study, we aimed to comprehensively investigate the evolutionary arms race between TE families and KRAB-ZFPs using multi-layered analyses: (1) identification of KRAB-ZFP targets by leveraging publicly available ChIP-seq data, (2) comparison of the evolutionary ages between TE subfamilies and KRAB-ZFPs, and (3) phylogenetic analyses of the TE family (Fig. 1). Accordingly, we reconstructed and characterized the evolutionary arms race with KRAB-ZFPs in several TE families, which have not been previously reported. Finally, we provide evidence that these arms race dynamics have potentially influenced the evolution of the host gene regulatory networks. Our findings illustrate the co-evolutionary relationship between TE and KRAB-ZFPs, offering fascinating insights into TE-host interactions.

Results

Characterization of the relationship between TE families and KRAB-ZFPs

To comprehensively characterize the relationship between TE families and KRAB-ZFPs, we first collected and processed a large dataset of KRAB-ZFP ChIP-seq experiments encompassing 361 out of 378 KRAB-ZFPs from 1,051 samples (Supplementary Fig. 1a,b and Supplementary Table 1). We obtained consensus sequences and metadata for the 1170 TE subfamilies belonging to 533 TE families from Dfam[26]. Given the association between alterations in TE sequences over evolution and KRAB-ZFP binding, we focused on the differences in sequences and KRAB-ZFP binding among the TE subfamilies[8,18,27]. To comprehensively explore the evolutionary arms race with KRAB-ZFPs, we developed a novel subfamily classification pipeline based on the genetic distance between TE copies and performed subfamily classification in TE families that previously lacked subfamily classification (344 families, 65%) (Supplementary Figs. 2, 3 and Methods). By applying our pipeline to the TE families, we successfully divided the 59 families that met our inclusion criteria into subfamilies (Supplementary Fig. 2c,d). The final dataset comprised 533 TE families and 1269 subfamilies (Supplementary Data 1).

We conducted enrichment analyses to examine the TE targets of KRAB-ZFPs and identified 616 primary and 888 secondary targets of KRAB-ZFPs (Supplementary Data 2 and Methods). Among 361 KRAB-ZFPs, 266 (74%) targeted at least one subfamily (Supplementary Fig. 4a). Additionally, 472 (37%) subfamilies and 146 (27%) families were targeted by at least one KRAB-ZFP (Supplementary Fig. 4b). Consistent with previous studies[9], LINEs, endogenous retroviruses (ERVs), and SINE-VNTR-Alus (SVAs) were enriched as targets of KRAB-ZFPs (Fig. 2a and Supplementary Fig. 4c).

Previous studies have described the co-emergence of KRAB-ZFPs and their target TE subfamilies in the same era[8,9,17]. Hence, we obtained the evolutionary ages of the KRAB-ZFPs from two previous studies and inferred the evolutionary ages of their target TE subfamilies (Supplementary Data 1, 3). Both the KRAB-ZFPs and TE subfamilies primarily emerged during the evolutionary period of the common ancestors of Eutherians around 105 million years ago (MYA), Simiiformes around 43 MYA, and Catarrhines around 29 MYA. (Supplementary Fig. 4d). Consistent with previous studies, the evolutionary ages of KRAB-ZFPs tended to coincide with those of their target TE subfamilies, suggesting that KRAB-ZFPs emerged in response to the emergence and expansion of

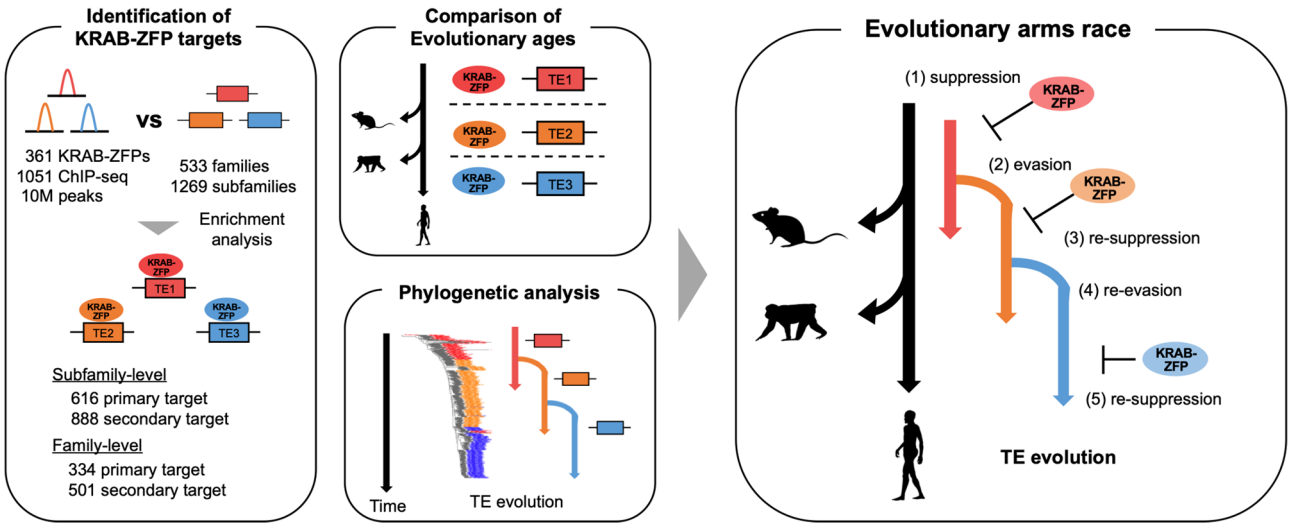

Fig. 1. Overview of the study design and the evolutionary arms race model between TE family and KRAB-ZFPs. Left panel: Schematic of the main analyses performed in this study, including identification of KRAB-ZFP targets, comparison of the evolutionary ages between KRAB-ZFPs and TE families, and phylogenetic analysis of TE families. Right panel: The evolutionary arms race between the TE family and KRAB-ZFPs. This arms race involves repeated cycles of (1) TE suppression by KRAB-ZFPs, (2) TE evasion from KRAB-ZFP suppression, (3) re-suppression by existing or newly emerged KRAB-ZFPs, (4) further TE evasion, and (5) re-suppression by additional KRAB-ZFPs.

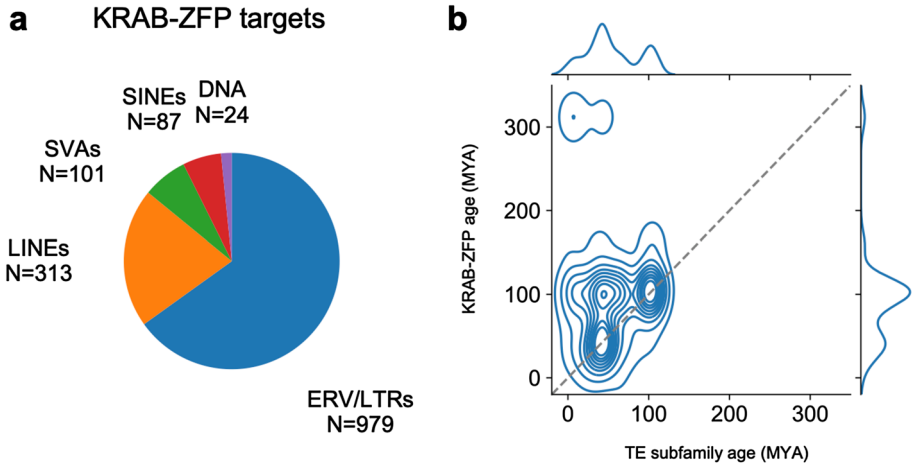

Fig. 2. Co-emergence of TE subfamilies and KRAB-ZFPs. (**a**) Proportion of TE classes in the identified KRAB-ZFP targets. The pie chart shows the distribution of TE classes, including ERVs (blue), LINEs (orange), SINEs (green), SVAs (red), and DNA transposons (purple). The number of TE subfamilies in each class is indicated by "N = ". (**b**) Comparison of evolutionary ages (in million years ago, MYA) between the TE subfamily (X-axis) and KRAB-ZFPs (Y-axis) for 616 primary KRAB-ZFP-TE subfamily pairs.

new TE subfamilies (Fig. 2b and Supplementary Fig. 4e–g). Unexpectedly, the KRAB-ZFPs that emerged in the common ancestor with non-primates targeted primate-specific TE subfamilies, whereas the reverse pattern was rarely observed. (Fig. 2b and Supplementary Fig. 4 h). These observations imply that KRAB-ZFPs that emerged from the common ancestor with non-primates specifically target and suppress newly emerged TE subfamilies. In summary, these findings support the hypothesis that the KRAB-ZFP family has co-evolved with TEs.

Several TE families have undergone evasion from KRAB-ZFPs

To investigate the evasion of KRAB-ZFP suppression by TE families, we developed a computational screening approach based on the differential binding of KRAB-ZFPs to young and old subfamilies within each TE family (Fig. 3a). Our rationale was that if a TE family evolved to evade KRAB-ZFP suppression, we would expect to see a significant reduction in KRAB-ZFP binding to its younger subfamilies compared to older ones. Applying this approach to our dataset, we identified 62 evasion candidates that showed reduced binding in younger subfamilies (Fig. 3b and Supplementary Data 4). As positive controls, loss of binding by ZNF765, ZNF649, and ZNF93 at the L1P_5end was also detected, which is consistent with previous studies (Supplementary Fig. 5a,b)[8,18,28]. In addition to the previously reported L1P_5end, evasion candidates were detected in several ERV families and SVAs, suggesting that the evasions of TE families from KRAB-ZFPs occurred across a broad range of TE families.

We focused on ERVs or long terminal repeats (LTRs) because HERVK, MER11, HERVH, and LTR7 were among the top results, in addition to LINE1. We observed many evasion candidates in the MER11 family (Fig. 3c). MER11, the LTR of HERVK11, first emerged from the common ancestor of Catarrhines (29.4 MYA) and continued to transpose to the common ancestor of Hominidae (9.1 MYA). Recent research has highlighted that the MER11 family comprises more subfamilies than previously identified[6]. Consequently, we reapplied our subfamily classification pipeline to the MER11 family and classified the MER11 family into 6 subfamilies (Supplementary Fig. 6a,b). In the MER11 family, evasions from seven KRAB-ZFPs were observed, and the binding of these KRAB-ZFPs gradually disappeared over the course of MER11 evolution (Fig. 3c and Supplementary Fig. 6c). Among these KRAB-ZFPs, ZNF611, ZNF440, and ZNF808 emerged in the same era as their target MER11 subfamilies, whereas ZNF445, ZNF33A, ZNF468, and ZNF433 appeared in eras older than their target MER11 subfamilies. Furthermore, we identified the binding sites of ZNF468 and ZNF433 and found that the disappearance of binding was attributable to mutations or indels at their binding sites (Fig. 3d and Supplementary Fig. 6d,e). At the binding site of ZNF468, position 1258 was mutated during the evolution of MER11_2, preventing ZNF468 from binding to MER11 copies via a 1258 C-to-T mutation (Fig. 3e). These findings suggest that MER11 evolved to evade KRAB-ZFP suppression through point mutations and indels. Additionally, the younger MER11 subfamilies were targeted by ZNF525, ZNF578, and ZNF727, which complemented the loss of KRAB-ZFP binding (Fig. 3c and Supplementary Fig. 6c). ZNF525 and ZNF727 bound to MER11 copies that were active in the same era as their emergences, suggesting that the MER11 family was targeted by the newly emerged KRAB-ZFPs after evasion from several KRAB-ZFPs. These findings illustrate the existence of an evolutionary arms race between the MER11 family and KRAB-ZFPs. A similar pattern was observed in the LTR5 family (Supplementary Fig. 7). In summary, these results suggest that evasion from KRAB-ZFPs occurred across a broader range of TE families than previously reported, leading to an evolutionary arms race with KRAB-ZFPs.

Evolutionary arms races between ERVs and KRAB-ZFPs

ERVs exist within the genome either as full-length ERVs or as solo LTRs[29]. ERVs are typically annotated by dividing them into LTRs and internal regions. Because ERVs expand through full-length ERVs, not solo-LTRs, to fully

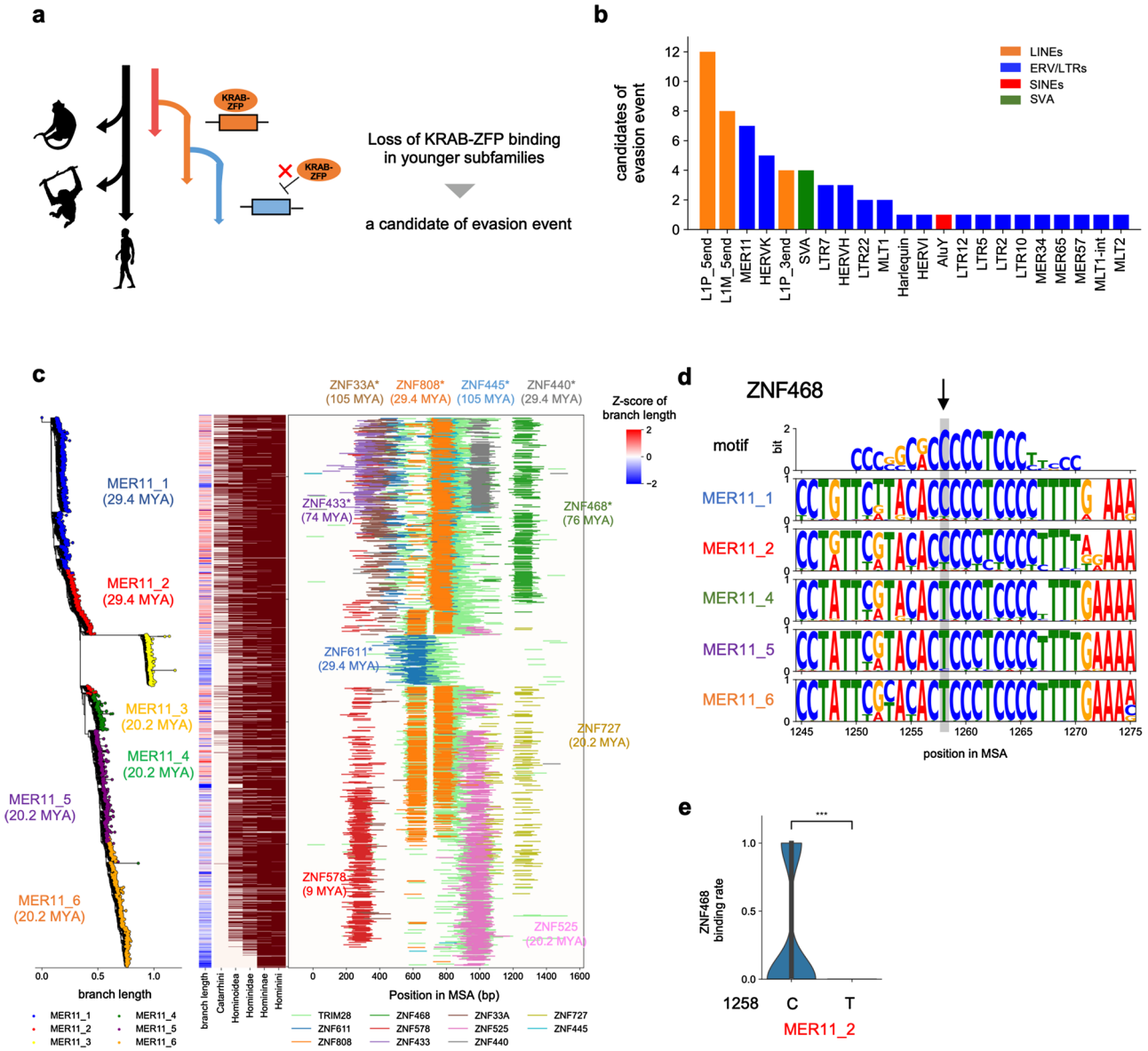

Fig. 3. Screening for TE evasion from KRAB-ZFPs and the evolutionary arms race in MER11. (**a**) Schematic of the screening approach for identifying potential TE evasion events from KRAB-ZFP repression. The screening was based on the loss of KRAB-ZFP binding in younger TE subfamilies compared with older subfamilies within the same family. (**b**) Number of candidate evasion events identified in each TE family. The bar plot shows the number of KRAB-ZFPs that each TE family potentially evaded in our screening approach, including ERVs (blue), LINEs (orange), SINEs (green), and SVA (red). (**c**) Evolutionary arms race in MER11 with 10 KRAB-ZFPs. 7 KRAB-ZFPs potentially evaded by the MER11 family and 3 KRAB-ZFPs primarily targeted the MER11 family. The asterisk (*) after gene symbol of KRAB-ZFPs indicates the significance of screening for evasion events. The phylogenetic tree (left) indicates the phylogenetic relationships between the MER11 subfamilies. The ages listed under subfamily names indicate the evolutionary ages of the MER11 subfamilies. The heatmap of the branch length and liftover in the center indicates the insertion date of the MER11 copies. The upper side of the y-axis indicates older MER11 copies and subfamilies. The plot on the right indicates the positions of TRIM28 and 10 KRAB-ZFP peaks for each MER11 copy. Peaks are depicted in different colors for each KRAB-ZFP gene. (**d**) Evasion of ZNF468 due to point mutations. The sequence logos indicate the de novo motif of ZNF468 (top) and the sequence of the ZNF468 binding site in the MER11 subfamily. Letters indicate the proportion of bases in each position. The arrow indicates the position potentially associated with the loss of ZNF468 binding. (**e**) Effect of the mutation on ZNF468 binding to the MER11 family. The violin plot on the right indicates the binding rate of ZNF468 to the MER11_2 subfamily with (left) and without (right) the 1258 C-to-T mutation. Statistical testing was conducted using the two-sided Mann–Whitney U test. ***$p < 0.001$.

understand the evolutionary arms race between ERVs and KRAB-ZFPs, it is necessary to analyze ERVs not as separate LTRs and internal regions, but as full-length ERVs. Therefore, we reconstructed full-length ERVs from the TE annotation data (Fig. 4a and Supplementary Fig. 8). As a result of this reconstruction, we identified 5518 full-length ERVs in 19 types of ERVs (Fig. 4b and Supplementary Data 5). We identified 20 5'-LTRs, 46 3'-LTRs, and 82 internal regions as targets of KRAB-ZFPs (Supplementary Fig. 9a and Supplementary Data 2). Because 5'-LTR and 3'-LTR are derived from the identical sequence, we integrated the results and identified 48 pairs of KRAB-ZFPs and LTRs. Interestingly, the binding rates of KRAB-ZFP varied not only in the LTRs but also in the internal region between subfamilies (Fig. 4c). This trend was also observed in TRIM28 (Supplementary Fig. 9b). These findings suggest that alterations in sequences and KRAB-ZFP binding occur not only in LTRs but also in internal regions, implying that both regions had been involved in the evolutionary arms race with KRAB-ZFPs.

Upon examining all full-length ERVs in detail, remarkable characteristics of the evolutionary arms race with KRAB-ZFPs were observed in LTR7_HERVH (Fig. 4d and Supplementary Fig. 10) and THE1_THE1-int (Supplementary Fig. 11). We focused on LTR7_HERVH because LTR7_HERVH is an ERV essential for the pluripotency of human embryonic stem cells (hESCs)[30,31]. LTR7_HERVH first emerged from the common ancestor of Catarrhines and remained active until the common ancestor of Hominidae. Our pipeline classified LTR7_HERVH into 8 subfamilies, which is consistent with the subfamily classification reported in a previous study (Supplementary Fig. 10a,b)[32]. In LTR7_HERVH, we observed 11 KRAB-ZFPs and 6 TRIM28 binding sites, most of which exhibited subfamily specificity (Fig. 4d). 5 KRAB-ZFPs targeted the old LTR7_HERVH subfamilies (LTR7_HERVH_1, 2), which emerged from the common ancestor of Catarrhines, whereas the other KRAB-ZFPs targeted the middle-aged LTR7_HERVH subfamilies (LTR7_HERVH_3–5) and young LTR7_HERVH subfamilies (LTR7_HERVH_6–8), which emerged from the common ancestor of Hominoidea and Hominidae, respectively (Fig. 4d and Supplementary Fig. 10c). While the old LTR7_HERVH subfamilies and 3 of KRAB-ZFPs emerged in the same era (Catarrhini, 29.4 Mya), most KRAB-ZFPs targeting middle-aged and young LTR7_HERVH subfamilies existed before these subfamilies (Fig. 4e). Finally, young LTR7_HERVH was targeted by ZNF90, which emerged from the common ancestor of Hominidae, suggesting a complementary response to the emergence of the new LTR7_HERVH subfamilies. Moreover, changes in the binding affinities of these KRAB-ZFPs were attributed to mutations and indels, similar to those observed in the MER11 family (Fig. 4f and Supplementary Fig. 10d, e). Specifically, the binding sites of ZNF600 were deleted in the LTR7_HERVH_3–8 subfamilies, along with the loss of its binding signal (Fig. 4f). This suggests that these middle-aged and young subfamilies evaded the regulatory control of ZNF600. These results suggest that LTR7_HERVH and some ERVs have undergone an evolutionary arms race with the KRAB-ZFP family.

Evolutionary arms race shapes regulatory landscape of LTR7_HERVH and nearby genes in hESCs

Finally, we aimed to reveal the relationship between the evolutionary arms race and the co-option of LTR7_HERVH in the gene regulatory networks of hESCs. Previous studies have shown that LTR7 is bound by pluripotency-related TFs (NANOG, KLF4, OCT4, SOX2, FOXP2, and FOXA1) and other TFs (FOXA2, GATA6, and EOMES)[2,32]. Therefore, we examined the binding patterns of KRAB-ZFPs, TRIM28, and these TFs, as well as the differences in chromatin state and expression between the LTR7_HERVH subfamilies in hESCs (Supplementary Fig. 12a and Supplementary Table 1)[23,33–35]. Consistent with the results of previous studies, young subfamilies, bound by KLF4, were highly expressed in hESCs (Supplementary Fig. 12a)[32,36]. In contrast, old and middle-aged subfamilies tended to overlap with the heterochromatin state (9_Het in ChromHMM 15 models) enriched with H3K9me3 modifications compared to that in the youngest LTR7_HERVH subfamily (Supplementary Fig. 12a). Although not statistically significant in some LTR7_HERVH subfamilies, LTR7_HERVH copies overlapping with 9_Het showed lower expression than those without an overlap (Supplementary Fig. 12b). To examine the roles of KRAB-ZFPs and TRIM28 in LTR7_HERVH expression in hESCs, we compared LTR7_HERVH expression in wild-type and TRIM28 knock out (KO) hESCs. Interestingly, the deficiency of TRIM28 induced the upregulation of the old and middle-aged subfamilies and the downregulation of the youngest subfamily (Fig. 5a). The expression levels of these TFs did not change significantly (Supplementary Fig. 12c). Moreover, LTR7_HERVH copies overlapping with 9_Het were more highly upregulated than other LTR7_HERVH copies (Fig. 5b). Consistent with these findings, TRIM28 peaks were observed in LTR7_HERVH copies in hESCs (Supplementary Fig. 12d). In the old LTR7_HERVH subfamilies, several TRIM28 binding sites associated with KRAB-ZFPs were observed. Together, these results suggest that the subfamily-specific binding patterns of KRAB-ZFPs and TRIM28 formed through the evolutionary arms race also contribute to shaping the subfamily-specific expression of LTR7_HERVH in hESCs.

Furthermore, many recent studies have demonstrated that the derepression of TEs affects the expression of nearby genes[21,23,24]. Therefore, we verified the impact of alterations in LTR7_HERVH expression on the expression of nearby genes within 50 kb of the LTR7_HERVH copy. Transcripts derived from LTR7_HERVH are sometimes annotated as lncRNA regardless of their function, hence only protein-coding genes were used in the analysis. Consistent with our hypothesis, the expression of genes near the LTR7_HERVH copies was significantly upregulated more than that in all the genes (Fig. 5c). Although not significant, the genes that were upregulated in TRIM28 KO hESCs tended to be closer to LTR7_HERVH copies than those that were not differentially expressed (Supplementary Fig. 12e). Specifically, in the LTR7_HERVH_2 subfamily, which exhibited the greatest increase in expression as shown in Fig. 5a, the proportion of upregulated genes was significantly higher than that of all the genes (21% vs. 2%). However, the proportion of downregulated genes did not follow this pattern (0% vs. 2%), suggesting that the derepression of LTR7_HERVH induced the upregulation of nearby genes. These findings suggest that KRAB-ZFPs/TRIM28 regulates gene expression via LTR7_HERVH. In summary, our observations suggest that the binding patterns of KRAB-ZFPs/TRIM28 formed through the evolutionary arms race suppresses

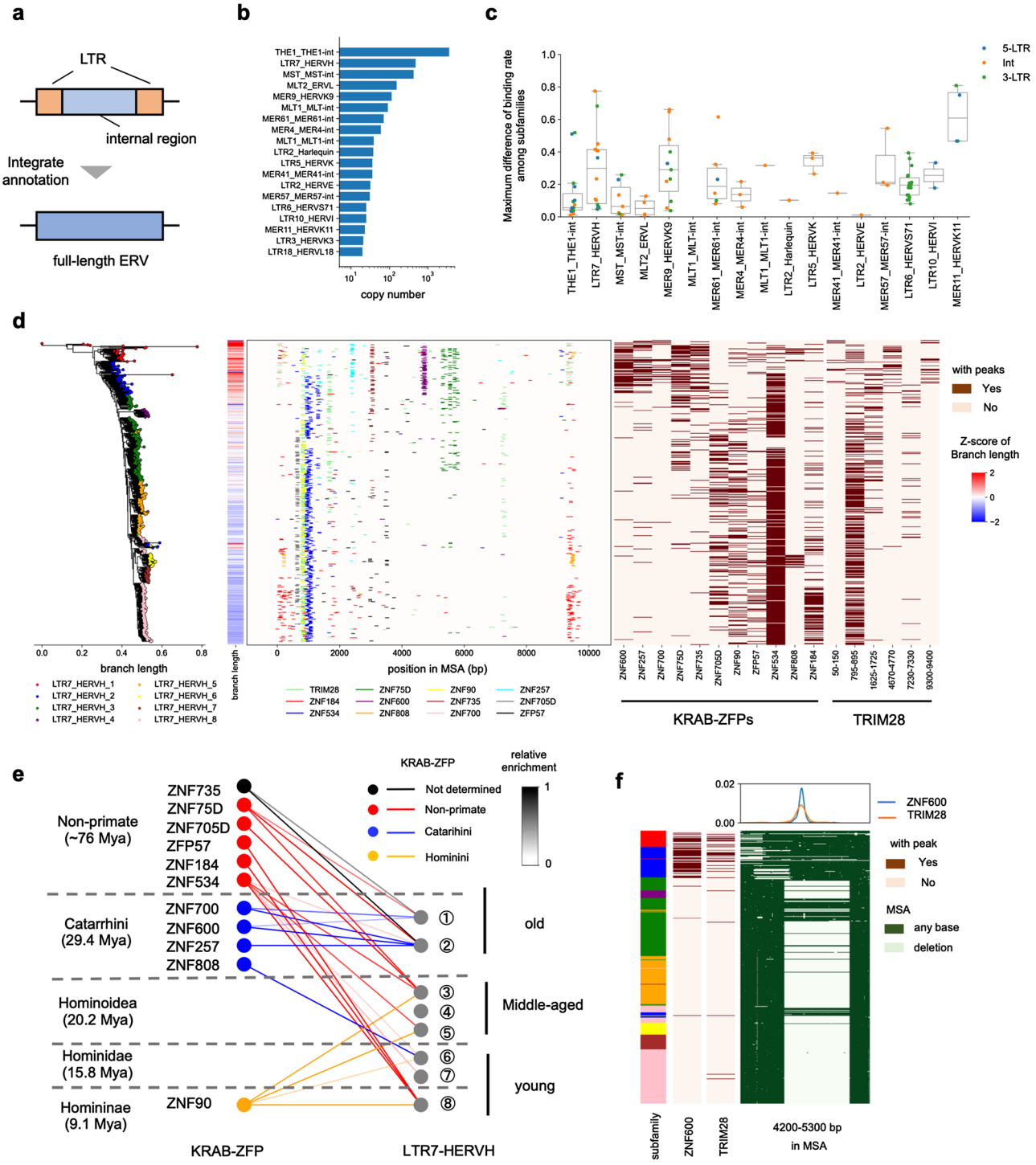

Fig. 4. Evolutionary arms race with KRAB-ZFPs in full-length ERVs and LTR7_HERVH. (**a**) Schematic representation of the reconstruction of full-length ERVs by integrating separately annotated LTR and internal region sequences. (**b**) The copy numbers of full-length ERVs in each ERV family. (**c**) Maximum difference in the binding rates of KRAB-ZFPs among subfamilies of each ERV family. Binding rates were estimated as the average proportion of ERV copies that overlapped with the KRAB-ZFP peaks. The maximum difference in the binding rate was defined as the difference between the maximum and minimum binding rates of the subfamilies in each KRAB-ZFP. The dotted colors indicate the region targeted by KRAB-ZFPs, including the 5'-LTR (blue), internal region (orange), and 3'-LTR (green). (**d**) Evolutionary arms race between LTR7_HERVH and 11 KRAB-ZFPs. The phylogenetic tree and the heatmap of the branch length on the left indicate the phyletic relationship and insertion date, respectively. Peaks and binding patterns of KRAB-ZFPs and TRIM28 are shown at the center and right, respectively. (**e**) Comparison of evolutionary ages between LTR7_HERVH subfamilies and KRAB-ZFPs. Dots indicate KRAB-ZFPs (left) and LTR7_HERVH subfamily (right). The lines between the dots show the relationship between the LTR7_HERVH subfamily and KRAB-ZFPs. The color and intensity of the lines indicate the evolutionary age of KRAB-ZFPs and the relative enrichment of their targets, respectively. (**f**) Deletion of the ZNF600 binding site. The color bar (left) shows the LTR7_HERVH linkage in each copy, based on the phylogenetic tree. The two heatmaps in the center indicate the binding of ZNF600 and TRIM28 between 4200 and 5300 bp, respectively. The top plot shows the positions of ZNF600 (blue) and TRIM28 (orange) peaks. The heatmap on the right shows the deletion of the sequence in each LTR7_HERVH copy. Light color indicates a deletion at each site.

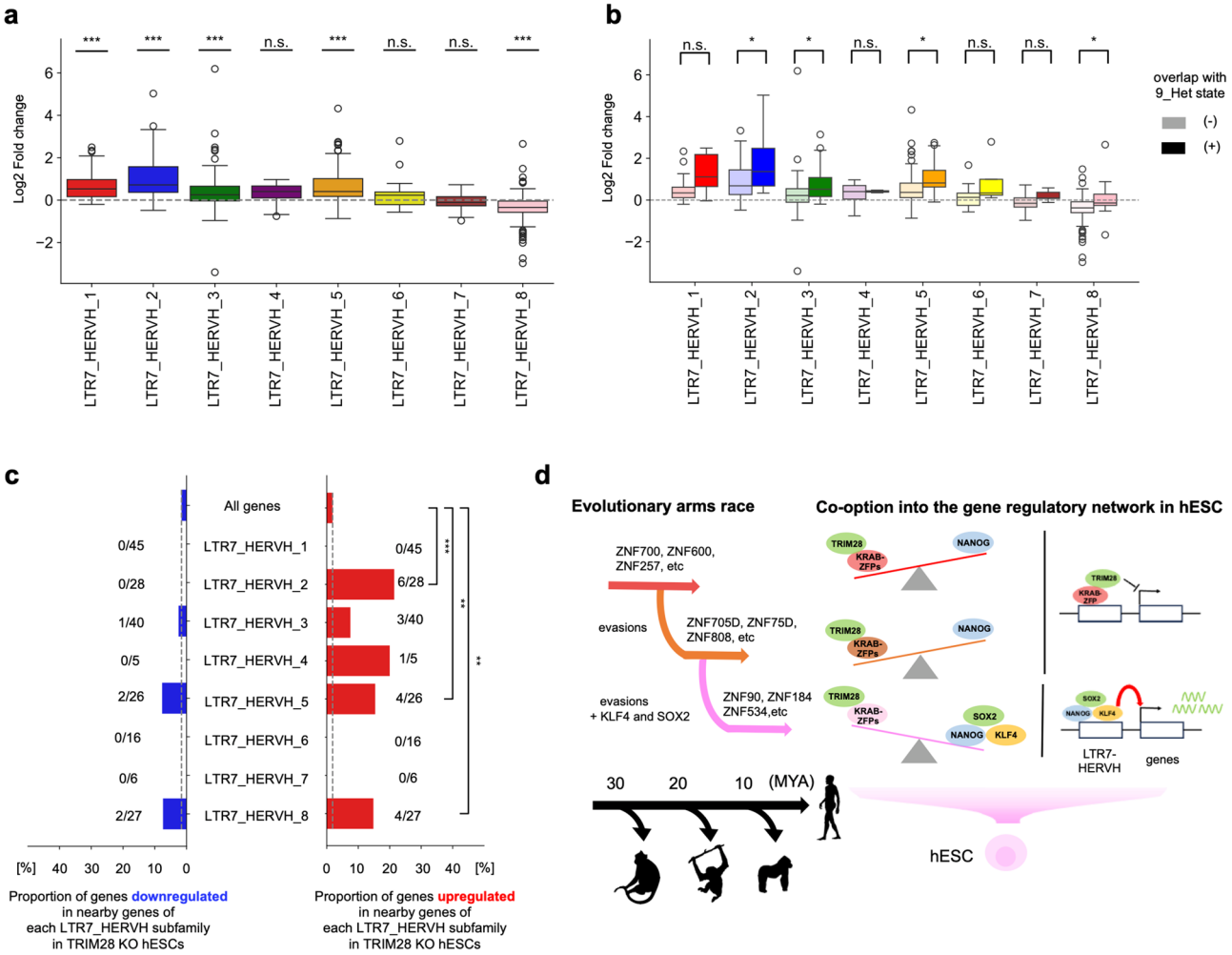

Fig. 5. TRIM28 regulates the expression of LTR7_HERVH and nearby genes in hESCs. (**a**) log2 fold change in the expression of LTR7_HERVH copies in TRIM28 knockout (KO) hESCs compared to wild-type (WT) hESCs for each LTR7_HERVH subfamily. Colors represent different subfamilies. Statistical analyses were performed using the two-sided Wilcoxon signed-rank test on the normalized read counts of LTR7_HERVH copies in wild-type and TRIM28 KO hESCs. The statistical significance indicates a tendency for LTR7_HERVH subfamily to be upregulated or downregulated due to TRIM28 depletion. P-values were adjusted for multiple testing using the Benjamini–Hochberg procedure ***FDR < 0.001; n.s., not significant. (**b**) Relationship between perturbation due to TRIM28 deficiency and the heterochromatin state (9 Het) of LTR7_HERVH. The intensity of the color indicates an overlap with the heterochromatin state. Statistical testing was conducted using the two-sided Mann–Whitney U test. *P*-values were adjusted using the Benjamini–Hochberg procedure. *FDR < 0.05. (**c**) Proportion of differentially expressed genes (DEG) in nearby genes of each LTR7_HERVH subfamily. Box plots on the left and right show the proportions of downregulated and upregulated protein-coding genes among the nearby genes, respectively. Statistical analyses were conducted using the two-sided binomial test and compared with the proportion of DEGs in all genes. *P*-values were adjusted using the Benjamini–Hochberg procedure. *FDR < 0.05, **FDR < 0.01, ***FDR < 0.001. (**d**) Schematic model illustrating the effect of the evolutionary arms race between LTR7_HERVH and KRAB-ZFPs (left) on the regulatory landscape of LTR7_HERVH and nearby genes in hESCs (right). As a result of the evolutionary arms race between LTR7_HERVH and KRAB-ZFPs, older LTR7_HERVH subfamilies are silenced by KRAB-ZFPs and TRIM28, whereas younger subfamilies are activated by pluripotent TFs. Silencing older subfamilies prevents the aberrant activation of nearby genes, whereas younger subfamilies function as cell type-specific regulatory elements.

and modulates LTR7_HERVH subfamilies and nearby genes, contributing to the gene regulatory networks in hESCs (Fig. 5d).

Discussion

In this study, we systematically investigated the evolutionary arms race relationship between TE families and KRAB-ZFPs (Fig. 1). We comprehensively identified the targets of KRAB-ZFPs and compared their evolutionary ages with those of their target TE subfamilies to characterize the chronological relationship between KRAB-ZFPs and TE subfamilies (Fig. 2). Furthermore, we identified and characterized novel instances of evolutionary

arms races in MER11, LTR5, LTR7_HERVH, and THE1_THE1-int (Fig. 3, 4 and Supplementary Fig. 7, 11). Remarkably, we also revealed that the evolutionary arms race in LTR7_HERVH was related to the regulation of LTR7_HERVH and nearby genes in hESCs (Fig. 5). In summary, our findings suggest that TEs, KRAB-ZFPs, and their host organisms have complex co-evolutionary relationships.

It has been hypothesized that KRAB-ZFPs have emerged to suppress newly emerging TE subfamilies[8,9,17]. Based on this hypothesis, we identified the TE targets of KRAB-ZFPs and obtained the evolutionary ages of both KRAB-ZFPs and TE subfamilies to re-examine the relationship between KRAB-ZFPs and TEs. Consistent with the results of previous studies, we observed that KRAB-ZFPs and their target TE subfamilies emerged concurrently (Fig. 2b). The expansion of KRAB-ZFPs and TE families occurred in the common ancestors of Eutherians, Simiiformes, and Catarrhines, re-emphasizing that KRAB-ZFPs engaged in evolutionary arms races through gene duplication during these periods (Supplementary Fig. 4d).

In addition to the gene duplication mechanisms discussed above, we observed that KRAB-ZFPs that emerged from the common ancestor with non-primates also targeted primate-specific TE subfamilies (Fig. 2b and Supplementary Fig. 4 h). While this phenomenon has been reported in the youngest L1, our observations suggest that it is more universal across TE classes (Supplementary Fig. 4f.)[8]. There are two possible explanations for this observation. First, ancient KRAB-ZFPs may have adapted to target and suppress newly emerged TE subfamilies. ZNF649, which emerged from the common ancestor of Eutherians, has been reported to have acquired new ZF domains and sequence specificity during the evolution of primates, enabling ZNF649 to suppress young L1 elements, thus supporting this hypothesis[28]. The second scenario is the opposite of the typical evolutionary arms race in which younger primate-specific TEs have evolved to be bound by KRAB-ZFPs. The uncontrolled expansion of TEs poses a threat to host survival and reproductive functions, suggesting a trade-off between TE proliferation and genetic persistence. Interactions with existing KRAB-ZFPs may have enabled the TEs to expand while maintaining host viability. However, these hypotheses require future research.

Next, we identified many candidates for evasion events of TE families from KRAB-ZFP suppression (Fig. 3a,b). In addition to the young L1 family, which has previously been reported to evade KRAB-ZFP, many other events have been observed in ERVs. The loss of KRAB-ZFP binding in younger subfamilies was attributed to substitutions and indels in the TE sequences, which supports evasion from KRAB-ZFPs over the evolution of TE families (Fig. 3d,e and Fig. 4f). As the deletion of KRAB-ZFP clusters in mice led to a slight increase in ERV insertions over the span of a few generations, these events could have potentially induced the expansion of ERVs over millions of years[16].

Moreover, we showed that several ERVs were targeted by new KRAB-ZFPs following their evasion from older KRAB-ZFPs (Figs. 3c and 4c). This retargeting by KRAB-ZFPs represents a complementary response to the emergence and expansion of young TE subfamilies that have evaded KRAB-ZFPs, suggesting the occurrence of evolutionary arms races between these ERVs and KRAB-ZFPs. It is important to note that, in many analyses, we limited ourselves to TE families that were active across multiple eras to examine the chronological relationship between the emergence of TE families and KRAB-ZFPs. In TE families that expanded over a short period, we were unable to examine TE events or evolutionary arms races. Nevertheless, our findings revealed that evolutionary arms races with KRAB-ZFPs occurred in more TE families than previously reported.

Additionally, we found that KRAB-ZFPs frequently bind not only to LTRs but also to internal regions (Supplementary Fig. 9a). While some KRAB-ZFPs have been reported to bind to the primer binding sites of ERVs, the binding sites of KRAB-ZFPs were found to be spread across the entire internal region (Fig. 4c and Supplementary Fig, 11a)[37]. Whether these KRAB-ZFP bindings suppress TEs will require further investigation, but since H3K9me3 induced by KRAB-ZFPs/TRIM28 can spread via HP1, it is possible that TEs are suppressed even when the binding sites of KRAB-ZFPs are distant from 5'-LTR[38]. Furthermore, we discovered that KRAB-ZFP bindings in the internal regions also varies among ERV subfamilies (Figs. 3b and 4b). Due to the higher sequence similarity in the internal regions compared to LTRs, subfamily classification is often absent even in TE families with many subfamilies, such as LTR7_HERVH. This finding implies that in order to fully understand the evolutionary arms race with KRAB-ZFPs and the co-option of ERVs, it is essential to conduct more detailed analyses using phylogenetic trees and subfamily classification methods.

Furthermore, we demonstrated that the expression differences of the LTR7_HERVH subfamily in hESCs were determined not only by transcription factors but also by KRAB-ZFPs/TRIM28. It has been suggested that the young subfamilies of LTR7_HERVH are specifically activated in hESCs because the transcription factors KLF4 and SOX2 bind specifically to these subfamilies[32]. However, our analysis showed that even older subfamilies could be activated in hESCs when the suppression by TRIM28 was inhibited (Fig. 5a,b). Additionally, our data suggest that young LTR7_HERVH subfamilies can be activated in hESCs despite being targeted by TRIM28 to the same extent as older subfamilies (Supplementary Fig. 12d). This phenomenon is likely due to KLF4 and SOX2, which bind to these young subfamilies and activate transcription, even within the heterochromatin, functioning as pioneer transcription factors[32,39]. In summary, our data suggested that the subfamily-specific expression patterns of LTR7_HERVH were formed by a balance between activation by transcription factors and suppression by KRAB-ZFPs/TRIM28 (Fig. 5d).

Finally, we investigated the effect of TRIM28 deficiency on the expression of genes near the derepressed LTR7_HERVH. Importantly, we discovered that the derepression of the older LTR7_HERVH subfamilies tended to promote the upregulation of nearby genes (Fig. 5c). This finding suggests that in the absence of TRIM28 suppression, the old LTR7_HERVH subfamilies can also act as cis-regulatory elements that influence nearby genes. Young LTR7_HERVH subfamilies, which are specifically activated in hESCs, play a crucial role in maintaining pluripotency in hESCs, likely through the regulation of neighboring gene expression. Overall, our data suggest that KRAB-ZFPs/TRIM28 modulates not only the expression patterns of LTR7_HERVH but also those of genes near the LTR7_HERVH copies (Fig. 5d).

This study has three limitations. The first was the inability to infer causality during the evasion of KRAB-ZFPs. Although our findings and those of previous studies support the concept of TE evasion, we have not examined whether the loss of KRAB-ZFP binding in younger subfamilies is due to random drift or selective pressure. In a future study, we need to examine whether evasion from KRAB-ZFPs accelerates the proliferation of TEs within the host genome using methods such as population analysis. The second limitation is the possibility that the binding patterns of human KRAB-ZFPs do not necessarily reflect those of their ancestors. Given the rapid evolution of KRAB-ZFP genes, it is possible that some ancestral KRAB-ZFPs were lost or optimized for co-option in the gene regulatory networks. Additionally, old TE families have lost KRAB-ZFP-binding sites because of the accumulation of mutations. This led to a decrease in KRAB-ZFP binding affinity, making it difficult to detect in analyses. Although the binding patterns of KRAB-ZFPs and TRIM28 in humans may retain aspects of their ancestral state, they may represent optimized or attenuated states. Third, the present study did not consider suppression mechanisms other than KRAB-ZFPs and TRIM28. Mechanisms that suppress TE transposition in host organisms are robust and complementary. For example, it has been reported that the youngest L1 is not repressed by TRIM28 but is instead repressed by DNA methylation associated with PIWI-interacting RNA[40]. Therefore, these mechanisms may complement the gaps in the evolutionary arms race with KRAB-ZFPs. Cross-species analyses that include various transposition suppression mechanisms are necessary to fully elucidate the evolutionary arms race with KRAB-ZFPs.

Despite these limitations, we have demonstrated that the evolutionary arms race between TEs and KRAB-ZFPs is not limited to LINE1 but is observed across a wide range of TEs and that such evolutionary arms races may have influenced the evolution of the host's gene expression regulatory networks. This study provides new insights into the complex co-evolutionary relationships among TE families, KRAB-ZFPs, and host gene regulatory networks.

Methods

Processing of ChIP-seq and ChIP-exo data

Raw reads from ChIP-seq and ChIP-exo were obtained from the NCBI GEO and Encyclopedia of DNA Elements (ENCODE) databases (Supplemental Table 1)[41,42]. The ChIP-seq dataset for KRAB-ZFPs included 1051 samples, encompassing 361 KRAB-ZFP genes, from five previous large-scale studies on KRAB-ZFPs (GSE58341, GSE76496, GSE78099, GSE120539, GSE200964)[8,9,17,43,44] and the ENCODE. The ChIP-seq data for TRIM28 were listed from ChIP-atlas and ENCODE[45]. The TRIM28 dataset consisted of 35 samples from 14 studies and the ENCODE. Datasets for TFs (NANOG, KLF4, SOX2, POU5F1, EOMES, FOXA1, FOXA2, and GATA6) in hESCs and differentiated cells were derived from GSE61475[33] and GSE130417[23]. Raw reads were trimmed using fastp (ver.0.2.3) and mapped to GRCh38/hg38 using Bowtie2 (ver.2.4.4) in an end-to-end --sensitive mode[46,47]. Unmapped reads were removed, and SAM files were converted to BAM using SAMtools (ver1.9)[48]. Multiple mapped reads were retained for analysis. PCR duplicates were removed only from ChIP-seq samples using Picard (ver.2.9.2) and SAMtools, because it is recommended to keep them in ChIP-exo samples[8,49]. Peak calling was performed using MACS2 (ver.2.2.7.1) with the options --keep-dup all -q 0.05[50]. The peaks were filtered using the following criteria: (1) does not overlap with the blacklist regions defined by the ENCODE (ENCFF419RSJ), (2) signal value > 2, (3) score > 50, and (4) peak length < 1000 bp. The ChIP-seq samples without input were also processed using this pipeline.

Subfamily classification and genome annotation of TE families and full-length ERVs

Genomic annotation data for the TE families were obtained using a two-step process (Supplementary Fig. 3). In the first step, subfamily classification was performed using the genomic annotation data of TEs. The GRCh38/hg38 genome was annotated using RepeatMasker with consensus sequences of the TE subfamilies downloaded from Dfam[51]. Fragmented repeats were reconstructed using OneCodeToFindThemAll.pl[52]. All extracted repeats were aligned to these consensus sequences using MAFFT (ver.7.520) with the options --addfragments --keeplength --retree 2 --mapout[53]. To filter out fragmented and truncated repeats, those that did not meet the following criteria were excluded from the analysis: alignment to the consensus sequence was less than 80% (or 60% if the consensus sequence length exceeded 4,000 bp); (2) more than 20% of the repeat were not aligned to the consensus sequence; and (3) the substitution rate exceeded 0.4 times per base pair. TE that met these criteria were defined as full-length copies. The integration of TE subfamilies into TE families followed the naming conventions for TE subfamilies[54]. Additionally, in accordance with Dfam, the 5′ end, orf, and 3′ end of LINEs were treated as different TE families, although they are parts of the full-length LINEs. TE families with more than 20 full-length sequences at the family level were aligned using MAFFT with the option --auto. To reduce the computational load on TE families with more than 10,000 full-length copies, 10,000 full-length copies were randomly sampled for analysis. Positions with more than 99% of the gaps in the multiple sequence alignment (MSA) were removed. Phylogenetic trees were constructed using iqtree2 with the options -m MFP -bb 1000 -alrt 1000[55–57]. The maximum likelihood distance matrix presented in the mldist file was dimensionally reduced using UMAP[58]. The latent space obtained from UMAP was clustered using K-means (k = 2–20). The optimal number of clusters was manually determined based on the number of clusters with the highest average silhouette coefficients. Of the 174 TE families without subfamily classification and with more than 20 full-length copies, new clustering was adopted for 59 families with an average silhouette coefficient exceeding 0.6; these clusters were defined as new subfamilies. From the alignment data, the consensus sequences for each subfamily were reconstructed based on the majority-rule consensus. If more than 50% of the bases were present, a specific base was used; otherwise, 'N' was assigned. The consensus sequences of the original family were replaced with those of the newly defined subfamilies.

Second, the GRCh38/hg38 genome was re-annotated using new consensus sequences. The process was performed using the same pipeline as in the first step. The full-length sequences were defined according to the first two criteria established in the first step. To eliminate the misannotation of Alu sequences, the third criterion was specifically applied to the SVA family. In all steps, repeats located on the Y chromosome, contigs, and unmapped regions (described later) were removed.

The genome annotation and subfamily classification of the MER11 family were constructed based on the annotation of 2nd step due to the heterogenicity of the MER11 subfamily highlighted by previous research and the redundancy of the annotation between the MER11A and MER11C subfamilies[6]. Duplications in the annotation of MER11 copies longer than 700 bp were examined using Bedtools intersect. Annotations of the MER11 copies that were longer and had higher scores were retained. A total of 2239 MER11 copies were used in the analyses. Phylogenetic tree construction and subfamily classification was performed using the pipeline described above.

The LTRs and internal regions constituting full-length ERVs were determined based on a previous study, Dfam and the naming conventions[26,54,59]. The LTRs and internal regions were assembled using OneCodeToFindThemAll.pl. The overlaps between the assembled annotations and the annotations obtained in the first step were examined using Bedtools intersect. Assembled sequences meeting the following criteria were defined as full-length: (1) overlap of one full-length internal region and two full-length LTRs; (2) both LTRs belonging to the same subfamily; and (3) the internal region flanked by the two LTRs. full-length ERV families were classified into subfamilies using the process outlined in the above subfamily classification pipeline.

Definition of KRAB-ZFP targets

The target of each KRAB-ZFP was determined using a binomial test, as described in a previous study[9]. Since KRAB-ZFPs also bind to repeat sequences other than TEs, such as satellites, rRNA, and tRNA, we included these repeats listed in Dfam in the enrichment analysis and extracted the KRAB-ZFPs that target TEs[8]. Overlaps between the summits of the KRAB-ZFP peaks and TE subfamilies were quantified using Bedtools intersect. The expected overlap probability was estimated by dividing the total length of each TE subfamily by the effective genomic length, which was defined as the total genome length, excluding the Y chromosome, contigs, and unmapped regions. Unmapped regions were inferred from 405 ChIP-seq samples. The coverage for each 100 bp bin was quantified using Bedtools coverage. Regions with zero coverage across all the ChIP-seq samples were defined as unmapped regions. The q value was calculated for each KRAB-ZFP using the Benjamini–Hochberg procedure. Targets were considered significant if the q value was < 0.05. A target was defined as the primary target if the quotient of the − log10 q value divided by the maximum value of the − log10 q value exceeded 0.9; otherwise, it was defined as a secondary target. In addition to the primary targets, secondary targets were included in the analysis if they met the following criteria: (1) − log10 q value > 10; (2) ratio > 2; (3) more than 5 peaks overlapping with full-length targets; and (4) peaks observed in more than 10% of the full-length targets. When a KRAB-ZFP targeted at least one subfamily belonging to a certain family, it was defined as a KRAB-ZFP targeting that TE family.

The full-length ERV targets of each KRAB-ZFP were determined for the 5'-LTR, internal region, and 3'-LTR. To match the copy number of the internal region, 5'-LTR and 3'-LTR were analyzed separately. Statistical testing was performed using the same pipeline as for the TEs, and KRAB-ZFP-ERV subfamily pairs that met the following criteria were used for analyses: (1) − log10 q value > 10; (2) ratio > 2; (3) more than 5 peaks overlapping with full-length targets; and (4) peaks observed in more than 10% of the full-length targets.

Evolutionary age inference of TE subfamilies, full-length ERVs, and KRAB-ZFPs

The evolutionary ages of the TE subfamilies and TE copies were estimated by liftover to the genomes of the 40 species (Supplementary Table 2). The chain files for the liftover were downloaded from UCSC[60]. All TE copies were lifted to other species using LiftOver -minMatch = 0.5. The evolutionary age of the TE copies was defined as the oldest era in which they could be lifted to at least three species, or more than half of the species within the same branch. The evolutionary age of TE subfamilies was defined as the oldest era in which more than 10% of the full-length copies first appeared. Because full-length ERVs can lose their internal region and one LTR due to homologous recombination, the evolutionary ages of full-length ERVs were defined using the evolutionary ages of the 5'-LTR rather than that of the full-length ERVs.

The evolutionary ages of the KRAB-ZFPs were obtained from two previous studies[8,9]. The evolutionary ages of KRAB-ZFPs were primarily determined based on the estimated ages of de Tribolet-Hardy et al[9]. Missing values were supplemented using data from Imbeault et al[8].

Screening of TE evasion from KRAB-ZFPs

For each significant combination of TE families and KRAB-ZFPs, differences in the binding rates of KRAB-ZFPs among all pairs of TE subfamilies were examined. The binding rate was calculated as the mean proportion of overlap in the ChIP-seq samples. Statistical testing was performed using Tukey's Honest Significant Difference test. A candidate for an evasion event of the TE family from KRAB-ZFP was defined as having at least one pair of subfamilies that met the following criteria: (1) adjusted p-value < 0.05, (2) difference in binding rates > 0.1, (3) decrease in binding rate in the younger subfamily, and (4) KRAB-ZFP being older than at least one TE subfamily. In Supplementary Data 4, the significant pair of the old TE subfamily and the young subfamily, which had a significant difference and the greatest variation in binding rates, is listed as a representative example.

Phylogenetic analyses of TE families
An unrooted phylogenetic tree was constructed using iqtree2 with the same options employed for subfamily classification. The oldest subfamily was defined as the subfamily with the oldest evolutionary age and the longest branch length. As the root of the phylogenetic tree, the copy that emerged during the evolutionary age of the oldest subfamily and had the longest branch length was selected.

MSA analyses of TE families
Consensus sequences of the TE subfamilies were aligned using MAFFT with the option --auto. All copies were aligned to the MSA of consensus sequences using MAFFT with the options --addfragments --keeplength --retree 2 --mapout. The positions of the ChIP-seq peaks in hg38 were converted to positions in MSA using map files.

De novo motif analyses and identification of KRAB-ZFP binding sites
For the de novo motif analysis, all peaks overlapping with TEs, satellites, and simple repeats were excluded. The top 500 peaks with the highest signal values were used. If there were fewer than 500 peaks, all the peaks that did not overlap with the repeats were used. Sequences 250 bp upstream and downstream of the peak summits were used as input. De novo motifs for each KRAB-ZFP experiment were generated using MEME (ver.5.3.0) with the options meme-chip -dna -minw 6 -maxw 30 -meme-nmotif 5 -meme-p 8 -meme-mod zoops[61]. The positions of the motifs within the peaks were identified using FIMO with options --thresh 0.0001 --no-qvalue[62]. Motif positions were aligned on the MSA. The discovery rate within the peaks was calculated for each motif position in MSA. Positions with a discovery rate of 50% or higher were defined as binding sites.

Processing of RNA-seq data and DEG analyses
Raw reads were downloaded from the GSE99215 dataset (Supplementary Table 3)[15]. The raw reads were trimmed using fastp and mapped to hg38 using STAR (ver. 2.7.8) with the options --outSAMtype BAM SortedByCoordinate --outFilterMultimapNmax 10,000,000 --outSAMmultNmax 1 --outMultimapperOrder Random[63]. Read counts of genes and LTR7_HERVH were separately counted using featureCounts (ver.2.0.1)[64]. Gene annotation of hg38 (GRCh38.13, release 40) was downloaded from GENCODE[65]. Differentially expressed gene (DEG) analyses between wild-type and TRIM28 KO hESCs were performed using DESeq2[66]. The read counts of LTR7_HERVH were normalized using the size factor estimated using DESeq2.

Nearby genes were defined as genes whose transcription start sites (TSS) were within 50 kb of the LTR7_HERVH copies because previous studies have reported that DEGs were enriched within 50 kb of perturbed TE copies[23]. The TSSs of the genes were extracted from the gene annotation. The distance between the gene and the LTR7_HERVH copy was defined as the shortest distance from either end of the LTR7_HERVH copy to the TSS. If the TSS overlapped with the LTR7_HERVH copies, the distance was defined as zero. To avoid the effects of pseudogenes, genes annotated as "protein_coding," were used for analyses.

Chromatin state analyses with ChromHMM
The chromatin state of the hESCs (E003) was downloaded from Roadmap Epigenetics[34,35]. A Core15-state model lifted to hg38 was used. The overlap between chromatin states and LTR7_HERVH copies was obtained using bedtools intersect.

Statistical analyses
Statistical analyses were performed using SciPy (ver.1.12.0) and statsmodels (ver.0.13.2)[67,68]. Multiple testing was performed using the Benjamini–Hochberg procedure.

Data availability
The accession codes for the data used in the analysis are listed in the Supplementary Table. The computational codes and processed data for reproduction are publicly available on GitHub (https://github.com/hmdlab/Evolutionary-Arms-Race). For further needs contact the corresponding author.

Received: 24 July 2024; Accepted: 20 September 2024
Published online: 07 October 2024

References
1. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* **10**, 691–703 (2009).
2. Ito, J. *et al.* Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. *PLoS Genet.* **13**, e1006883 (2017).
3. Garcia-Perez, J. L., Widmann, T. J. & Adams, I. R. The impact of transposable elements on mammalian development. *Development* **143**, 4101–4114 (2016).
4. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: From conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).
5. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083–1087 (2016).
6. Chen, X. *et al.* Cryptic endogenous retrovirus subfamilies in the primate lineage. *bioRxiv*, 2023.2012. 2007.570592 (2023).
7. Ecco, G., Imbeault, M. & Trono, D. KRAB zinc finger proteins. *Development* **144**, 2719–2729 (2017).
8. Imbeault, M., Helleboid, P.-Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550–554 (2017).
9. de Tribolet-Hardy, J. *et al.* Genetic features and genomic targets of human KRAB-zinc finger proteins. *Genome Res.* **33**, 1409–1423 (2023).

10. Huntley, S. *et al.* A comprehensive catalog of human KRAB-associated zinc finger genes: Insights into the evolutionary history of a large family of transcriptional repressors. *Genome Res.* **16**, 669–677 (2006).
11. Schultz, D. C., Ayyanathan, K., Negorev, D., Maul, G. G. & Rauscher, F. J. SETDB1: A novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. *Genes Dev.* **16**, 919–932 (2002).
12. Schultz, D. C., Friedman, J. R. & Rauscher, F. J. Targeting histone deacetylase complexes via KRAB-zinc finger proteins: The PHD and bromodomains of KAP-1 form a cooperative unit that recruits a novel isoform of the Mi-2α subunit of NuRD. *Genes Dev.* **15**, 428–443 (2001).
13. Turelli, P. *et al.* Interplay of TRIM28 and DNA methylation in controlling human endogenous retroelements. *Genome Res.* **24**, 1260–1270 (2014).
14. Haring, N. L. *et al.* ZNF91 deletion in human embryonic stem cells leads to ectopic activation of SVA retrotransposons and up-regulation of KRAB zinc finger gene clusters. *Genome Res.* **31**, 551–563 (2021).
15. Tao, Y. *et al.* TRIM28-regulated transposon repression is required for human germline competency and not primed or naive human pluripotency. *Stem Cell Rep.* **10**, 243–256 (2018).
16. Wolf, G. *et al.* KRAB-zinc finger protein gene expansion in response to active retrotransposons in the murine lineage. *Elife* **9**, e56337 (2020).
17. Najafabadi, H. S. *et al.* C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat. Biotechnol.* **33**, 555–562 (2015).
18. Jacobs, F. M. *et al.* An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature* **516**, 242–245 (2014).
19. Trono, D. *Cold Spring Harbor Symposia on Quantitative Biology.* 281–288 (Cold Spring Harbor Laboratory Press).
20. Yang, P., Wang, Y. & Macfarlan, T. S. The role of KRAB-ZFPs in transposable element repression and mammalian evolution. *Trends Genet.* **33**, 871–881 (2017).
21. Ito, J. *et al.* Endogenous retroviruses drive KRAB zinc-finger protein family expression for tumor suppression. *Sci. Adv.* **6**, eabc3020 (2020).
22. Turelli, P. *et al.* Primate-restricted KRAB zinc finger proteins and target retrotransposons control gene expression in human neurons. *Sci. Adv.* **6**, eaba3200 (2020).
23. Pontis, J. *et al.* Hominoid-specific transposable elements and KZFPs facilitate human embryonic genome activation and control transcription in naive human ESCs. *Cell Stem Cell* **24**, 724–735 (2019).
24. Brattås, P. L. *et al.* TRIM28 controls a gene regulatory network based on endogenous retroviruses in human neural progenitor cells. *Cell Rep.* **18**, 1–11 (2017).
25. Rowe, H. M. *et al.* TRIM28 repression of retrotransposon-based enhancers is necessary to preserve transcriptional dynamics in embryonic stem cells. *Genome Res.* **23**, 452–461 (2013).
26. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA* **12**, 1–14 (2021).
27. Fukuda, K. *et al.* Potential role of KRAB-ZFP binding and transcriptional states on DNA methylation of retroelements in human male germ cells. *elife* **11**, e76822 (2022).
28. Fernandes, J. D. *et al.* KRAB zinc finger proteins coordinate across evolutionary time scales to battle retroelements. *bioRxiv*, 429563 (2018).
29. Jern, P. & Coffin, J. M. Effects of retroviruses on host genome function. *Ann. Rev. Genet.* **42**, 709–732 (2008).
30. Lu, X. *et al.* The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat. Struct. Mol. Biol.* **21**, 423–425 (2014).
31. Wang, J. *et al.* Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature* **516**, 405–409 (2014).
32. Carter, T. A. *et al.* Mosaic cis-regulatory evolution drives transcriptional partitioning of HERVH endogenous retrovirus in the human embryo. *Elife* **11**, e76257 (2022).
33. Tsankov, A. M. *et al.* Transcription factor binding dynamics during human ES cell differentiation. *Nature* **518**, 344–349 (2015).
34. Bernstein, B. E. *et al.* The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).
35. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
36. Ohnuki, M. *et al.* Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. *Proc. Natl. Acad. Sci.* **111**, 12426–12431 (2014).
37. Yang, B. *et al.* Species-specific KRAB-ZFPs function as repressors of retroviruses by targeting PBS regions. *Proc. Natl. Acad. Sci.* **119**, e2119415119 (2022).
38. Groner, A. C. *et al.* KRAB–zinc finger proteins and KAP1 can mediate long-range transcriptional repression through heterochromatin spreading. *PLoS Genet.* **6**, e1000869 (2010).
39. Soufi, A., Donahue, G. & Zaret, K. S. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. *Cell* **151**, 994–1004 (2012).
40. Castro-Diaz, N. *et al.* Evolutionally dynamic L1 regulation in embryonic stem cells. *Genes Dev.* **28**, 1397–1409 (2014).
41. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (2012).
42. Luo, Y. *et al.* New developments on the encyclopedia of DNA elements (ENCODE) data portal. *Nucleic Acids Res.* **48**, D882–D889 (2020).
43. Schmitges, F. W. *et al.* Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome Res.* **26**, 1742–1752 (2016).
44. Helleboid, P. Y. *et al.* The interactome of KRAB zinc finger proteins reveals the evolutionary history of their functional diversification. *EMBO J.* **38**, e101220 (2019).
45. Zou, Z., Ohta, T., Miura, F. & Oki, S. ChIP-Atlas 2021 update: A data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data. *Nucleic Acids Res.* **50**, W175–W182 (2022).
46. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
47. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
48. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
49. Broad Institute, Picard Toolkit. *Broad Institute, Github Repository.* http://broadinstitute.github.io/picard/ (2018)
50. Liu, T. Use model-based analysis of ChIP-Seq (MACS) to analyze short reads generated by sequencing protein–DNA interactions in embryonic stem cells. In *Stem Cell Transcriptional Networks: Methods and Protocols* (ed. Kidder, B. L.) 81–95 (Springer, 2014).
51. Smit, A., Hubley, R & Green, P. *RepeatMasker Open-4.0.* http://www.repeatmasker.org (2013–2015).
52. Bailly-Bechet, M., Haudry, A. & Lerat, E. "One code to find them all": A perl tool to conveniently parse RepeatMasker output files. *Mob. DNA* **5**, 1–15 (2014).
53. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evolut.* **30**, 772–780 (2013).
54. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).
55. Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evolut.* **37**, 1530–1534 (2020).
56. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., Von Haeseler, A. & Jermiin, L. S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

57. Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evolut.* **35**, 518–522 (2018).
58. McInnes, L., Healy, J. & Melville, J. Umap *Uniform manifold approximation and projection for dimension reduction*. arXiv:1802.03426 (2018).
59. Kojima, K. K. Human transposable elements in Repbase: Genomic footprints from fish to humans. *Mobile DNA* **9**, 2 (2018).
60. Hinrichs, A. S. *et al.* The UCSC genome browser database: Update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
61. Machanick, P. & Bailey, T. L. MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696–1697 (2011).
62. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
63. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
64. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
65. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).
66. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).
67. Virtanen, P. *et al.* SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
68. Seabold, S. & Perktold, J. Statsmodels: Econometric and statistical modeling with python. *SciPy* **7**, 1 (2010).

Acknowledgements

We thank the members of Trono lab, Laboratory of Virology and Genetics at EPFL, for their discussions and insights that enriched our research. We also thank Junna Kawasaki for meaningful comments regarding this study. We thank all contributors to the public databases and data used in this study. We also extend our gratitude to all contributors who developed and maintained the bioinformatics tools and packages utilized in the present study. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics. We would also like to thank PHYLOPIC (https://www.phylopic.org/) and the contributors for providing the animal silhouette illustrations used in the figures of this study under the CC0 license. This study was supported in part by JST SPRING (JPMJSP2128, to Masato Kosuge) and MEXT/JSPS KAKENHI (Grant Numbers JP24K21326, JP23H00509, JP22H04925, JP20H00624 to Michiaki Hamada).

Author contributions

M.K. performed all bioinformatics analyses. M.K., J.I., and M.H. designed the analyses and interpreted the results. M.K., J.I., and M.H. wrote the original manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-73752-7.

Correspondence and requests for materials should be addressed to M.H.