Multi-species analysis of inflammatory response elements reveals ancient and lineage-specific contributions of transposable elements to NF-κB binding

Liangxi Wang[1,2], Azad Alizada[3], Kumaragurubaran Rathnakumar[2,4,12], Nadiya Khyzha[5], Tiegh Taylor[6], Laura F Campitelli[1,7], Zain M Patel[1,7], Lina Antounians[8,9], Timothy Hughes[1,7], Sushmita Roy[10], Jennifer A Mitchell[6], Jason E Fish[11,12,13] and Michael D Wilson[1,2]

1 Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada
2 Genetics and Genome Biology, Sickkids Research Institute, Toronto, ON, Canada
3 Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK
4 Present address: Division of Developmental Biology, Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA
5 Basic Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA, USA
6 Department of Cell and Systems Biology, University of Toronto, Toronto, ON, Canada
7 Donnelly Center for Cellular and Biomolecular Research, Toronto, ON, Canada
8 Developmental and Stem Cell Biology, Sickkids Research Institute, Toronto, ON, Canada.
9 Division of General and Thoracic Surgery, The Hospital for Sick Children, Toronto, ON, Canada
10 Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA
11 Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada
12 Toronto General Hospital Research Institute, University Health Network, Toronto, ON, Canada
13 Peter Munk Cardiac Centre, University Health Network, Toronto, ON, Canada

ABSTRACT

Nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB) is an essential and evolutionarily conserved transcription factor complex primarily involved in innate immunity and inflammation. Transposable elements (TEs) can be co-opted to innovate immune transcriptional regulatory networks; however, the extent to which TEs have contributed to the modulation of NF-κB response in different mammalian lineages is not well established. Here we performed a multi-species analysis of TEs bound by the NF-κB subunit RELA (p65) in response to the pro-inflammatory cytokine TNFα (Tumor Necrosis Factor alpha). Using endothelial cell RELA ChIP-seq data from human, mouse and cow, we found that 55 TE subfamilies were enriched within NF-κB bound regions. These RELA-bound transposons possess multiple active epigenetic features and reside near TNFα-induced genes. A prominent example of lineage-specific contribution of transposons comes from the bovine SINE subfamilies Bov-tA1/2/3 which collectively contributed over 14,000 NF-κB bound regions in cow. By comparing NF-κB binding data across species, we found several examples of NF-κB motif-bearing TEs that appeared to colonize the genome prior to the divergence of the selected mammals, including a DNA transposon MER81, whose ancestral sequence contains two intact RELA motifs. We demonstrate that one NF-κB bound MER81 element can control the TNFα-induced expression of *INFGR2* (Interferon Gamma Receptor 2) in human. Lastly, the presence of RELA motifs within MER81 elements appeared to stabilize during human evolution, indicative of purifying selection acting on a subset of these NF-κB bound ancient DNA transposons. Taken together, our results implicate multiple transposons in establishing NF-κB mediated regulatory networks during mammalian evolution.

INTRODUCTION

Transposable elements (TEs) can act as a dynamic and powerful force to drive the evolution of gene regulation (Chuong et al. 2017; Sundaram and Wysocka 2020; Senft and Macfarlan 2021; Fueyo et al. 2022). One prevalent mechanism by which TEs can influence gene regulation is through exaptation into cis-regulatory elements (CREs). Indeed an increasing number of studies have clearly demonstrated how TE-derived cis-regulatory elements impact gene regulatory networks involved in development and innate immunity (Lynch et al. 2011; Notwell et al. 2015; Chuong et al. 2016; Todd et al. 2019; Bogdan et al. 2020; Kelly et al. 2022). From a sequence perspective, the co-option of TEs can be largely explained by transcription factor (TF) binding motifs they contain. Understanding which transcription factors interact with TEs under a given condition is an ongoing challenge in genome biology (Bourque et al. 2008; Macfarlan et al. 2011; Emera and Wagner 2012; Sundaram et al. 2014; Ito et al. 2020).

Nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB) is an essential mediator of innate immune/inflammatory responses (Zhang et al. 2017), whose activity has been conserved across metazoans (Sen and Baltimore 1986; Satriano and Schlondorff 1994; Correa et al. 2004; Hetru and Hoffmann 2009). In vertebrates, NF-κB is composed of hetero-dimers of five proteins: NFKB1 (p50), NFKB2 (p52), RELA (p65), RELB and REL, with the most predominant combinations being NFKB1:RELA which can be found in most cell types (Bhatt and Ghosh 2014). The NFKB1:RELA complex mediates canonical NF-κB signaling pathways. This process relies on NFKB1:RELA being sequestered in its inactive form in the cytoplasm through its association with the inhibitory IκB proteins under unstimulated conditions. In response to upstream stimuli, such as a pro-inflammatory cytokine, the IκB inhibitors are degraded and NF-κB rapidly translocates into the nucleus where it binds proximal and distal cis-regulatory elements that drive the expression of multiple innate immunity/inflammation related genes (Zhang et al. 2017).

While the components of NF-κB are typically conserved across cell types, the binding of NF-κB to the genome is specialized by chromatin context (aka the epigenome) (Natoli et al. 2005; Ostuni et al. 2013). For example, abundant NF-κB binding occurs at the pre-existing SPI1 (PU.1) transcription factor bound regions in macrophages (Kaikkonen et al. 2013) whereas NF-κB binding in aortic endothelial cells occurs at ERG and JUN bound regions (Hogan et al. 2017). In addition to comparisons of NF-κB binding across cell types, comparing its genome wide binding across species can reveal NF-κB regulatory logic and identify key CREs that drive NF-κB target gene expression. For example, an inter-species comparison of NF-κB binding in aortic endothelial cells found that bound regions that occur at orthologous locations between species possess multiple characteristics of strong NF-κB enhancers including multiple NF-κB motifs (Alizada et al. 2021). Furthermore, this comparative epigenomic analysis of NF-κB showed that conserved orthologous NF-κB bound regions are active across diverse cell types and overlap non-coding genetic variation associated with both inflammatory and cardiovascular phenotypes.

While this previous work focused on the global properties of NF-κB binding between species, it did not address the extent to which TEs, the majority of which are lineage specific, contributed to NF-κB binding.

Here we characterized the contribution of TEs to NF-κB bound regions in human, mouse and cow endothelial cells following stimulation with the potent pro-inflammatory cytokine, Tumor Necrosis Factor alpha (TNFα) (Alizada et al. 2021). Among these three species, we found 55 TE subfamilies to be overrepresented within the RELA-bound regions and these RELA-bound TEs were associated with active epigenetic signatures. The majority of TE contributions to RELA binding were species-specific, with the most notable example being the bovine-specific SINE subfamilies Bov-tA1/2/3 which harbor RELA motifs and contribute over 14,000 RELA binding events. We also found several TE subfamily-derived RELA binding shared across mammals, one of which is the mammalian DNA transposon MER81. Suggestive of a functional role of MER81 in the NF-κB response, we demonstrate that one distal MER81 instance interacts with and is required for the activation of *IFNGR2*. Lastly, using ancestral genome reconstructions, we infer a constant MER81 copy number in combination with the stabilization of RELA motif occurrences during the last ~100 million years of human evolution. Overall, our results implicate the potential involvement of multiple transposons in establishing NF-κB mediated regulatory networks and highlight the functional contribution of an ancient DNA transposon during mammalian evolution.

RESULTS

Transposable elements contribute to mammalian NF-κB bound regions.

To identify the contribution of transposable elements to NF-κB bound regions relative to other transcription factors, we first interrogated genome-wide binding data for 435 human and 266 mouse TFs found in the Gene Transcription Regulation Database (GTRD) (Kolmykov et al. 2021). The GTRD database contains a wide variety of uniformly processed, cumulative ChIP-seq data across many contexts and cell types, including 47 datasets from 10 cell types for the NF-κB subunit RELA (Kolmykov et al. 2021). We calculated the fraction of TF peak summits that overlap RepeatMasker annotated TEs (Bao et al. 2015) for each TF. While transcription factor bound regions are globally depleted for TEs (median TE overlap was 25% for human and 18% for mouse), TE-derived RELA bound regions was above average of all transcription factors (human: 84th percentile; mouse: 77th percentile), with a median overlap of 33% and 21% for human and mouse respectively (Supplementary Figure 1A). This result suggests that indeed TEs have contributed substantially to human and mouse NF-κB binding.

To investigate the evolutionary dynamics of TE-derived RELA binding between species, we then focused our analysis on our previously published study that mapped RELA binding by ChIP-seq in human, mouse and cow (Alizada et al. 2021). This comparative epigenomics study was performed in primary aortic endothelial cells treated acutely (45 min) with a species-

matched pro-inflammatory cytokine tumor necrosis factor alpha (TNFα). This dataset gave a robust number of RELA bound regions for each species (human: 59,735; mouse: 23,623; and cow: 93,511) after 45 minutes of TNFα treatment (Alizada et al. 2021). We reasoned that since this dataset also contained matched chromatin accessibility and post translational histone modification data, it was an ideal system for dissecting the contribution of ancestral and lineage-specific TE-derived RELA binding (Figure 1).

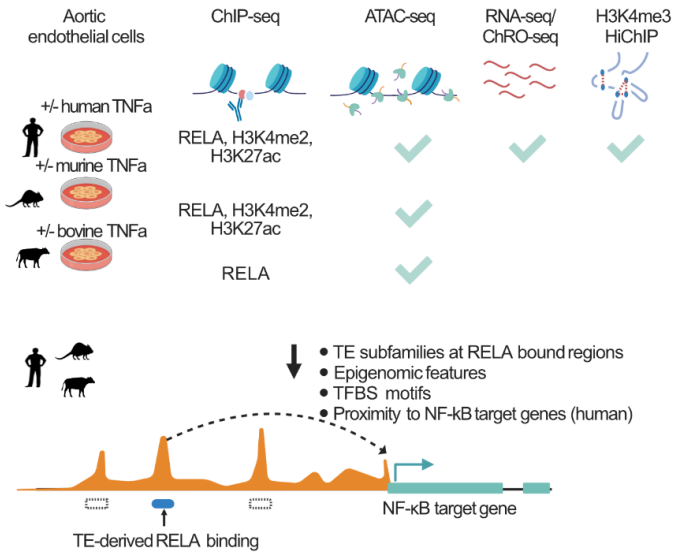

Figure 1. Schematic of the experimental data and computational workflow. Primary aortic endothelial cells from three mammalian species were cultured with and without 45-minute treatment of the pro-inflammatory cytokine TNFα. These cells were subjected to chromatin immunoprecipitation followed by DNA sequencing (ChIP-seq) assays using antibodies against RELA, H3K4me2 and H3K27ac. Chromatin accessibility was assessed using ATAC-seq. Immortalized human aortic endothelial cells (TeloHAEC) were used for ChRO-seq and H3K4me3 HiChIP. Here we performed H3K4me3 HiChIP to identify active enhancer-promoter connections in human. Transposable element derived RELA-bound regions were computationally. In particular, the contribution of transposable elements was examined across three species and four cell types. These elements exhibit unique sequence and epigenetic features and functionally associate with target gene expression.

We observed a robust number, albeit a slightly lower overlapping fraction of TEs within RELA-bound regions in human (19%) and mouse (12%) aortic endothelial cells compared to the GTRD database. Nonetheless, we also observed a striking fraction (~31%) of RELA bound regions in cow to be TE derived (Figure 2A). The proportions of the four major TE classes (LINE, SINE, DNA transposon, LTR) that overlap RELA peak summits were also significantly different from background expectations for each species (defined by the genome coverage of each TE class; chi-squared test, corrected p values $< 1\times10^{-7}$ for all tests; Figure 2A). The most striking example of this difference was for the SINEs in cow, which overlapped 20,188 RELA peaks, at a frequency of 22% (genome background of 17%; chi-squared test, 1.3-fold enriched, corrected p value= 2.6×10^{-267}). In terms of the relative composition change of TE-derived sequences, we observed that the LTR class was close to two-fold enriched within RELA-bound regions for both human and mouse, but not in cow (chi-squared test; human: 1.8-fold enriched, p

value< 2×10^{-308}; mouse: 1.8-fold enriched, p value< 2×10^{-308}; cow: 0.99-fold enriched, p value=0.53. Supplementary Figure 2A).

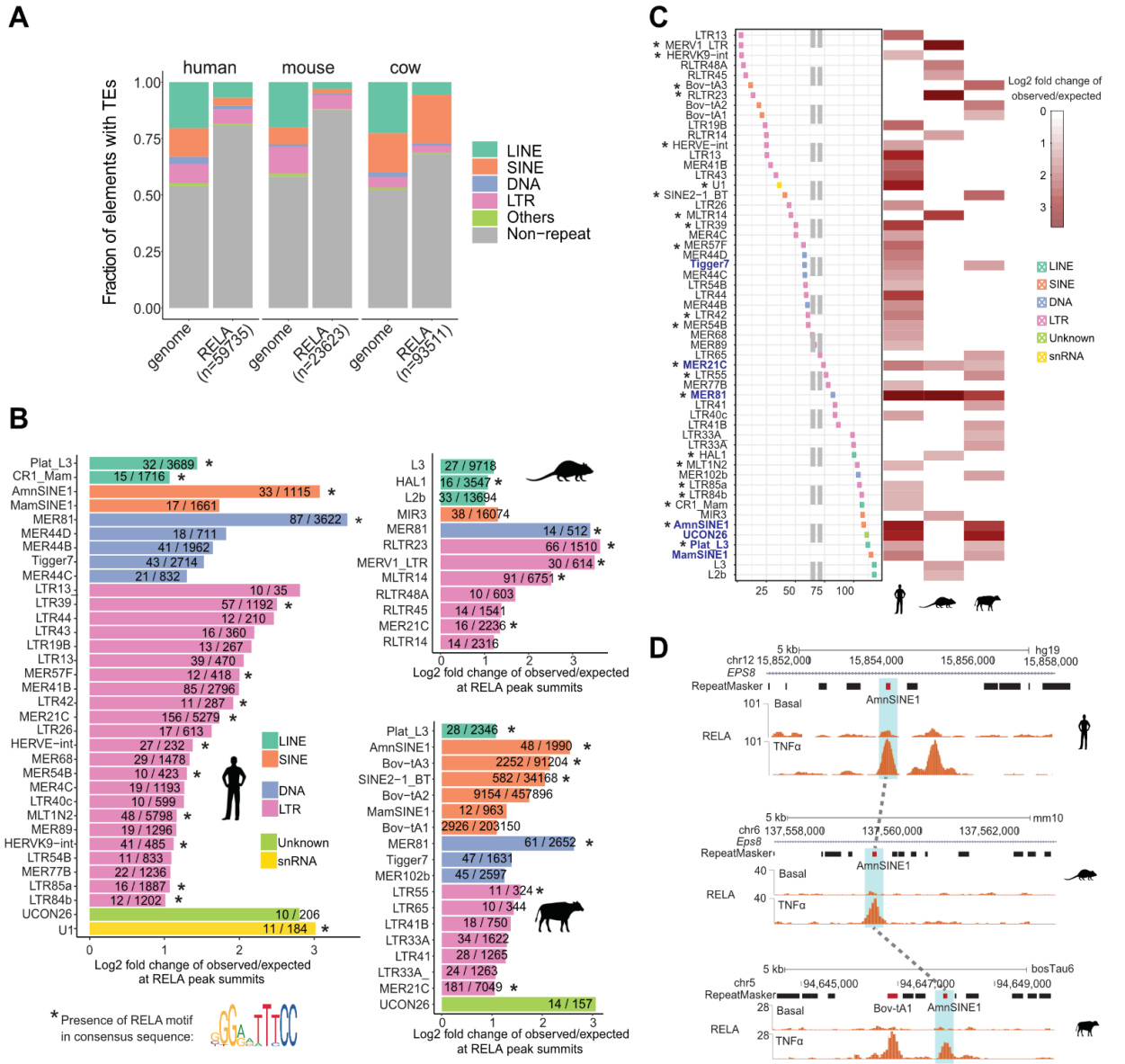

Figure 2. Transposable elements contribute to mammalian NF-κB bound regions. RELA ChIP-seq data with/without TNFα treatment for 45 min from human/mouse/cow aortic endothelial cells were used to characterize RELA-bound regions in the three species. A) The overlap between RELA-bound regions and four major TE classes. RELA bound regions/peaks were identified by inputting both stimulated and unstimulated data via MACS2 (Zhang et al. 2008) for each species. We define an overlap if a RELA peak summit intersects with a TE. Genome coverage of TE classes in each species serves as a reference. B) TE subfamilies that were significantly enriched within RELA-bound regions in human, mouse and cow. The enrichment values were calculated by comparing observed and expected/simulated overlap events at RELA peak summits. The cutoffs to define enriched TE subfamilies: FDR < 0.05, log2 fold change > 1 and the number of observed overlap events >= 10. Observed overlap number/ total number of each TE subfamily is displayed for each bar. Bars marked with an asterisk indicate the presence of canonical RELA motifs in the consensus sequence. C) Estimated evolutionary age of all the enriched TE

subfamilies. The left panel depicts the median of estimated age of these elements, while the right panel presents enrichment values. Dashed gray lines represent estimated human-mouse and mouse-cow divergence time obtained from TIMETREE database (Kumar et al. 2017). Bars marked with an asterisk indicate the presence of canonical RELA motifs in the consensus sequence. TE subfamilies shared by more than one species are highlighted in blue. D) An example (AmnSINE1) of a three-species-conserved TE-derived RELA-bound region. Red block from RepeatMasker track represents RELA-bound TE and light blue shades show orthologous regions of the three species.

An important aspect of our analysis is that we incorporated ChIP-seq data from multiple species. Using global multiple sequence alignments, we previously identified conserved orthologous RELA binding (Alizada et al. 2021). By applying the same framework to TE-derived RELA binding, we found the vast majority of TE-derived RELA-bound regions to be species-specific (chi-squared test; human: 90%, mouse: 92%, cow: 95%; corrected p values $<3.6\times10^{-207}$ for all three species; Supplementary Figure 2B). While we would expect the minority of TE-derived transcription factor bound regions to be shared across species, it was unclear whether the counterpart of a TE-derived conserved orthologous RELA binding in another species would also be TE-derived. Interestingly, the majority of conserved orthologous TE-derived RELA-bound regions were only TE-derived in one species (human: 901/1122; mouse:143/273 and cow: 972/1479). This observation suggests that these TE-derived RELA bound regions may represent cases where conserved RELA binding turned over into adjacent TE-derived DNA sequences with RELA binding ability.

To gain more insight into the specific transposon features that contributed to TE-derived RELA-bound regions, we then annotated each RELA peak summit with its corresponding TE subfamily using RepeatMasker (see Methods for details). By calculating the statistical enrichment of these overlapped TEs, we identified 55 significantly enriched TE subfamilies among all three species, 7 of which were shared in at least one additional species (Figure 2B). These enriched subfamilies collectively make up 9.0% (1020/11365), 13% (369/2903) and 52% (15475/29576) of TE-derived RELA-bound regions in human, mouse and cow, respectively. Notably, 22 subfamilies were detected with the presence of canonical RELA motifs in their consensus sequences (Figure 2B). For the 55 identified subfamilies, members from LTRs account for the majority of TE-derived RELA-bound regions in both human and mouse (human: 68%, mouse: 65%). Only one of the 35 enriched LTRs, MER21C, is shared across more than one species which is in line with the documented contribution of LTRs to lineage-specific evolutionary processes (Chuong et al. 2013, 2017; Jacques et al. 2013; Sundaram et al. 2014; Trizzino et al. 2017). An interesting example of a lineage-specific LTR subfamily is MER41B, a dominant human LTR subfamily previously reported by Chuong *et al.* 2016 to propagate STAT1 and IRF1 binding sites to regulate interferon signaling. The expanded function of MER41B to NF-κB signaling could indicate an even more widespread role of this subfamily in shaping immune/inflammatory regulatory landscapes during human evolution. In contrast, the SINE class of TEs makes up 68% of TE-derived RELA-bound regions in the cow genome. Strikingly four SINE subfamilies Bov-tA1, Bov-tA2, Bov-tA3, and SINE2-1_BT account for the majority (14,914 or 74%) of these RELA-bound SINEs. The presence of a RELA motif in the consensus

sequence of Bov-tA3, which introduced 2,252 RELA binding events in the bovine genome, provides a plausible explanation for these SINE-derived RELA-bound regions (Figure 2B).

We next estimated the evolutionary age of the 55 enriched TE subfamilies using the sequence divergence of subfamily instances relative to their consensus. We found these elements span a large range of evolutionary time that encompasses the divergence of the three selected mammals (~25-140 million years ago (MYA); Figure 2C). Given the estimated age and recurrent enrichment signals across species, we defined a conserved repertoire of seven ancient subfamilies (Tigger7, MER21C, MER81, AmnSINE1, UCON26, Plat_L3, MamSINE1), which likely colonized the mammalian genome near or prior to the divergence of the three species (~90 MYA; Figure 2C). The unknown subfamily UCON26, AmnSINE1, and MER81 differentiated themselves from the other ancient elements with high cross-species enrichment signals (UCON26: human and cow; AmnSINE1: human and cow; MER81: human, mouse, and cow). AmnSINE1 is presumed to have colonized the mammalian genome before the radiation of amniote animals (Nishihara et al. 2006) and has contributed to conserved gene regulatory networks impacting brain development (Sasaki et al. 2008; Hirakawa et al. 2009). While AmnSINE1 was not considered to be a significantly enriched subfamily in mouse we found 9 copies bound by RELA with 4.6-fold enrichment (one example is shown in Figure 2D). MER81 is a 114 bp non-autonomous DNA transposon derived from MER91 sequence (Bao et al. 2015). MER81 was strongly enriched for RELA binding in all three species (10.9-fold change in human, 10.6-fold change in mouse and 6.2-fold change in cow). In summary, our cross-species analysis has revealed a substantial contribution of multiple lineage-specific and ancient TEs to mammalian NF-κB binding.

RELA-bound TEs possess active enhancer features for target gene regulation

To explore the regulatory potential of the TE-derived RELA-bound regions, we assessed epigenomic proxies for active enhancers in TNFα-treated and untreated endothelial cells in each species. Starting with the 55 TE subfamilies in human, we investigated the dynamic changes in RELA binding, H3K27ac (active enhancers/promoters), H3K4me2 (active or latent enhancers/promoters), and chromatin accessibility (ATAC-seq) before and 45 minutes after TNFα treatment (Alizada et al. 2021). We first subcategorized these elements into "old-TEs" and "young-TEs" where "old" TE subfamilies were designated based on their colonization of mammalian genomes prior to the divergence of the selected mammals (90 MYA (Kumar et al. 2017)). As conserved orthologous RELA binding is closely linked to enhancer activity and NF-κB target genes (Alizada et al. 2021), we additionally included conserved RELA-bound regions (regardless of whether or not they were derived from TEs) as a "positive control" reference set containing the strongest NF-κB mediated enhancers. The remaining human-specific bound regions (regardless of whether or not they were derived from TEs) were included as a benchmark for baseline enhancer features at RELA bound regions.

We found that the old and young TE-derived RELA-bound regions were associated with high RELA binding signal after TNFα treatment and were comparable to that of conserved RELA bound regions (Figure 3A; Supplementary Figure 3). High chromatin accessibility was also observed for both TE-derived groups. While a clear signature of active enhancers, characterized by H3K4me2 and H3K27ac, was associated with TE-derived groups, this TE-derived epigenomic signal was far lower than what was observed for the conserved RELA bound regions and slightly lower than the human-specific group. To further assess the enhancer features of TE-derived subsets in human, we analyzed data from a genome-wide global RNA polymerase run-on assay, which we previously generated from a karyotypically normal, immortalized human aortic endothelial cell line (TeloHAEC) (Fish et al. 2017). This profiling was achieved using ChRO-seq (Chromatin Run-On sequencing (Chu et al. 2018)) which reveals bi-directional transcription of enhancer RNAs (eRNAs), a hallmark of active enhancers (Kaikkonen et al. 2013). We found a clear pattern of bi-directional transcription of eRNAs in all four RELA groups (Figure 3B). Conserved RELA-bound regions showed the highest eRNA signal followed by old TE-derived RELA-bound regions. Importantly, evidence for TE-derived RELA bound regions being associated with active enhancer characteristics was also observed for mouse and cow TE-derived RELA bound regions (Supplementary Figure 4). Overall this analysis supports the possibility that transposons derive active enhancers which are involved in NF-κB mediated gene regulation.

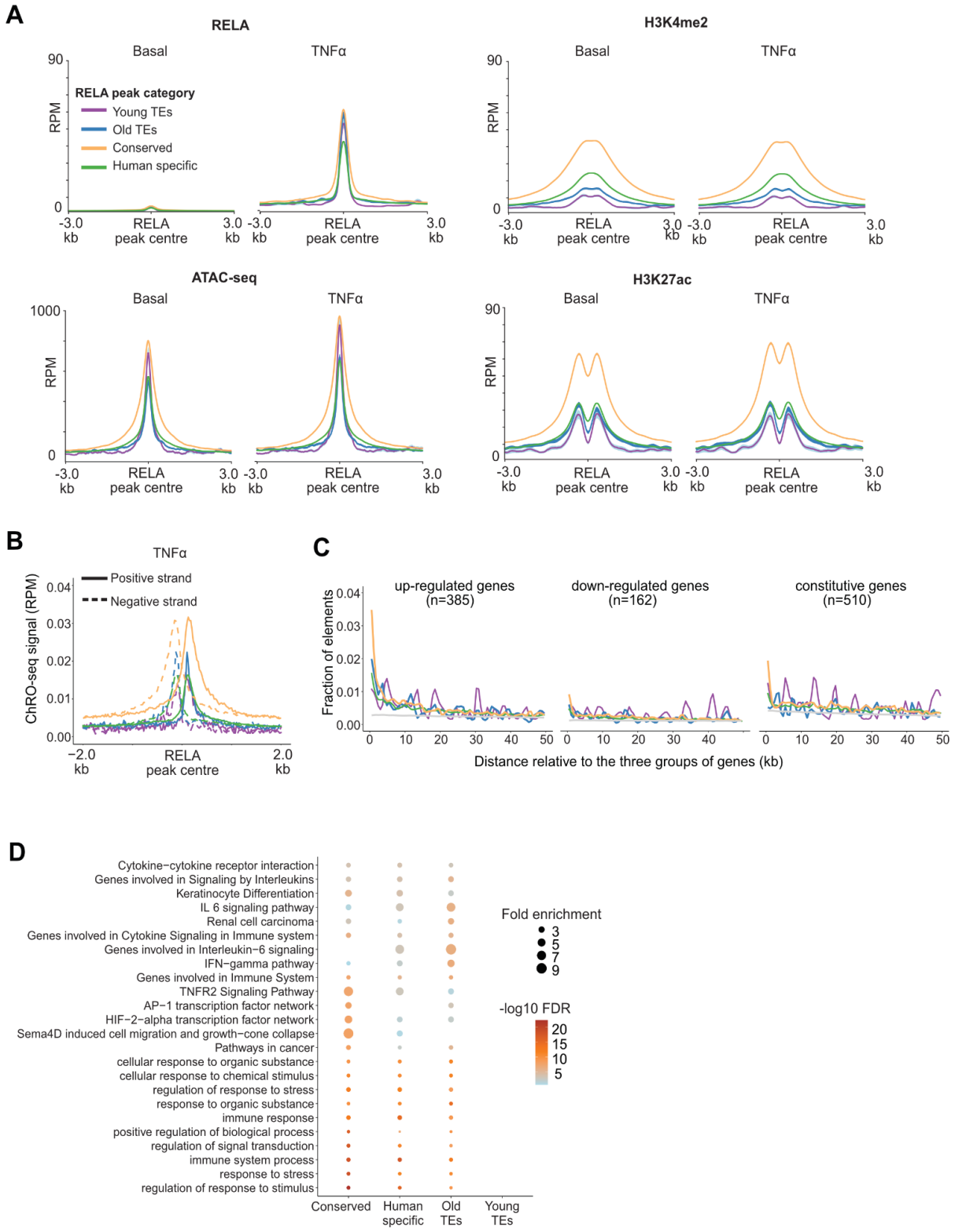

Figure 3. RELA-bound TEs possess active enhancer features for target gene regulation. Four categories of RELA-bound regions were interrogated: young TE-derived RELA-bound regions, old TE-derived RELA-bound regions, conserved RELA-bound regions and human specific RELA-bound regions. A) Epigenetic profile around

RELA-bound TEs before and after TNFα stimulation. Included are aggregated signals of RELA ChIP-seq, H3K4me2 and H3K27ac ChIP-seq and ATAC-seq data from human aortic endothelial cells. B) eRNA signal from ChRO-seq of TeloHAEC around RELA-bound TEs with TNFα stimulation for 45 min. C) The distribution of RELA-bound TEs near up-regulated, down-regulated and constitutive genes. The three groups of genes were defined using matched RNA-seq data (see Methods). "Background" represents the distribution of all the genomic TEs near the three gene groups, shown in gray. D) Functional enrichment of RELA-bound TEs by GREAT (McLean et al. 2010).

To have a functional impact, we would expect TE-derived RELA bound regions to be found in the vicinity of NF-κB target genes. To address this question we reanalysed the RNA-seq data that was paired with the RELA ChIP-seq data obtained from primary human aortic endothelial cells before and after TNFα treatment (Alizada et al. 2021). We assessed the likelihood that RELA-bound TEs influenced NF-κB target gene expression by comparing the distribution of the four categories of RELA bound regions around up-regulated, down-regulated, and constitutive genes (Figure 3C; see Methods). We observed a significant enrichment of RELA-bound regions near the promoters of up-regulated genes compared to constitutive genes, with the old TE-derived RELA group showing the highest enrichments, followed by the conserved RELA group (binomial test for occurrences within 10kb relative to up-regulated vs constitutive gene TSS. Old TEs: 3.6-fold, 6.9×10^{-9}; Conserved: 1.8-fold, p-value$< 2.2\times10^{-16}$; Human specific: 1.6-fold, p-value$< 2.2\times10^{-16}$; and Young TEs: 0.9-fold, p-value=0.85). Accordingly, TE-derived RELA-bound regions were enriched for annotated immune/inflammatory functions (Figure 3D). While there was no enrichment of the young TE-derived RELA group (n=561), we found comparable enrichment signals in functionally related terms for the other three groups. We noted several terms particularly enriched in the old TE-derived RELA group, including "IFN-gamma pathway" (binomial FDR of 2.2×10^{-7}) and "IL6 signaling pathway" (binomial FDR of 1.06×10^{-6}). These observations suggest the possibility that ancient TE-derived RELA bound elements have preferentially contributed to IFN-gamma/IL6 signaling during human evolution.

To obtain more direct evidence that these distal RELA-bound regions are involved in long-range interactions with NF-κB target genes, we performed HiChIP (Fang et al. 2016; Mumbach et al. 2016) in TeloHAECs before and 45 min after TNFα stimulation using an antibody against H3K4me3, a post-translational histone modification found at active gene promoters. We found a union of 19,372 significant chromatin interactions before and 45 min after TNFα stimulation, 718 of which involved NF-κB induced genes (n=385). We found that conserved RELA-bound regions have the highest fraction of overlap with the portion of distal elements from the identified loop sets (14%), followed by old TE-derived (13%), human-specific (11%) and young TE-derived (9%) RELA bound regions. Overall, these TE-gene proximity results further support the possibility that TE-derived RELA-bound regions have been incorporated into NF-κB gene regulatory networks.

TE-derived RELA bound regions coincide with cell-type specific TF motifs

The NF-κB signaling pathway is active across diverse cell types and involves common and cell-type specific target gene expression (Tabruyn and Griffioen 2007; Gerondakis and Siebenlist 2010; Liu et al. 2017). While cell-type specific NF-κB binding is known to occur in collaboration with cell-type specific TFs (Kaikkonen et al. 2013; Hogan et al. 2017; Alizada et al. 2021), it was not known if specific cell types would be enriched for distinct TE subfamilies. To address this question, we reanalysed data from three additional cell types which all showed an appreciable number of RELA bound regions after TNFA stimulation: human umbilical vein endothelial cells (HUVECs; n = 12,442 (Brown et al. 2014)), lymphoblasts (LCLs; n = 52563 (Kasowski et al. 2010)) and adipocytes (n = 58,043 (Schmidt et al. 2015)) (Figure 4A). After calculating the overlap of RELA peak summits with RepeatMasker annotated TE subfamilies, we observed both cell-specific and cell-shared enrichments (Figure 4B and 4C). We found that HUVECs showed a subset of TE-derived RELA binding that we observed in our human aortic endothelial cell (HAEC) data. In contrast we found enrichments for multiple TE subfamilies unique to adipocytes, as well as a few highly enriched endogenous retroviruses specific to LCLs (e.g., LTR43, LTR43B, MER57E1, and MER73). Notably MER81 and AmnSINE1 subfamilies, which we previously identified as being enriched in human, mouse and cow aortic endothelial cells, were also found to be enriched across multiple cell types. MER81 showed the highest enrichment values in almost all four selected cell types (10.6-, 9.2-, 18.4-, and 8.6- fold enrichment for HAEC, HUVEC, Adipocytes, and LCL, respectively). AmnSINE1 shows both high conservation degree and cross-cell usage of RELA-bound instances when compared across cell type and species (with 60% and 53% respectively; Supplementary Figure 5).

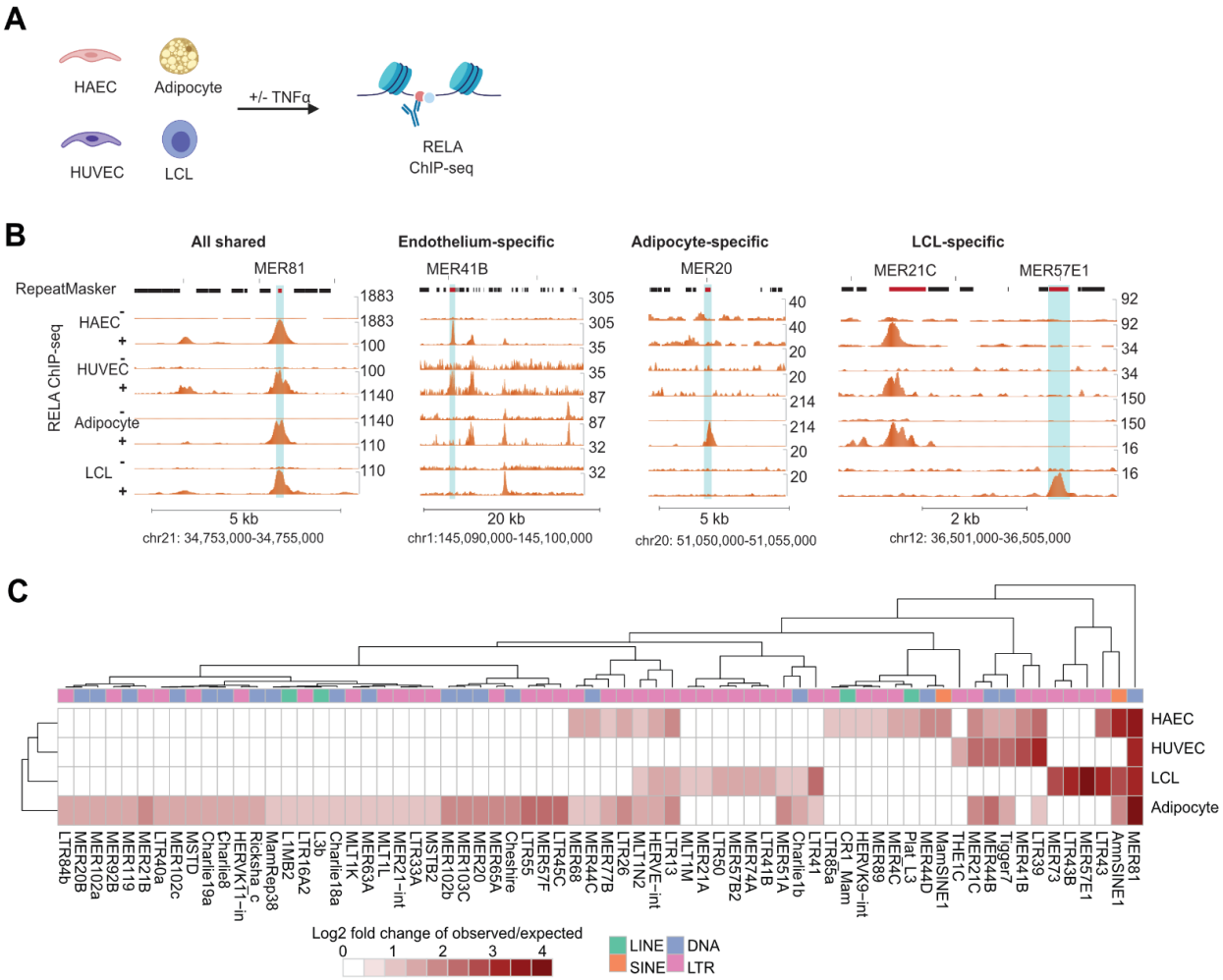

Figure 4. Multiple TE subfamilies were enriched within RELA-bound regions in different human cell types. A) A schematic showing the use of RELA ChIP-seq data from four human cell types. Together with RELA ChIP-seq data from this study, another three public RELA ChIP-seq datasets were incorporated to assess TE contribution to RELA-bound regions across cell types, which covers HAEC (Human Aortic Endothelial Cells), HUVEC (Human Umbilical Vein Endothelial Cells), LCL (Lymphoblastoid Cells) and Adipocytes. B) Examples of cell-type-shared and cell-type-specific TE-derived RELA-bound regions. C) An enrichment heatmap of identified TE subfamilies in the four cell types (cutoffs: FDR <0.05, log2 fold change > 1 and observed overlap events >= 15).

To gain further insight into the nature of the shared and cell-specific TE-derived RELA binding we investigated TF motif usage within TE-derived RELA-bound regions from each species and cell type individually. Supporting a direct role of DNA sequence features in NF-κB recruitment, motifs from RELA and other NF-κB subunits (RELB, REL, NFKB1, and NFKB2) were consistently detected in the majority of TE subfamilies from all the species and cell types. For example, 16 out of 22 subfamilies in HAEC showed strong NF-κB motif enrichments (Figure 5A, 5B and Supplementary Figure 6). In endothelial cells, NF-κB collaborates with AP-1 for inducible transcriptional regulation (Oeckinghaus et al. 2011; Dorrington and Fraser 2019). Accordingly we found enrichments for motifs from AP-1 subunits (e.g., JUN and FOS) in specific TE subfamilies enriched in RELA binding data from HAEC, HUVEC and cow aortic

endothelial cells (Figure 5A and Supplementary Figure 6B). We also found multiple transposable elements with high AP-1 motif enrichment that lacked NF-κB motif enrichment (such as Tigger7 and MER44B/C/D in HAEC). This suggests that a subset of TE-derived RELA bound elements arise from co-binding of NF-κB and AP-1 to non-canonical/degenerate NF-κB sites (Kolovos et al. 2016). Additional evidence for the collaboration of ETS-related factors in TE-derived RELA bound regions was observed in aortic endothelial cells from human and cow (7/22 and 11/18 enriched TE subfamilies in human and cow, respectively). This finding could be driven by ERG binding, a known ETS factor that plays a critical role in endothelial function and inflammation (Shah et al. 2016; Hogan et al. 2017). Further supporting the importance of collaborative TF binding in TE-derived RELA bound regions, we found enrichments of 9 subfamilies (e.g., MER20 and L1MB2) for the adipocyte-specific master regulator CEBPA in human adipocyte RELA ChIP-seq data (Freytag et al. 1994; Darlington et al. 1998) (Figure 5B).

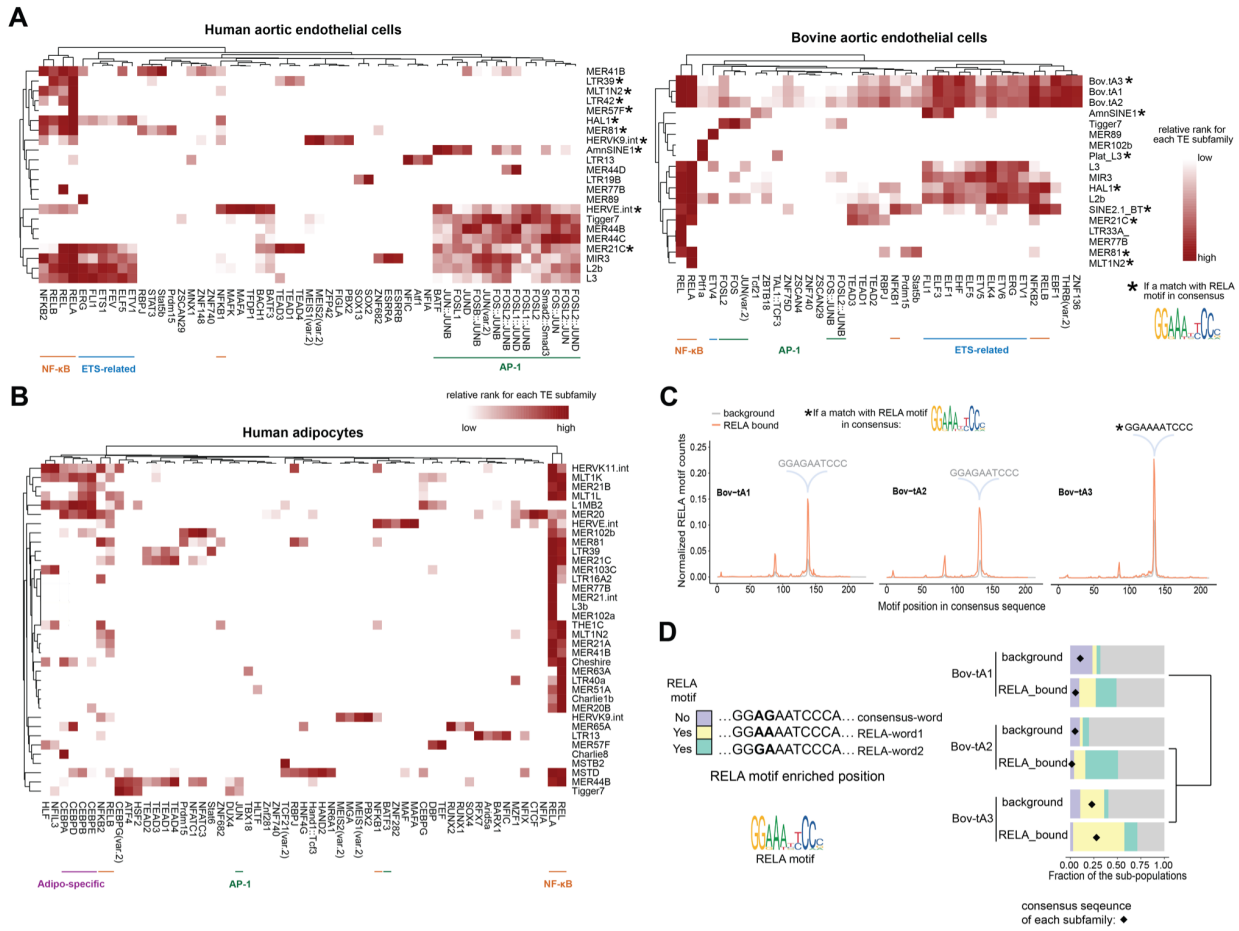

Figure 5. Multiple TF motifs contribute to TE-derived RELA bound regions across species and cell types. Heatmaps of significantly enriched TF motifs from A) human aortic endothelial cells as well as cow aortic endothelial cells, and B) human adipocytes. Significantly enriched TE subfamilies from species/ cell type analyses were investigated and only TEs detected with significant TF motifs are shown in the heatmaps. Rows and columns of the heatmaps indicate TE subfamilies and TF motifs, respectively. The intensity of heatmaps represent the rank of significantly enriched TF motifs for each TE (see more details in Methods). TF members of NF-κB, ETS, AP-1 and

adipocyte-specific factors have been highlighted. C) RELA motif distribution relative to the consensuses of Bov-tA1, Bov-tA2 and Bov-tA3. Gray and orange represent all instances and RELA-bound instances, respectively. D) Fine stratification of Bov-tA1/2/3 instances based on abundant RELA motif-words. The left panel shows three investigated motif-words, including a consensus word of Bov-tA1/2 (consensus-word), and the top two abundant RELA motif-words (RELA-word1 and RELA-word2). Note RELA-word1 is the consensus motif of Bov-tA3. The right panel displays the fraction of instances that match with each word for all instances and RELA-bound instances. Purple, yellow, green and gray represent consensus-word, RELA-word1, RELA-word2 and all other sequences, respectively. The dendrogram depicts the phylogenetic relationship of Bov-tA1/2/3, estimated by the similarity of their consensus sequences.

Bovine SINEs mediate RELA binding via a restricted repertoire of RELA motifs

To further investigate the role of TF binding motifs in TE-derived RELA bound regions we investigated the sequence features within the ~14,000 RELA-bound regions observed in the bovine SINE subfamilies Bov-tA1, Bov-tA2, and Bov-tA3. These three bovine SINE subfamilies make up ~15% of all RELA binding observed in bovine aortic endothelial cells. Bov-tA elements originated through recombination between tRNA-Glu derived SINEs and a fragment of BovB element before the radiation of Pecoran ruminants (~25-30 MYA) (Nijman et al. 2002; Nilsson et al. 2012). Collectively, Bov-tA1/2/3 give rise to more than 750,000 copies in the modern cow genome with a consensus sequence length of ~210 bp. Given the sequence divergence of transposon instances relative to the subfamily consensus, Bov-tA1 are estimated to be the oldest subfamily, followed by Bov-tA2 with Bov-tA3 being the youngest one (Figure 2C). Unlike the Bov-tA1 and Bov-tA2 consensus sequences, the Bov-tA3 consensus sequence contains a RELA motif. We found that a one nucleotide difference at the 130-140 bp position of the aligned consensus sequences can explain the presence of the RELA motif in Bov-tA3 (Supplementary Figure 7A).

To investigate the how the presence of a RELA motifs relates to RELA binding observed by ChIP-seq, we compared RELA motif density of RELA-bound Bov-tA1, Bov-tA2, and Bov-tA3 instances to unbound instances of the same repeat subfamily. Indeed we observed a significant increase in motif density at the 130-140 bp position in RELA bound regions (chi-squared test; 4.4-fold, 3.8-fold, 2.0-fold enrichment for Bov-tA1/2/3, respectively; P-value$<2.2\times10^{-16}$ for all three subfamilies; Figure 5C). Given that a single mutation at the 130-140 bp position of Bov-tA1 or Bov-tA2 can generate a RELA motif, which in turn could be spread through the genome by retrotransposition, we next asked if the RELA bound instances were biased towards specific motif words. We tabulated all 10 bp "words" in the 130-140 bp region that matched the RELA motif. The most abundant RELA motif matching word ("RELA-word1") corresponded to the consensus sequence of Bov-tA3 and accounted for 18%, 12%, and 54% of RELA-bound Bov-tA1, Bov-tA2, and Bov-tA3 elements, respectively (Supplementary Figure 7B). The second most frequent RELA motif containing word ("RELA-word2") was also found in all three subfamilies and accounted for 22%, 35%, and 14% of RELA-bound Bov-tA1, Bov-tA2, and Bov-tA3 elements, respectively. The presence of "RELA-word1" and "RELA-word2" in RELA-bound instances was significantly above background expectations for all three

subfamilies (Supplementary Figure 7B; chi-squared test; p value $<2.2\times10^{-16}$ for all three subfamilies; Figure 5D). Overall this analysis shows that while the consensus sequences for Bov-tA1 and Bov-tA2 do not possess a canonical RELA motif, a single base pair mutation can create one for RELA recruitment in Bov-tA1/2 instances. The prevalence of RELA motif-containing instances of these SINE subfamilies suggests NF-κB played a role in their genome-wide expansion and evolution.

While the bovine SINE expansion is the most prominent example of specific mutations in the ancestral TE sequence leading to de novo RELA binding sites, we also observed an enrichment of RELA motifs at a specific position (378-387 bp) of the consensus sequence and an increased fraction of RELA motif-word-matched subpopulations in RELA-bound instances in MER41B in human aortic endothelial cells (Supplementary Figure 8A and 8B). The other potential example of this phenomenon is LTR13, which lacks a RELA motif in the consensus but enriches for RELA binding. However, we did not find evidence for a site-specific RELA motif within this subfamily.

DNA transposon MER81 carries two RELA motifs responsible for enhancer activity

MER81 was the only TE subfamily enriched in RELA bound regions from all tested species and cell types. Using RepeatMasker annotations from a representative set of 24 mammals (Chuong *et al*. 2016), we found MER81 to be present in all 23 placental mammals and absent in the non-placental mammal representative opossum (Supplementary table 1). On average, 2800 copies (median value of all 24 species) were identified in each placental mammal. There is a notable lineage-specific reduction of annotated MER81 elements in Eulipotyphla (hedgehog) and Rodentia (mouse, rat and prairie vole) species (~500 copies in each species). In human, 3639 MER81 elements were annotated in the genome, among which 734 had RELA motifs. From our HAEC ChIP-seq data, we identified 87 extant MER81 copies bound by RELA in response to TNFα stimulation. Consistent with the general features of RELA-bound TEs, RELA-bound MER81s also exhibit active enhancer characteristics measured by H3K4me2, H3K27ac and ATAC-seq signal before and after TNFα stimulation (Figure 6A, Supplementary Figure 9A). Furthermore, RELA-bound MER81s were enriched for immune-related functions (Supplementary Figure 9B), such as "increased hematopoietic cell number" (binomial p value= 4.4×10^{-6}) and "IFN-gamma pathway" (binomial p value= 5.5×10^{-5}). In particular, we observed the association with multiple immune/inflammatory genes. For example, the proximal promoter region (0.3kb upstream of the TSS) of the gene *CXCR5*, which plays an important role in B cell maturation and migration (Pereira et al. 2010), has one instance of MER81 element with RELA binding that is shared across all four human cell types. We also found several interferon-signaling related genes in the vicinity of RELA-bound MER81 instances, including *IFNAR1* (surface receptor of IFN-α signaling), *TYK2* (downstream kinase required for IFN-α signaling), *IFNGR2* (surface receptor of IFN-γ signaling), *JAK2* (downstream kinase required for IFN-γ

signaling), *IRF1* (known transcription factor activated by IFN-γ signaling) (Supplementary table 2).

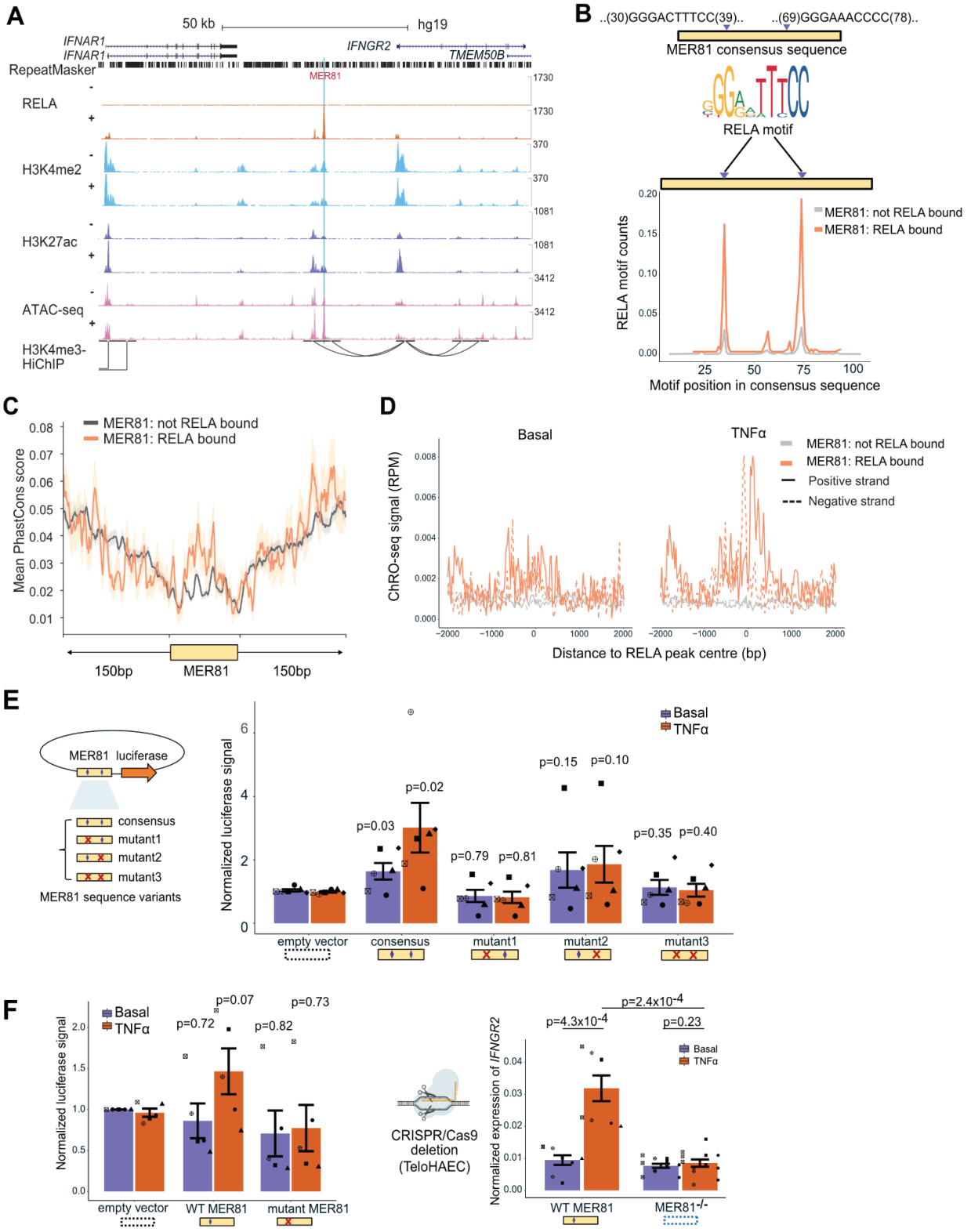

Figure 6. DNA transposon MER81 carries two RELA motifs responsible for enhancer activity. A) An instance of a MER81-derived RELA-bound region. The genome browser view covers tracks of RELA, H3K27ac ChIP-seq, ATAC-seq and H3K4me3 HiChIP. "-" and "+" denote basal and TNFα stimulation, respectively. B) Two canonical RELA motifs in the MER81 consensus sequence appear to function for RELA binding. The left panel shows the two RELA motifs in the consensus sequence, while the right panel depicts normalized motif counts of RELA motifs in extant human MER81 instances. Orange and gray represent RELA-bound and no RELA-bound instances, respectively. C) Conservation scores around RELA-bound vs no RELA-bound MER81 instances. PhastCons score was used to evaluate the conservation degree. D) eRNA signal around MER81-derived RELA sites before and after TNFα stimulation. eRNA signal was obtained from TeloHAEC ChRO-seq data. E) Functional validation of the enhancer activity of MER81 consensus sequence. Luciferase reporter assays were performed to determine the enhancer activity with/without TNFα stimulation for 3 hours in HEK293T cells. The figure shows luciferase signals of MER81 consensus sequence and three sequence variants. F) Functional validation of the enhancer activity of a native MER81 sequence. The left panel displays luciferase signals of the native MER81 sequence shown in A) and its variant sequence. Luciferase reporter assays were performed with/without TNFα stimulation for 3 hours in HEK293T cells. This selected MER81 instance was then deleted using CRISPR/Cas9 in TeloHAECs. The right panel shows the qPCR-measured expression level of *IFNGR2* in wild-type and the MER81 knockout TeloHAECs (with/without TNFα simulation for 45 min). For E) and the left panel of F), we have performed unpaired one-sided t-test for each group compared to the empty vector under the corresponding condition. Significance values of the statistical tests are shown on top of each bar. For the right panel of F), we have performed paired one-sided t-test for comparisons that involve the same genotype (e.g., stimulated vs unstimulated mutants) and unpaired one-sided t-test for ones that involve different genotypes (e.g., TNFα stimulated wild-type vs mutants). Error bars indicate standard error of the mean. Each biological replicate is shown with a unique shape in the figure.

To understand how MER81 could recruit RELA for binding, we examined its sequence features. We found two canonical RELA motifs in its 114-bp consensus sequence, which could explain RELA binding (Figure 6B). To expand the repertoire of RELA-bound MER81 elements, we additionally searched the pool of annotated RELA-bound regions from the GTRD database, which contains data from several human cell lines (Supplementary material 1). Together with our HAEC data, we obtained 327 MER81 instances that overlapped a RELA peak summit in at least one human cell type. We observed a clear enrichment of normalized RELA motif counts in RELA-bound MER81 elements compared to unbound ones (0.16 vs 0.03 for MER81 RELA-motif 1 and 0.19 vs 0.03 for MER81 RELA-motif 2) (Figure 6B). In other words, the loss of RELA binding in extant MER81 copies could be largely explained by the loss of RELA motifs. In addition, RELA-bound MER81 sequences show increased sequence constraints relative to unbound ones, further indicating their potential functional importance (Figure 6C).

Further support that these RELA-bound MER81 sequences can serve as enhancers, was obtained by ChRO-seq analysis revealing the bi-directional signal associated with active enhancers which are modulated by TNFα stimulation. This feature was not seen at MER81 elements that were not bound by RELA (Figure 6D). To further explore the functionality of these MER81 elements, we performed luciferase reporter assays to assess the enhancer activity of the sequences. We first tested the activity of the MER81 consensus sequence, which serves as an "ancestral" representative of all MER81 instances. We found a moderate enhancer activity of the MER81 consensus sequence compared to an empty vector control under basal condition (one-sided t-test; 1.6-fold, p-value=0.03), which was much enhanced upon TNFα stimulation (one-sided t-test; 3.0-fold, p-value=0.02). Interestingly, the enhancer activity of the sequence appeared

to be lost when the RELA motifs within the sequence were disrupted (one-sided t-test; p-values of 0.81, 0.10 and 0.40 for TNFα-stimulated mutant1, mutant2 and mutant3, respectively; Figure 6E). Altogether, our results implicate that the two RELA motifs confer the enhancer activity of the MER81 consensus sequence.

Next, we selected one native RELA-bound MER81 element in human that resides in the vicinity of interferon signaling genes to explore its enhancer function (Figure 6A). This instance is located between two interferon receptor genes, *IFNAR1* and *IFNGR2* (+58 kb and -19 kb relative to *IFNAR1* and *IFNGR2* TSS, respectively), both of which we identified as RELA target genes in our HAEC RNA-seq data. Our epigenomic data indicate that this MER81 sequence coincides with the largest RELA peak at the locus and is within an accessible chromatin region with a strong H3K4me2/H3K27ac signal. Using our H3K4me3 HiChIP data from TeloHAECs we identified a chromatin loop connecting this instance and the downstream target gene *IFNGR2*. We first validated the enhancer activity of this sequence via luciferase reporter assay. This instance only retained the first RELA motif and appeared to be moderately inducible in response to TNFα stimulation despite a borderline p-value (one-sided t-test for TNFα-stimulated native sequence, 1.5-fold, p-value=0.07). By contrast, the RELA motif-shuffled counterpart exhibited loss of enhancer activity, further supporting the importance of the RELA motif (one-sided t-test for TNFα-stimulated mutant sequence, 0.91-fold, p-value = 0.73; Figure 6F).

To explore the function of this enhancer in its chromatin context, we deleted the genomic region that spans from 213 bp upstream to 209 bp downstream of this instance in TeloHAECs using CRISPR/Cas9 (Supplementary Figure 10). We then stimulated wild-type cells and deletion mutants with and without TNFα for 45 minutes, and quantified the pre-mRNA transcript expression of *INFGR2* normalized to *TUBA1A* using qPCR. We confirmed the TNFα-dependent activation of *INFGR2* in wild-type cells with a 3.36-fold change of pre-mRNA transcript expression (paired one-sided t-test, p-value of 4.3×10^{-4}). By contrast, the deletion of this instance almost fully abolished the inducibility of *INFGR2* (1.1-fold change, paired one-sided t-test, p-value of 0.23; Figure 6F). The result indicates that this native MER81 instance functions as a bona fide enhancer, which is required for the regulation of the downstream target gene *INFGR2*. Altogether, our results support that MER81 sequences have functional RELA motifs responsible for the enhancer activity to regulate target gene expression.

Ancestral reconstruction reveals a stabilization of RELA motifs during MER81 evolution

The enrichment of RELA binding in extant MER81 copies appears to be a common feature across cells and mammalian lineages. However, the extant copies of MER81 with RELA binding ability tend to be species-specific. For example, MER81 primarily derives species-specific RELA binding events (79%) despite being an ancient mammalian transposon. Furthermore, there is a mouse-specific loss of MER81 copies, with only 500 elements remaining in the mouse genome, compared with ~3600 in human and ~2600 in cow (Supplementary Figure 5; Supplementary material 2).

To further delineate the evolutionary trajectories of the MER81 elements over the course of human evolution, we utilized ancestral whole-genome reconstructions of human and ten other mammalian species which cover an evolutionary time window of ~100 million years (Campitelli et al. 2022). With this strategy, we were able to recover the genome sequences of each node/common ancestor (Figure 7A). By pooling all ten reconstructed genomes together with the modern human genome, we achieved 11 nodes of genome sequences and were able to study "human evolution" in a "time-course" manner. We found the total copy number of MER81 was highly stable during evolution; in particular, ~2800 MER81 elements were detected in each node. The stability of MER81 copies suggests no or low level of transposition activity in the recent ~100 million years. Secondly, a large proportion (20-45%) of MER81 elements were identified with RELA motifs and the proportion containing a RELA motif decreased steadily over time, with >45% of MER81 elements containing a RELA motif in node 10 (or the human-star-nosed-mole common ancestor) to only ~20% containing a RELA motif in modern human (Figure 7B).

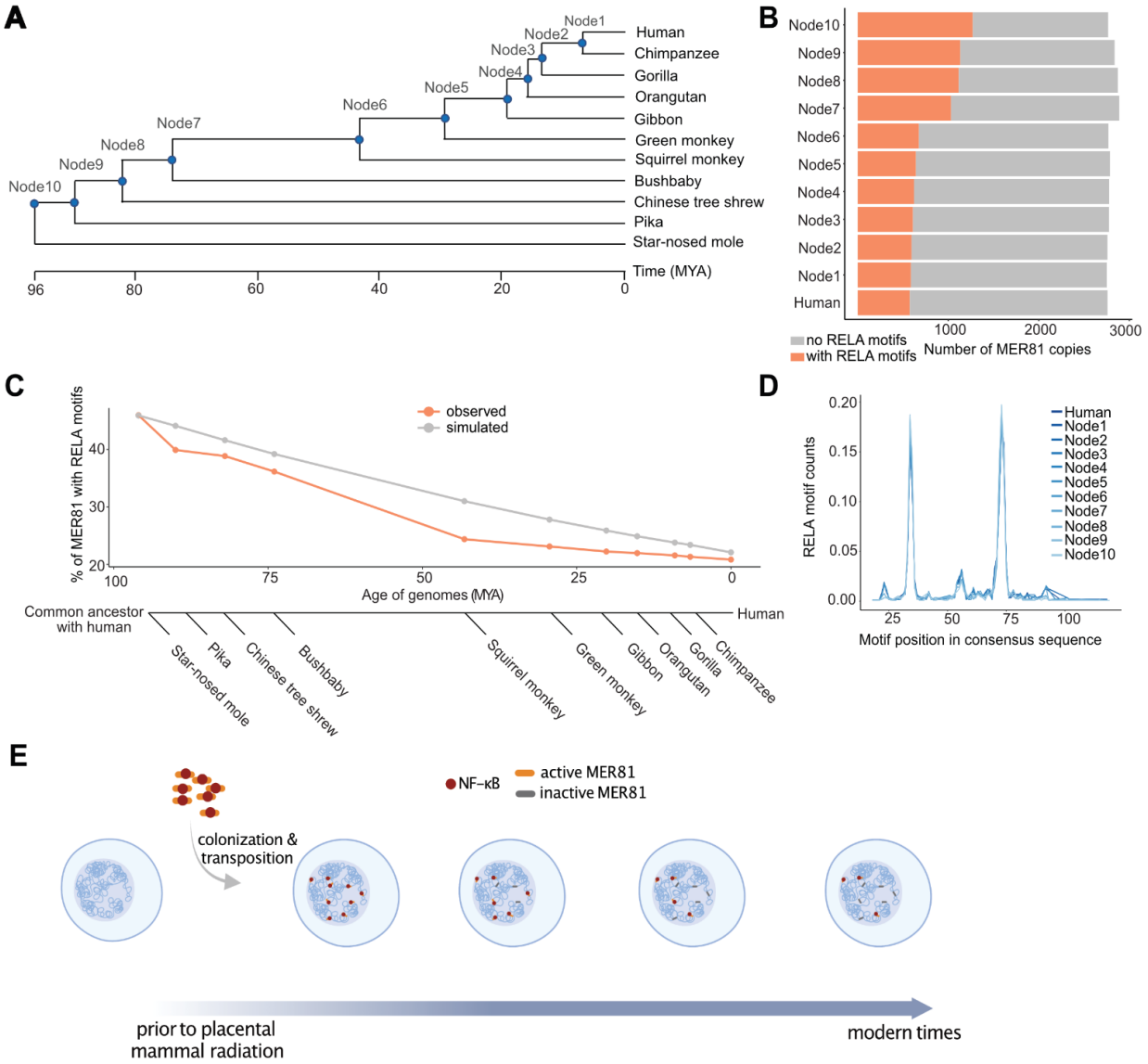

Figure 7. Ancestral reconstruction reveals a stabilization of RELA motifs during MER81 evolution. Modern human together with ten related mammalian species were selected to reconstruct a series of ancestral genomes based on multiple sequence alignments. A) Phylogenetic tree of the 11 selected species to reconstruct ancestral genomes. Each node in the tree represents a reconstructed genome of the corresponding common ancestor. All the nodes were collected to form a series of representative time points in the process of human evolution. The phylogenetic tree was obtained by querying TIMETREE database (Kumar et al. 2017). B) The number of MER81 copies in each human ancestor genome. Gray and orange represent total numbers and copies containing RELA motifs, respectively. C) RELA motif change in MER81 along human evolution. The percentage of MER81 proportion containing RELA motifs is shown as a function of the estimated evolutionary age of each node/common ancestor (divergence time estimation obtained from TIMETREE database (Kumar et al. 2017)). A simulation of MER81 sequence change based on nucleotide substitution was shown in gray (see Methods). D) The distribution of RELA motifs in MER81 instances in all human ancestors. Only instances with detected RELA motifs were investigated. E) Schematics of a MER81 co-option model during human evolution. MER81 colonized into the genome at a very early stage. These elements were then subject to a long-term selection with no or low level of expansion by active transposition. Functional units were potentially co-opted as enhancers to regulate target gene expression.

Next, we simulated RELA motif change in these MER81 elements over the course of the recent ~100 million years, by taking into account estimated nucleotide substitution rate and evolutionary time (see Methods). The simulation indicates a linear decrease of the RELA motif over time. In contrast to the simulation, we observed a rapid decrease of RELA motifs in the early stage of evolution, followed by a stabilizing trend in recent years (Figure 7C). The deceleration of the RELA motif decay rate in MER81 copies is likely to be a signature of negative selection for a subset of MER81 elements. We further investigated the distribution of RELA motifs within MER81 elements that contain the motif, and confirmed that these motifs still centered on the two positions from the consensus sequence and that no apparent motif turnover or shift occurred within these elements (Figure 7D). Together, our analysis supports a model where the DNA transposon MER81 colonized the genome prior to the radiation of placental mammals, followed by a long-term selection for the NF-κB motifs it contained (Figure 7E). At least a subset of these elements may have been co-opted as active enhancers to regulate target gene expression in an NF-κB-dependent manner.

DISCUSSION

In this study, we explored the contribution of transposable elements to genome-wide binding profiles of NF-κB in three mammalian lineages. Our comparative analyses of RELA binding, together with epigenetic and transcriptional profiles, revealed conserved and lineage-specific TE subfamilies that are likely to participate in mammalian NF-κB regulatory networks. The expansion of RELA-bound regions in the bovine genome largely results from Bovine SINE subfamilies Bov-tA1/2/3, which contribute over 14,000 RELA-bound regions to the cow genome (Figure 2A, 2B). This bovine-specific expansion even exceeds the ~10,000 SINE-B2 derived CTCF sites observed in the mouse genome (Bourque et al. 2008; Schmidt et al. 2012), making it one of the most striking examples of lineage-specific TF binding expansions observed in mammals to date.

The remarkable lineage-specific expansion of Bov-tA1/2/3 may imply a more general strategy to shape regulatory networks in ruminants. Kelly *et al*. 2022 recently discovered a substantial expansion of IFNγ-inducible elements by a ruminant-specific SINE, Bov-A2 (~12,000 instances), for the regulatory evolution of ruminant innate immunity. Interestingly, Bov-A2 originated from Bov-tA elements via dimerization of the Bov-tA tail regions after the radiation of Pecoran ruminants (>10 million years ago) (Nijman et al. 2002; Nilsson et al. 2012; Bao et al. 2015). Bov-A2 and Bov-tA1/2/3 constitute the family BovA, and the related evolutionary trajectory highlights a potentially pervasive function of the TE family BovA in ruminant-specific innate immunity. Together, our findings have strengthened the possibility that ruminants have taken advantage of SINEs (in particular, BovA), to largely expand the repertoire of immunity-related cis-regulatory elements, which may have a large functional impact on ruminant immune regulatory networks.

For both Bov-tA1 and Bov-tA2, we observed evidence of "maturation" of transposon instances from imperfect "pre-sites" or "proto-motifs" in ancestral sequences to perfectly matched RELA binding motifs. This phenomenon, currently described as the "proto-motif" model (MacArthur and Brookfield 2004; Bourque et al. 2008; Zemojtel et al. 2011; Sundaram and Wysocka 2020; Judd et al. 2021), involves individual TEs acquiring their TF-binding motifs after insertion. Observing this phenomenon in Bov-tA1 and Bov-tA2 suggests a model where the ancestral Bov-tA1 lineage repeatedly evolved the two prevalent and spatially constrained RELA-motif words. Through their ability to bind NF-κB, one or more of these elements increased their transcription and hence the opportunity for retrotransposition. This gave rise to the Bov-tA1 subfamily heterogenous for their affinity for NF-κB. Subsequently, the ancestral Bov-tA2 element, which also lacks an ancestral RELA binding motif, evolved the same two common RELA motif-words, which in turn facilitated their expansion in the cow genome. Finally, the ancestral sequence of the Bov-tA3 element, which contains the most prevalent RELA-motif word, contributed disproportionately to the observed Bov-tA derived RELA binding events (Figure 2B). Notably and consistent with this model, Bov-tA3 also shows evidence of expansions of RELA-word2 suggesting the gain and loss of RELA motifs was an ongoing process for Bov-tA elements during the evolution of ruminants.

Our comparative epigenomic study of TE-derived RELA binding allowed us to identify 24 TE subfamilies enriched within RELA-bound regions which appear to predate the human-mouse and mouse-cow divergence times (Figure 2C). Of these TE subfamilies, AmnSINE1, UCON26, Plat_L3, MamSINE1 were found to be enriched for RELA binding in two species. The ~300 million-year-old AmnSINE1 subfamily has been shown to contribute TF binding sites to mammalian gene regulatory networks (Sasaki et al. 2008; Hirakawa et al. 2009; Chuong et al. 2016) and our results suggest AmnSINE1 derived NF-κB binding may contribute its enhancer function across species and cell types. We also found two ancient TEs that were enriched for RELA binding in all three species and across distinct cell types: the LTR retrotransposon MER21C and DNA transposon MER81. While there is one documented example of a co-opted MER21C element forming part of a human gene promoter (Romanish et al. 2007), its influence on NF-κB mediated gene regulation remains to be seen.

One of the central discoveries in this study is the pervasive contribution of the ancient DNA transposon, MER81, to NF-κB mediated regulatory networks. The majority of MER81-derived RELA-bound regions are species-specific (79%) indicating the ancient MER81 elements may play core NF-κB functions in a human-specific manner. Growing evidence has indicated that TEs, especially evolutionarily young TEs, act as a potent source of cis-regulatory elements, which primarily innovate regulatory programs in a species- and tissue-specific manner (Buecker and Wysocka 2012; Chuong et al. 2016; Fuentes et al. 2018; Todd et al. 2019; Choudhary et al. 2022; Du et al. 2022; Kelly et al. 2022; Lee et al. 2022). On the other hand, ancient TEs, though few, are also able to shape regulatory landscapes through either species-specific (Notwell et al. 2015) or conserved mechanisms (Lynch et al. 2011). As a case of an ancient transposon, MER81 colonized the mammalian genome prior to the radiation of placental mammals. In human, the

majority (>80%) of MER81-derived RELA-bound regions are also shared across cell types, which increases the potential for MER81 to influence core components of NF-κB mediated transcriptional regulation. Noticeably, MER81 was also found to be inducible upon both bacterial and viral infection in human macrophages with strong NF-κB motif signals (Bogdan et al. 2020; Chen et al. 2022). We found MER81 displayed enhancer activity at the subfamily level and demonstrated the bona fide enhancer function conferred by one MER81 instance to control the activation of one IFN-γ signaling gene, *IFNGR2* (Figure 6A and 6F). It is possible that MER81 plays a larger functional role in IFN-γ signaling, given the distribution of instances near multiple IFN-γ signaling genes. However, the in vivo functions of the MER81 element and others remain to be seen.

MER81 also serves as an example to understand the evolutionary impact of a previously under-appreciated DNA transposon which is featured by much lower replicative efficiency via a "cut-and-paste" mechanism (Muñoz-López and García-Pérez 2010; Wells and Feschotte 2020). By tracing sequence changes of MER81 along the recent ~100 million years of evolutionary history, we propose a model whereby this ancient DNA transposon impacted regulatory evolution. In this model, limited waves of expansion occur over the course of evolution. Integrated instances, or rather, functional sequences they contain (such as TF motifs), then go through a long time of selection, where a functional subset has been repurposed to wire regulatory networks.

Given the clear link between TE activation and cellular stress responses (Bourque et al. 2018), their recurrent association with NF-κB may represent opportunistic means to increase their copy number in the mammalian genome. However, the NF-κB motif alone is unlikely to be the only sequence feature that facilitates RELA binding. This is because most functional mammalian RELA binding occurs in the context of accessible chromatin (e.g., (Alizada et al. 2021)), which is established through direct and indirect cooperation with other transcription factors (e.g., (Michida et al. 2020)). Similarly in this study, we found consistent evidence for collaborative TFs from the AP-1 and ETS family across species and cell types. For example, the absence of canonical NF-κB motifs in several AP-1 enriched TE subfamilies MER44B/C/D and Tigger could result from the co-binding of AP-1 and NF-κB at degenerate NF-κB motifs (Kolovos et al. 2016). Notably, both AP-1 and NF-κB motifs were also found to be significantly enriched within several TE subfamilies (such as MER41B and THE1C) from macrophages in response to bacterial infection (Bogdan et al. 2020), which is mainly modulated by NF-κB signaling. While additional experiments focusing on the collaborating TFs are needed, our results highlight the involvement of collaborating TFs in establishing TE-derived NF-κB enhancers.

Overall, our work highlights diverse and pervasive contributions of TEs to NF-κB binding across mammalian species and cell types. Many of these elements appear to have come into the genome with either pre-built NF-κB motifs or proto-motifs which facilitated their expansion. Given the active enhancer features and close proximity to multiple NF-κB target genes, it is likely that a subset of these numerous TE-derived RELA-bound elements have been incorporated into the host regulatory networks.

MATERIALS AND METHODS

ChIP-seq. To annotate mouse TE elements we performed H3K4me2 and H3K27ac ChIP-seq with and without TNFα treatment from mouse aortic endothelial cells. One biological replicate of primary mouse aortic endothelial cells (Cell Biologics, cat# C57-6052, lot# B092913T2MP) was selected and cultured following the methods of Alizada et al. 2021. Cells were then treated with 10 ng/mL recombinant mouse TNFα (Cell Applications, cat# RP2031-20) for 45 min in basal Endothelial Cell Growth Media MV2 (PromoCell) without supplements. The unstimulated control samples were treated with vehicles (i.e., an equivalent amount of water in MV2 media without supplements). For the ChIP-seq experiments, the primary aortic ECs were starved (basal media without supplements) for 16 h prior to TNFα stimulation. Next, cells were crosslinked in 1% formaldehyde (FA) solution (50-mM Hepes-KOH, 100-mM NaCl, 1-mM EDTA, 0.5-mM EGTA, 1% FA) for 10 min at room temperature and were then lysed for nuclei isolation as described in Schmidt *et al*., 2009 (Schmidt et al. 2009). Chromatin was sheared into 100–500-bp DNA fragments by sonication (Misonix Sonicator) at 27–28 W for 26 cycles (10-s ON, 1-min OFF, 28 W). We used ~1.7% of chromatin as input. ChIPs were performed at 4 °C overnight with 10 µg of antibody in a final volume of 250-µl block solution (0.5% BSA (w/v) in PBS): mouse anti-H3K27ac monoclonal (Millipore # 05-1334), rabbit anti-H3K4me2 polyclonal (Millipore # 07-030). The cross-links were reversed at 65 °C for 16 h. Proteins and RNA were enzymatically digested, and DNA was purified using phenol-chloroform extraction and ethanol precipitation. For library preparation, all of the ChIP DNA and 220 ng of input DNA were mixed with 3 units of T4 DNA polymerase (NEBNext® DNA Library Prep Master Mix Set for Illumina, #E6040L) to create blunt ends. To generate 5′-phosphates, 10 units of T4 PNK (NEBNext® DNA Library Prep Master Mix Set for Illumina, #E6040L) were added to the blunt-ended DNA in NEBNext End Repair Reaction Buffer (NEB, #E6042A) and incubated for 30 min at room temperature. DNA was purified using DNA Clean and Concentrator (Zymogen, cat# D4014). To add dAMP to the 3′ ends, the end-repaired DNA was mixed with Klenow Fragment (3′ → 5′ exo−) (NEB, #E6044A) in NEBNext dA-Tailing Reaction Buffer (NEB, #E6045A) and incubated for 40 min at 37 °C. The dA-tailed DNA was then purified. Illumina sequencing adapters containing uracil hairpin loop structure and 3′ T overhangs (NEB, #E7337A) were ligated to the DNA by adding 150 units of Quick T4 DNA Ligase (NEB, #E6047A) in Quick Ligation Reaction Buffer (NEB, #E6048A) and incubation for 15 min at room temperature. Looped adapter sequences were opened by removal of uracil from hairpin structures by adding 3 units of USER enzyme (Uracil-Specific Excision Reagent) (NEB, M5505S) and incubation at 37 °C for 15 min. This made DNA accessible for PCR amplification with barcoded primers for Illumina sequencing (NEB, cat# E7335L). PCR amplifications were carried out for 16 cycles [98 °C 30 s, (98 °C 10 s, 65 °C 30s, 72 °C 30 s) × 17 cycles, 72 °C 5 min, 4 °C hold]. The amplified and barcoded library was then selected for 200–350-bp fragments

using Pippin Prep (Sage Science) and sequenced on Illumina HiSeq2500 with a 100-bp single-end run to obtain ~20–25 million single-end reads per sample.

H3K4me3 based HiChIP. We prepared TeloHAECs with/without TNFα stimulation for HiChIP experiments (two replicates per condition). Briefly, TeloHAECs were cultured following the protocol from Alizada et al. 2021. For TNFα and vehicle treatment, the protocol was as described at the beginning of the ChIP-seq method, except here we used 10-ng/mL recombinant human TNFα (Cell Applications, cat# RP1111-50). HiChIP was performed from a modified PLAC-seq protocol (Fang et al. 2016) in three steps as follows: **(1)** In situ proximity ligation. 5 million TeloHAECs per replicate were crosslinked in 37% formaldehyde as per instructions of the Arima Hi-C kit (cat# A510008) and frozen down as 10^6 cells per aliquot. Briefly, after DNA estimation, cells corresponding to 3 ug of total DNA were thawed on ice and lysed in 20 ul of cold lysis buffer for 15 min, followed by permeabilization, digestion, and proximity ligation as per the instructions of the kit. We skipped step 10 of the Enzyme D step in the Arima kit and proceeded to ChIP. **(2)** Chromatin immunoprecipitation (ChIP). We adapted the protocol from Fang et al. 2016 with slight modifications. After proximity ligation, the nuclei were spun down at 2,500 g for 5 min at 4°C and the supernatant was discarded. The nuclei were then resuspended in 130 µl RIPA buffer (10 mM Tris, pH 8.0, 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.1% sodium deoxycholate) with proteinase inhibitors. The nuclei were lysed on ice for 10 min and then sonicated using Diagenode Bioruptor® Pico with the following setting: 10 Cycles of 30 Secs On/Off. After sonication, 370 ul of RIPA buffer was added to the samples before they were cleared by centrifugation at 20,000 G for 20 min and the supernatant was collected. The clear cell lysate was mixed with 500ul of antibody (2.5 µg of H3K4me3, Millipore cat #04-745) coated Dynabeads M-280 (Sheep anti-Rabbit Thermo Scientific cat #11201D) and rotated at 4°C overnight. Next day, the beads were collected by centrifugation at 2,000 rpm for 1 min and then resuspended in 1 ml of RIPA buffer and washed for three times in 1x RIPA buffer, twice in high-salt RIPA buffer (10 mM Tris, pH 8.0, 300 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.1% sodium deoxycholate), once with LiCl buffer (10 mM Tris, pH 8.0, 150 mM LiCl, 1 mM EDTA, 0.5% IGEPAL CA-630, 0.1% sodium deoxycholate), and twice with TE buffer (10 mM Tris, pH 8.0, 0.1 mM EDTA). Washed beads were first treated with 10 µg Rnase A in extraction buffer (10 mM Tris, pH 8.0, 350 mM NaCl, 0.1 mM EDTA, 1% SDS) by shaking at 1000 rpm for 1 h at 37°C. Then 20 µg proteinase K was added and reverse crosslinking was performed by shaking at 1000 rpm overnight at 65°C or at least 2 h. The fragmented DNA was purified with DNA Concentrator (Zymo Research, cat# D4014). **(3)** Biotin pull-down and library construction. The biotin pull-down procedure was performed according to Swift Biosciences Accel-NGS 2S Plus DNA Library Kit (cat# 21024). An optimal number of PCR cycles were determined by qPCR using the KAPA library quantification kit (cat# KK4824). Libraries were checked on a bioanalyzer for quality and then sequenced on a NovaSeq6000 S4 for PE150.

Luciferase reporter assay. For the investigated sequences, we designed 4 MER81 consensus-related sequences and 2 sequences from a native MER81 instance (located at chr21:34755496-34755596 of hg19). The consensus sequence of MER81 was obtained from

Repbase (Bao et al. 2015) and the native MER81 sequence was downloaded from the UCSC genome browser for coordinates of chr21:34755496-34755596 (hg19). The mutant sequences were designed by randomly shuffling the RELA motifs in consensus/native sequences. A motif scan test was performed using 746 known vertebrate TF motifs documented in JASPAR (Castro-Mondragon et al. 2022) to make sure no new TF motifs occur within the newly designed sequences. All these sequences of interest were directly synthesized and cloned in pGL3 luciferase construct backbone (service from General Biosystems). Since NF-κB signaling is pervasive across cell types, we chose a simple cell line, HEK293T, as the model to test the enhancer activity of the sequences. HEK293T cells (ATCC, cat# CRL-1573) were cultured in DMEM media (Gibco, cat# 11965092) supplemented with 10% FBS (Gibco, cat# 12483020). For luciferase assays, cells were seeded in 12-well dishes with a technical triplicate for each tested condition. HEK293Ts were transfected at 80% confluency with 1ug of pGL3 luciferase constructs and 0.1ug of pRenilla using Lipofectamine 2000 (2ul/1ug DNA, Invitrogen) per well in 1mL of optiMEM. After 5 hours, optiMEM was replaced with 2mL of DMEM + 10% FBS. For TNFα treatment, 10 ng/ml recombinant human TNFα (Cell Applications, cat# RP1111-50) was added for 3 hours. Luciferase reporter assay was performed using the Dual-Luciferase Reporter Assay System (Promega). 24h post-transfection, media was aspirated and 250 µl of passive lysis buffer (1X) was added to each well. Lysis was performed by shaking at 20 min covered in aluminum foil at RT. The GloMax20/20 Luminometer (Promega) was used to measure dual luciferase (Renilla and Firefly). Prior to taking measurements, lysed samples were diluted 10-fold to be within a detectable range of the luminometer. Afterward, 20 µl of diluted sample was added to each Eppendorf tube and samples were processed one by one. First 100 µl of the LARII reagent was added and the Firefly luciferase reading was obtained. Immediately after, 100 µl of the S&G reagent was added and the Renilla luciferase reading was obtained. To analyze the data, the Firefly/Renilla luciferase value was obtained. Renilla only background values were subtracted. All data points were then normalized to empty pGL3 control. We included a synthetic sequence with tandem 5x RELA motifs as a positive control and generally observed > 20-fold change with TNFα stimulation compared to empty pGL3 control for all experiments (data not shown in Figure 6E/6F). Overall, we prepared 6 biological replicates with 3 technical replicates each for each investigated sequence to ensure reproducibility.

CRISPR/Cas9 deletions. Cas9 targeting guide RNAs (gRNAs) were selected flanking the selected MER81 instance (chr21:34755496-34755596 of hg19) for deletion. On and off-target specificity of the gRNAs was calculated as described in Doench et.al 2016 (Doench et al. 2016) and Hsu et al. 2013 (Hsu et al. 2013) respectively, to choose optimal guides. Guide RNA plasmids were assembled with gRNA sequences using the protocol described by (Mali et al. 2013). Briefly, two partially complementary 61-bp oligos were annealed and extended using Phusion polymerase (New England Biolabs). The resulting 100-bp fragment was assembled into an AflII-linearized gRNA empty vector (Addgene, ID#41824) using the Infusion assembly protocol (TaKaRa Bio). The sequence of the resulting guide gRNA plasmid was confirmed by sequencing with either T7 or SP6 primers. TeloHAECs were transfected with 5 µg each of the 5′

gRNA, 3′ gRNA, and pCas9_GFP plasmids (Addgene, ID#44719; (Ding et al. 2013)) using the Neon Transfection System (Life Technologies). Forty-eight hours post-transfection, GFP-positive cells were isolated on a BD FACSAria. After recovery, single cells were seeded in 96-well plates and inspected for colonies after 1–2 weeks. Upon 60% confluency, the colonies were split into replicate 96-well plates. Genotyping of the deletions was performed by amplifying products internal to and surrounding the target deletion. All deleted clones identified from the initial screen were sequenced across the deletion to confirm deletion boundaries. Four homozygously deleted clones were selected for the following experiments (as shown in Supplementary Figure 10).

RNA isolation and qPCR. TeloHAECs were treated with/without TNFα for 45 min. Total RNA was purified from single wells of >85% confluent six-well plates using the RNeasy Plus Mini Kit (Qiagen), and an additional DNase I step was used to remove genomic DNA. RNA was reverse-transcribed with random primers using the high-capacity cDNA synthesis kit (Thermo Fisher Scientific). Primary *IFNGR2* gene expression was detected by primers targeting the fourth intron. The standard curve method was used to calculate expression levels using TeloHAEC genomic DNA to generate the standard curves. Levels of TUBA1A RNA were used as an internal control to normalize expression values.

Next-generation sequencing data processing. ChIP-seq, ATAC-seq, RNA-seq and ChRO-seq processing were the same as Alizada et al. 2021, except for RELA ChIP-seq data (including all cross-species and cross-cell-type RELA ChIP-seq data) where we chose a relatively lenient MAPQ cutoff of 3 to filter out multiple mapping reads. H3K4me3 HiChIP was processed using the MAPS pipeline (Juric et al. 2019) with default parameters. Significant loops were called using MAPS with default parameters for each sample. Since chromatin loops are expected to stay largely unchanged within a rapid induction system (in our case, TNFα treatment for 45 minutes), we merged the loops identified from basal and TNFα treatment conditions to generate a union set of active chromatin loops (in total 19372) for TeloHAECs. These loops vastly map active enhancer-promoter or active promoter-promoter connections (as the example in Figure 6A shows).

Enrichment analysis for TEs overlapping RELA-bound regions. We used the GAT tool (https://gat.readthedocs.io/en/latest/contents.html) to determine the potential overrepresentation of TEs within RELA-bound regions. To run the enrichment analysis, we input the peak summits of RELA peaks, RepeatMasker annotated TE subfamilies. We set the workspace (available genomic regions for simulation) as all non-random chromosomes. 10000 iterations were requested for empirical distribution estimation and other parameters were set as default. The input TE annotation files (matched annotation for hg19, mm10 and btau6) by RepeatMasker were downloaded using the table browser under the UCSC genome browser (https://genome.ucsc.edu). Simple repeats were excluded from all the TE annotations. In brief, the GAT tool shuffles RELA peak summits across the given workspace for 10000 iterations to obtain the distribution of expected overlap numbers for each TE subfamily. Empirical p-value and enrichment/fold change values are calculated by combining both observed and expected

overlap numbers. The cutoffs for significantly enriched subfamilies across species were: FDR <0.05, log2 fold change >1 and observed overlap >=10 events. For cross-cell-type analysis, slightly more stringent cutoffs were used to control potential false positives: FDR <0.05, log2 fold change >1 and observed overlap >=15 events.

Estimation of evolutionary ages of TEs. Age was calculated for each TE instance using percent divergence of sequence (obtained from the aforementioned TE annotation files from UCSC table browser) divided by estimated nucleotide substitution rates of the corresponding species from (Bulmer et al. 1991; Pace et al. 2008). The median value of all instances from the same subfamily was calculated as the estimation of the subfamily age. Two point estimations of human-mouse and mouse-cow divergence time (Figure 2C) were acquired from TIMETREE database (Kumar et al. 2017). Human-mouse divergence time was later used to divide TEs into "young" and "old" categories based on whether the predicted age of TEs was younger or older than when humans and mice diverged (in particular, the two categories were denoted as "Group Y" or "Group O", respectively, as shown in Figure 3).

Analysis of TE overlap with TF bound regions. We took advantage of all the processed public TF ChIP-seq data from GTRD (Kolmykov et al. 2021) (v2018.3) to investigate TE and TF connections globally. The bound regions of TFs documented in GTRD were generated by merging ChIP-seq peaks from diverse conditions and cell types. This feature represents a comprehensive repertoire of bound regions for each TF, which allows us to study the general relationship between TEs and TFs in an unbiased manner. In particular, we downloaded all human/mouse peak clusters called by MACS2. The peak summits were then extracted to investigate the overlap between TEs and TF bound regions (Supplementary Figure 1). For this analysis, we only kept TFs with peak clusters of more than 10,000, which resulted in 435 human and 266 mouse TFs. The fraction of TF peak summits that overlapped with TEs was then calculated for each TFs. Note that later in Figure 6, in order to capture all possible functional RELA-bound MER81 elements in human, we again used the RELA peak clusters documented in GTRD to expand RELA-bound MER81 instances in HAEC (n=87) to all the reported ones (n=327) for functional assessments (e.g., Figure 6B, 6C and 6D).

TF Motif analysis. In this study, we have performed two types of motif analysis: (1) motif enrichment analysis for a group of sequences; (2) motif scan analysis for individual TE instance or consensus sequence. The vertebrate TF motifs documented in JASPAR CORE (746 TFs in total; (Castro-Mondragon et al. 2022)) were selected as the database for both analyses. (1) For motif enrichment analysis, CentriMo from MEME suite (Bailey et al. 2015) (version 4.12.0) was used with the threshold "-ethresh" set to 1. The query sequences, for example, from Figure 5, were from RELA-bound TE instances. The overlapped RELA peak summits were first extracted for all these instances and extended with 250 bp in both directions. CentriMo was then performed for instances of each TE subfamily that was identified to be significant as shown in Figure 2. From the enrichment results, top 10 enriched TFs based on E values were saved for each TE subfamily and only TEs with significant TF motifs were considered for the following analysis. To make the enrichment results comparable across different TE subfamilies, we chose

to use the rank of TF motifs in each TE subfamily as the metric and visualized these values in heatmaps as shown in Figure 5A, 5B and Supplementary Figure 6. (2) In terms of motif scan analysis, FIMO from the MEME suite (version 4.12.0) was used with default parameters. For example, FIMO was used to determine RELA motifs in each TE consensus sequence (Figure 2B) and MER81 instances from all the reconstructed ancestral genomes (Figure 7D). In addition, one major downstream analysis from motif scan results was the visualization of motif distribution relative to the consensus sequence (e.g., Figure 5C, 6B and 7D). With regard to this analysis, RepeatMasker annotation was used to determine the corresponding position in the consensus sequence for each identified motif from instances. The occurrences of motifs in each position were then counted and normalized by the total number of instances.

Motif-word analysis for several example TE subfamilies. Motif-word analysis has been performed for Bov-tA1/2/3 and MER41B. According to the motif distribution plots (e.g., Figure 5C and Supplementary Figure 8A), the position of the most prominent "suboptimal" RELA motif (sequence that is only one or two nucleotides away from a matched RELA motif) in the consensus sequence was determined for each TE subfamily. Next, the RELA motif was scanned over all the instances and only matches with RELA motifs in instances that corresponded to the position of the "suboptimal" RELA motif in consensus were recorded. The recorded RELA motif-words were then counted for occurrences and the most abundant motif-words were picked for closer investigation based on the elbow plot (Supplementary Figure 7B). Finally, the occurrences of these motif-words in the queried position were counted for all the instances (denoted as "background") and RELA-bound instances (Figure 5D and Supplementary Figure 8B).

Visualization of sequence constraints and epigenetic signal profile. We used PhastCons scores (Siepel et al. 2005) (downloaded from hg38.phastCons100way.bw) to evaluate the sequence constraints around MER81 instances (Figure 6C). For epigenetic data, we used UCSC genome browser (Tyner et al. 2017) to visualize the signal from ChIP-seq, ATAC-seq and loops identified from H3K4me3 HiChIP on the whole genome scale. As to the aggregate plot of ChIP-seq and ATAC-seq signal around RELA-bound TEs (e.g., Figure 3A, 6D and Supplementary Figure 4), we used computeMatrix and plotProfile functions from deepTools2 suite (Ramírez et al. 2016) (v3.1.3). To visualize stranded eRNA signals from ChRO-seq, we used HOMER (Heinz et al. 2010) (v4.11). Details and parameter choices were as Alizada et al. 2021.

Target gene association and functional enrichment analysis for RELA-bound TEs. Human RNA-seq data have been used to determine the significantly regulated (up-regulated and down-regulated) and constitutive genes. The three gene sets used in this study were previously defined in Alizada *et al.* 2021 using both intronic and exonic reads. The distance of the four groups of RELA-bound regions relative to the three gene groups (Figure 3C) was determined using the TSS annotation of GENCODE (v19). The occurrences were counted in 10 kb bins and then converted to the corresponding fraction value for the distribution in Figure 3C. For the background reference, all the annotated TE instances were considered and the corresponding

fraction was calculated accordingly. As to the functional enrichment analysis, GREAT (McLean et al. 2010) (v3.0.0) was performed for the genomic coordinates of regions of interest. For Figure 3D, to make the enrichment results comparable for the four groups of regions, a number-matched strategy was used. In brief, we selected the top 1492 RELA peaks from the conserved and human-specific groups based on peak intensity to match the total number of group O (n=1492) as input for GREAT. Ontology terms from Biological Process (BP) and MsigDB pathway were extracted for each peak category. The top enriched ontology terms shown in Figure 3D were generated by combining the top 5 enriched terms (from BP/ MsigDB pathway; ranked by binomial FDR) of the four peak categories.

Cross-species and cross-cell-type comparison of RELA-bound regions. We define conserved RELA binding events in one species as the ones that can be detected in the orthologous regions of at least another one species with at least 1bp overlap. The workflow to identify the conservation degree (either conserved or species-specific) of any given RELA-bound regions is as Alizada et al. 2021. Based on this rule, human RELA peaks were divided into conserved and human specific groups, as shown in Figure 3. Similarly, when compared across human cell types, a RELA-bound region is shared if another site is identified from at least one other cell type with at least 1 bp overlap.

Ancestral genome reconstruction, MER81 annotation and RELA motif detection. Ancestral genomes were reconstructed for each common ancestor (nodes) shown in Figure 7A. RepeatMasker (Tempel 2012) with Dfam libraries (Hubley et al. 2016) was run for each reconstructed genome to annotate all human transposable elements. Details were extensively described in Campitelli et al. 2022 (Campitelli et al. 2022). To acquire high quality of MER81 annotations, we additionally required MER81 hits to be marked as non-redundant and contain a minimal length of 90 bp with RepeatMasker. For the detection of RELA motifs, we extracted sequences of all the valid MER81 instances from these reconstructed genomes and used FIMO to scan RELA motifs (as described in the TF motif analysis section). The motif occurrences within identified MER81 instances for each genome were counted and visualized in Figure 7B and 7C. Furthermore, the corresponding position of motif matches relative to the consensus sequence was also recorded for each genome (Figure 7D).

Simulation for neutral decay of MER81 elements during evolution. MER81 elements were extracted from the oldest reconstructed genome (common ancestor of human and star-nosed mole, or node 10). A primate specific estimate of nucleotide substitution rate (2.2*e-9/ (year*site); (Bulmer et al. 1991; Pace et al. 2008)) was used to simulate sequence change for the following 10 evolutionary time points. Starting from the collection of the oldest MER81 elements (real observations), all the following time points were estimated solely on nucleotide substitution. For example, each nucleotide of a MER81 instance was calculated for the likelihood of substitution based on the rate and evolutionary time, and the value of the nucleotide in the following time point/node was chosen based on the likelihood. For each node/ time point, all the simulated MER81 elements were scanned for RELA motifs using FIMO and the percentage of

elements with significant motifs was recorded. We have iterated 1,000 times to obtain the mean values for each node, which were used to plot the gray line in Figure 7C.

Data availability

Mouse H3K4me2 and H3K27ac ChIP-seq as well as TeloHAEC H3K4me3 HiChIP data generated in this study have been deposited into the ArrayExpress database under accession numbers E-MTAB-12212 and E-MTAB-12213, respectively. To access these two datasets, use username and password sets of "Reviewer_E-MTAB-12212" + "jfpzzzrs" and "Reviewer_E-MTAB-12213" + "FH2ftWII", respectively. The publicly available human/mouse/cow ChIP-seq and ATAC-seq data were downloaded from ArrayExpress with accession numbers E-MTAB-7889 and E-MTAB-7878, respectively. HAEC RNA-seq and TeloHAEC ChRO-seq were downloaded from ArrayExpress with accession numbers E-MTAB-7896 and E-MTAB-9425. The RELA ChIP-seq data for HUVEC (Brown et al. 2014), LCL (Kasowski et al. 2010) and Adipocytes (Schmidt et al. 2015) were downloaded from GEO database under accession numbers: GSE54000, GSE19486 and GSE64233.

CODE AVAILABILITY

Source code for main figures and supplementary figures is available on Wilson lab GitHub repository (https://github.com/wilsonlabgroup/RELA_TE).

ACKNOWLEDGMENTS

We would like to thank Alan Moses, Dustin Sokolowski, Cadia Chan, Mohamed Hawash and Huayun Hou for helpful discussions and the Centre for Applied Genomics (TCAG) for assistance with DNA sequencing. Funding for this project was provided by the Canadian Institutes of Health Research (CIHR) (201603PJT-364832) to M.D.W. and J.E.F. and the National Institutes of Health (S.R., M.D.W and J.A.M; R01-HG010045-01). J.E.F. and M.D.W. were supported by an Early Researcher Awards from the Ontario Ministry of Research and Innovation and Tier 2 Canada Research Chairs from CIHR. N.K. received a Canada Graduate Scholarship from the Natural Sciences and Engineering Research Council of Canada (NSERC) and an Ontario Graduate Scholarship. L.A. was supported by a NSERC CGS-M. K.R. was partially supported by a postdoctoral fellowship from the Ted Rogers Centre for Heart Research. L.C. was supported by a NSERC PGS-D scholarship.

CONTRIBUTIONS

L.W. performed data analysis, wrote the manuscript. A.A. R.K. L.A. designed and performed experiments; N.K. performed luciferase assays; T.T. performed CRISPR/Cas9 experiments; L.C., Z.P. performed ancestral genome analyses; T.H., S.R. J.M. and J.F. provided funding and supervision. M.D.W. designed experiments, wrote the manuscript, acquired funding, and supervised the project. All authors edited and approved the manuscript.

REFERENCES

Alizada A, Khyzha N, Wang L, Antounians L, Chen X, Khor M, Liang M, Rathnakumar K, Weirauch MT, Medina-Rivera A, et al. 2021. Conserved regulatory logic at accessible and inaccessible chromatin during the acute inflammatory response in mammals. *Nat Commun* **12**: 567.

Bailey TL, Johnson J, Grant CE, Noble WS. 2015. The MEME Suite. *Nucleic Acids Res* **43**: W39-49.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**: 11.

Bhatt D, Ghosh S. 2014. Regulation of the NF-κB-Mediated Transcription of Inflammatory Genes. *Front Immunol* **5**: 71.

Bogdan L, Barreiro L, Bourque G. 2020. Transposable elements have contributed human regulatory regions that are activated upon bacterial infection. *Philos Trans R Soc Lond B Biol Sci* **375**: 20190332.

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, et al. 2018. Ten things you should know about transposable elements. *Genome Biol* **19**: 199.

Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew J-L, Ruan Y, Wei C-L, Ng HH, et al. 2008. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res* **18**: 1752–1762.

Brown JD, Lin CY, Duan Q, Griffin G, Federation A, Paranal RM, Bair S, Newton G, Lichtman A, Kung A, et al. 2014. NF-κB directs dynamic super enhancer formation in inflammation and atherogenesis. *Mol Cell* **56**: 219–231.

Buecker C, Wysocka J. 2012. Enhancers as information integration hubs in development: lessons from genomics. *Trends Genet* **28**: 276–284.

Bulmer M, Wolfe KH, Sharp PM. 1991. Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian

orders. *Proc Natl Acad Sci USA* **88**: 5974–5978.

Campitelli LF, Yellan I, Albu M, Barazandeh M, Patel ZM, Blanchette M, Hughes TR. 2022. Reconstruction of full-length LINE-1 progenitors from ancestral genomes. *Genetics* **221**.

Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Lemma RB, Turchi L, Blanc-Mathieu R, Lucas J, Boddie P, Khan A, Manosalva Pérez N, et al. 2022. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **50**: D165–D173.

Chen X, Pacis AS, Aracena KA, Gona S, Kwan T, Groza C, Lin YL, Sindeaux RHM, Yotova V, Pramatarova A, et al. 2022. Transposable elements are associated with the variable response to influenza infection. *BioRxiv*.

Choudhary MN, Quaid K, Xing X, Schmidt H, Wang T. 2022. Widespread contribution of transposable elements to the rewiring of mammalian 3D genomes and gene regulation. *BioRxiv*.

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet* **18**: 71–86.

Chuong EB, Elde NC, Feschotte C. 2016. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**: 1083–1087.

Chuong EB, Rumi MAK, Soares MJ, Baker JC. 2013. Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nat Genet* **45**: 325–329.

Chu T, Rice EJ, Booth GT, Salamanca HH, Wang Z, Core LJ, Longo SL, Corona RJ, Chin LS, Lis JT, et al. 2018. Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme. *Nat Genet* **50**: 1553–1564.

Correa RG, Tergaonkar V, Ng JK, Dubova I, Izpisua-Belmonte JC, Verma IM. 2004. Characterization of NF-kappa B/I kappa B proteins in zebra fish and their involvement in notochord development. *Mol Cell Biol* **24**: 5257–5268.

Darlington GJ, Ross SE, MacDougald OA. 1998. The role of C/EBP genes in adipocyte differentiation. *J Biol Chem* **273**: 30057–30060.

Ding Q, Regan SN, Xia Y, Oostrom LA, Cowan CA, Musunuru K. 2013. Enhanced efficiency of human pluripotent stem cell genome editing through replacing TALENs with CRISPRs. *Cell Stem Cell* **12**: 393–394.

Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I, Tothova Z, Wilen C, Orchard R, et al. 2016. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotechnol* **34**: 184–191.

Dorrington MG, Fraser IDC. 2019. NF-κB Signaling in Macrophages: Dynamics, Crosstalk, and Signal Integration. *Front Immunol* **10**: 705.

Du AY, Zhuo X, Sundaram V, Jensen NO, Chaudhari HG, Saccone NL, Cohen BA, Wang T. 2022. Functional characterization of enhancer activity during a long terminal repeat's evolution. *Genome Res*.

Emera D, Wagner GP. 2012. Transformation of a transposon into a derived prolactin promoter with function during human pregnancy. *Proc Natl Acad Sci USA* **109**: 11246–11251.

Fang R, Yu M, Li G, Chee S, Liu T, Schmitt AD, Ren B. 2016. Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res* **26**: 1345–1348.

Fish JE, Cantu Gutierrez M, Dang LT, Khyzha N, Chen Z, Veitch S, Cheng HS, Khor M, Antounians L, Njock M-S, et al. 2017. Dynamic regulation of VEGF-inducible genes by an ERK/ERG/p300 transcriptional network. *Development* **144**: 2428–2444.

Freytag SO, Paielli DL, Gilbert JD. 1994. Ectopic expression of the CCAAT/enhancer-binding protein alpha promotes the adipogenic program in a variety of mouse fibroblastic cells. *Genes Dev* **8**: 1654–1663.

Fuentes DR, Swigut T, Wysocka J. 2018. Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. *eLife* **7**.

Fueyo R, Judd J, Feschotte C, Wysocka J. 2022. Roles of transposable elements in the regulation of mammalian transcription. *Nat Rev Mol Cell Biol* **23**: 481–497.

Gerondakis S, Siebenlist U. 2010. Roles of the NF-kappaB pathway in lymphocyte development and function. *Cold Spring Harb Perspect Biol* **2**: a000182.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**: 576–589.

Hetru C, Hoffmann JA. 2009. NF-kappaB in the immune response of Drosophila. *Cold Spring Harb Perspect Biol* **1**: a000232.

Hirakawa M, Nishihara H, Kanehisa M, Okada N. 2009. Characterization and evolutionary landscape of AmnSINE1 in Amniota genomes. *Gene* **441**: 100–110.

Hogan NT, Whalen MB, Stolze LK, Hadeli NK, Lam MT, Springstead JR, Glass CK, Romanoski CE. 2017. Transcriptional networks specifying homeostatic and inflammatory programs of gene expression in human aortic endothelial cells. *eLife* **6**.

Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, Fine EJ, Wu X, Shalem O, et al. 2013. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* **31**: 827–832.

Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AFA, Wheeler TJ. 2016. The Dfam database of repetitive DNA families. *Nucleic Acids Res* **44**: D81-9.

Ito J, Kimura I, Soper A, Coudray A, Koyanagi Y, Nakaoka H, Inoue I, Turelli P, Trono D, Sato

K. 2020. Endogenous retroviruses drive KRAB zinc-finger protein family expression for tumor suppression. *Sci Adv* **6**.

Jacques P-É, Jeyakani J, Bourque G. 2013. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genet* **9**: e1003504.

Judd J, Sanderson H, Feschotte C. 2021. Evolution of mouse circadian enhancers from transposable elements. *Genome Biol* **22**: 193.

Juric I, Yu M, Abnousi A, Raviram R, Fang R, Zhao Y, Zhang Y, Qiu Y, Yang Y, Li Y, et al. 2019. MAPS: Model-based analysis of long-range chromatin interactions from PLAC-seq and HiChIP experiments. *PLoS Comput Biol* **15**: e1006982.

Kaikkonen MU, Spann NJ, Heinz S, Romanoski CE, Allison KA, Stender JD, Chun HB, Tough DF, Prinjha RK, Benner C, et al. 2013. Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Mol Cell* **51**: 310–325.

Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE, et al. 2010. Variation in transcription factor binding among humans. *Science* **328**: 232–235.

Kelly CJ, Chitko-McKown CG, Chuong EB. 2022. Ruminant-specific retrotransposons shape regulatory evolution of bovine immunity. *Genome Res*.

Kolmykov S, Yevshin I, Kulyashov M, Sharipov R, Kondrakhin Y, Makeev VJ, Kulakovskiy IV, Kel A, Kolpakov F. 2021. GTRD: an integrated view of transcription regulation. *Nucleic Acids Res* **49**: D104–D111.

Kolovos P, Georgomanolis T, Koeferle A, Larkin JD, Brant L, Nikolicć M, Gusmao EG, Zirkel A, Knoch TA, van Ijcken WF, et al. 2016. Binding of nuclear factor κB to noncanonical consensus sites reveals its multimodal role during the early inflammatory response. *Genome Res* **26**: 1478–1489.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. Timetree: A resource for timelines, timetrees, and divergence times. *Mol Biol Evol* **34**: 1812–1819.

Lee HJ, Hou Y, Maeng JH, Shah NM, Chen Y, Lawson HA, Yang H, Yue F, Wang T. 2022. Epigenomic analysis reveals prevalent contribution of transposable elements to cis-regulatory elements, tissue-specific expression, and alternative promoters in zebrafish. *Genome Res* **32**: 1424–1436.

Liu T, Zhang L, Joo D, Sun S-C. 2017. NF-κB signaling in inflammation. *Signal Transduct Target Ther* **2**.

Lynch VJ, Leclerc RD, May G, Wagner GP. 2011. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat Genet* **43**: 1154–1159.

MacArthur S, Brookfield JFY. 2004. Expected rates and modes of evolution of enhancer

sequences. *Mol Biol Evol* **21**: 1064–1073.

Macfarlan TS, Gifford WD, Agarwal S, Driscoll S, Lettieri K, Wang J, Andrews SE, Franco L, Rosenfeld MG, Ren B, et al. 2011. Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. *Genes Dev* **25**: 594–607.

Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. 2013. RNA-guided human genome engineering via Cas9. *Science* **339**: 823–826.

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**: 495–501.

Michida H, Imoto H, Shinohara H, Yumoto N, Seki M, Umeda M, Hayashi T, Nikaido I, Kasukawa T, Suzuki Y, et al. 2020. The Number of Transcription Factors at an Enhancer Determines Switch-like Gene Expression. *Cell Rep* **31**: 107724.

Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, Chang HY. 2016. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods* **13**: 919–922.

Muñoz-López M, García-Pérez JL. 2010. DNA transposons: nature and applications in genomics. *Curr Genomics* **11**: 115–128.

Natoli G, Saccani S, Bosisio D, Marazzi I. 2005. Interactions of NF-kappaB with chromatin: the art of being at the right place at the right time. *Nat Immunol* **6**: 439–445.

Nijman IJ, van Tessel P, Lenstra JA. 2002. SINE retrotransposition during the evolution of the Pecoran ruminants. *J Mol Evol* **54**: 9–16.

Nilsson MA, Klassert D, Bertelsen MF, Hallström BM, Janke A. 2012. Activity of ancient RTE retroposons during the evolution of cows, spiral-horned antelopes, and Nilgais (Bovinae). *Mol Biol Evol* **29**: 2885–2888.

Nishihara H, Smit AFA, Okada N. 2006. Functional noncoding sequences derived from SINEs in the mammalian genome. *Genome Res* **16**: 864–874.

Notwell JH, Chung T, Heavner W, Bejerano G. 2015. A family of transposable elements co-opted into developmental enhancers in the mouse neocortex. *Nat Commun* **6**: 6644.

Oeckinghaus A, Hayden MS, Ghosh S. 2011. Crosstalk in NF-κB signaling pathways. *Nat Immunol* **12**: 695–708.

Ostuni R, Piccolo V, Barozzi I, Polletti S, Termanini A, Bonifacio S, Curina A, Prosperini E, Ghisletti S, Natoli G. 2013. Latent enhancers activated by stimulation in differentiated cells. *Cell* **152**: 157–171.

Pace JK, Gilbert C, Clark MS, Feschotte C. 2008. Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. *Proc Natl Acad Sci USA* **105**: 17023–17028.

Pereira JP, Kelly LM, Cyster JG. 2010. Finding the right niche: B-cell migration in the early phases of T-dependent antibody responses. *Int Immunol* **22**: 413–419.

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**: W160-5.

Romanish MT, Lock WM, van de Lagemaat LN, Dunn CA, Mager DL. 2007. Repeated recruitment of LTR retrotransposons as promoters by the anti-apoptotic locus NAIP during mammalian evolution. *PLoS Genet* **3**: e10.

Sasaki T, Nishihara H, Hirakawa M, Fujimura K, Tanaka M, Kokubo N, Kimura-Yoshida C, Matsuo I, Sumiyama K, Saitou N, et al. 2008. Possible involvement of SINEs in mammalian-specific brain formation. *Proc Natl Acad Sci USA* **105**: 4220–4225.

Satriano J, Schlondorff D. 1994. Activation and attenuation of transcription factor NF-kB in mouse glomerular mesangial cells in response to tumor necrosis factor-alpha, immunoglobulin G, and adenosine 3':5'-cyclic monophosphate. Evidence for involvement of reactive oxygen species. *J Clin Invest* **94**: 1629–1636.

Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves A, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**: 335–348.

Schmidt D, Wilson MD, Spyrou C, Brown GD, Hadfield J, Odom DT. 2009. ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. *Methods* **48**: 240–248.

Schmidt SF, Larsen BD, Loft A, Nielsen R, Madsen JGS, Mandrup S. 2015. Acute TNF-induced repression of cell identity genes is mediated by NFκB-directed redistribution of cofactors from super-enhancers. *Genome Res* **25**: 1281–1294.

Senft AD, Macfarlan TS. 2021. Transposable elements shape the evolution of mammalian development. *Nat Rev Genet* **22**: 691–711.

Sen R, Baltimore D. 1986. Inducibility of kappa immunoglobulin enhancer-binding protein Nf-kappa B by a posttranslational mechanism. *Cell* **47**: 921–928.

Shah AV, Birdsey GM, Randi AM. 2016. Regulation of endothelial homeostasis, vascular development and angiogenesis by the transcription factor ERG. *Vascul Pharmacol* **86**: 3–13.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–1050.

Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T. 2014. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res* **24**: 1963–1976.

Sundaram V, Wysocka J. 2020. Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. *Philos Trans R Soc Lond B Biol Sci* **375**: 20190347.

Tabruyn SP, Griffioen AW. 2007. A new role for NF-kappaB in angiogenesis inhibition. *Cell Death Differ* **14**: 1393–1397.

Tempel S. 2012. Using and understanding RepeatMasker. *Methods Mol Biol* **859**: 29–51.

Todd CD, Deniz Ö, Taylor D, Branco MR. 2019. Functional evaluation of transposable elements as enhancers in mouse embryonic and trophoblast stem cells. *eLife* **8**.

Trizzino M, Park Y, Holsbach-Beltrame M, Aracena K, Mika K, Caliskan M, Perry GH, Lynch VJ, Brown CD. 2017. Transposable elements are the primary source of novelty in primate gene regulation. *Genome Res* **27**: 1623–1633.

Tyner C, Barber GP, Casper J, Clawson H, Diekhans M, Eisenhart C, Fischer CM, Gibson D, Gonzalez JN, Guruvadoo L, et al. 2017. The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res* **45**: D626–D634.

Wells JN, Feschotte C. 2020. A field guide to eukaryotic transposable elements. *Annu Rev Genet* **54**: 539–561.

Zemojtel T, Kielbasa SM, Arndt PF, Behrens S, Bourque G, Vingron M. 2011. CpG deamination creates transcription factor-binding sites with high efficiency. *Genome Biol Evol* **3**: 1304–1311.

Zhang Q, Lenardo MJ, Baltimore D. 2017. 30 Years of NF-κB: A Blossoming of Relevance to Human Pathobiology. *Cell* **168**: 37–57.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.