Widespread roles of enhancer-like transposable elements in cell identity and long-range genomic interactions

Yaqiang Cao,[1,3] Guoyu Chen,[1,3] Gang Wu,[1,3] Xiaoli Zhang,[1,3] Joseph McDermott,[1] Xingwei Chen,[1] Chi Xu,[1] Quanlong Jiang,[1] Zhaoxiong Chen,[1] Yingying Zeng,[1,2] Daosheng Ai,[1] Yi Huang,[1] and Jing-Dong J. Han[1]

[1]*CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China;* [2]*School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China*

A few families of transposable elements (TEs) have been shown to evolve into *cis*-regulatory elements (CREs). Here, to extend these studies to all classes of TEs in the human genome, we identified widespread enhancer-like repeats (ELRs) and find that ELRs reliably mark cell identities, are enriched for lineage-specific master transcription factor binding sites, and are mostly primate-specific. In particular, elements of MIR and L2 TE families whose abundance co-evolved across chordate genomes, are found as ELRs in most human cell types examined. MIR and L2 elements frequently share long-range intra-chromosomal interactions and binding of physically interacting transcription factors. We validated that eight L2 and nine MIR elements function as enhancers in reporter assays, and among 20 MIR-L2 pairings, one MIR repressed and one boosted the enhancer activity of L2 elements. Our results reveal a previously unappreciated co-evolution and interaction between two TE families in shaping regulatory networks.

[Supplemental material is available for this article.]

Transposable elements (TEs) are widespread throughout the genome, covering more than 50% of the human genome (International Human Genome Sequencing Consortium 2001; de Koning et al. 2011). TEs are important contributors to genome complexity (Erwin et al. 2014), evolutionary variation (Biémont and Vieira 2006), and environmental adaptation (Chénais et al. 2012). TEs can be broadly classified into two types: retrotransposons, which copy and paste; and DNA transposons, which cut and paste (Seberg and Petersen 2009). Retrotransposons are more abundant than DNA transposons in the human genome (International Human Genome Sequencing Consortium 2001). Among retrotransposons, short interspersed elements (SINEs) and long interspersed elements (LINEs) are the two most prolific TEs in higher eukaryotes. The genomic distribution for specific families of SINEs and LINEs are associated with each other, and in the human genome, only retrotransposons such as L1, *Alu*, and SVA appear to have transposon activity (Mills et al. 2007). Previous evidence has shown that TEs can be adopted as *cis*-regulatory elements (CREs), donating enhancers, promoters, and insulators to the host genome (Chuong et al. 2016). These TE CREs, which harbor abundant motifs and transcription factor binding sites (TFBSs) (Sundaram et al. 2014), contribute to the gene regulatory networks (Kunarso et al. 2010).

Several genome-wide analyses of epigenetic activities and functions of TEs have been carried out (Xie et al. 2013; Su et al. 2014; Sundaram et al. 2014; Goke and Ng 2016), but an extensive cataloging and survey of TEs as CREs across different human cell types and tissues is still lacking. We therefore carried out a comprehensive TE-centric integrated analysis using data from the ENCODE Project (The ENCODE Project Consortium 2012), NIH Roadmap Epigenomics (Roadmap Epigenomics Consortium et al. 2015), and individual studies.

Results

Integrated framework for *cis*-regulatory TE detection

Accurately assigning reads from ChIP-seq data to TE regions is a key technical issue in the TE analysis field, especially for short reads (Derrien et al. 2012). The low mappability issue for short reads makes TEs genomic "dark matter" (Lee and Schatz 2012). The simplest computational strategy is only using uniquely mapped reads and relying on high mappability regions at boundaries of TEs, which could lead to signal loss at highly repetitive sequences. The other strategy focuses on assigning ambiguous mapped reads to TE families/subfamilies based mainly on mapping quality scores (Day et al. 2010; Wang et al. 2010; Chung et al. 2011; Xie et al. 2013). An alternative method is based on the uniquely mapped reads and whole genome-wide mappability score to carry out signal correction (Cheung et al. 2011; Harmanci et al. 2014).

Binding of co-activators such as EP300 and CBP has been thought of as the golden standard for identifying genome-wide active enhancers (Heintzman et al. 2007; Visel et al. 2009; Wang et al. 2009). However, such co-activator ChIP-seq is not generally feasible (Rajagopal et al. 2013), and currently the epigenetic mapping community such as NIH Roadmap Epigenomics (Roadmap

[3]**These authors contributed equally to this work.**
Corresponding author: jdhan@picb.ac.cn
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.235747.118. Freely available online through the *Genome Research* Open Access option.

29:40–52 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/19; www.genome.org

Epigenomics Consortium et al. 2015) and BLUEPRINT (Adams et al. 2012) prefer using histone modification ChIP-seq data to detect enhancers. Enhancer-like repeats (ELRs) can be obtained from widely used genome segmentation tools such as ChromHMM (Ernst and Kellis 2012) by overlapping putative enhancers and TEs. One problem for this method is that the potential ELRs may locate at the boundary of putative enhancers and thus actually have low coverage of H3K27ac or H3K4me1 signal. The classification model RFECS using the random forest algorithm (Liaw and Wiener 2002) with histone modifications used as features has been shown to detect enhancers with high accuracy (Rajagopal et al. 2013). Therefore, based on TE-centric mappability signal correction, normalization to input data (Supplemental Fig. S1; Supplemental Materials) and the random forest classification model, we selected histone modifications as features from the ENCODE data sets and built the FOFM (forest of forest model) (details in Methods) to detect *cis*-functional TEs. Our repeats-centric classification method FOFM consists of two classification layers. The first layer is used to train a model in cell lines with curated positive and negative training data sets (Methods), and the second layer is used to integrate the classification results from each trained model in the first layer (Methods).

As expected, H3K4me1, H3K27ac, H3K4me2, and H3K4me3 were the most important features (feature importance is defined as the decrease of accuracy if the values for that feature are randomly permuted) for enhancer-like repeats detected by the FOFM, consistent with H3K4me1 being an established mark for active and poised enhancers (Heintzman et al. 2007), H3K27ac a mark for the active regulatory element (Creyghton et al. 2010), H3K4me2 for promoters and enhancers, and H3K4me3 for active transcription start sites (TSSs) (Supplemental Fig. S2A,B; Heintzman et al. 2007). Models built in one cell line showed high performance in other cell lines for each model (all AUCs were >0.9) (Supplemental Fig. S2C). We further trained a second layer random forest classification model to integrate the output from the first layer (Supplemental Fig. S2D). We also carried out the same analysis to detect promoter-like repeats (PLRs) (Supplemental Fig. S3), and as expected, the most important features were H3K4me3, H3K4me2, and H3K9ac (Supplemental Fig. S3B), which are known to be associated with promoters (Karmodiya et al. 2012). The above analysis was all based on unique mapping. To rule out the possibility that different mapping strategies could affect the results, we also built a classification model based on random hit (RH) mapping (Methods), and there was little difference between the performance of models using these two mapping strategies (Supplemental Fig. S2E). When using 10 histone modification features, the FOFM for ELRs has an AUC of 0.942, and for PLRs the AUC is 0.970 (Supplemental Fig. S2E). We also compared the performance of our FOFM model to a multiple forest model (MF). The first layer of the MF model is the same as the FOFM, while in the second layer, the MF model requires more than half of the first layer outputs to support the classification. Although the MF model achieved a little higher AUC for ELRs, according to the zero-one-loss (ZOL) measurement (normalized number of mismatches between true labels and predictions), the FOFM performed significantly better (fewer false positives) and obtained more ELRs (for example, the MF model detects 50,908 ELRs for GM12878, while the FOFM detects 69,013) (Supplemental Fig. S2F). Using the FOFM with unique mapped data, we identified 491,656 unique/nonoverlapping ELRs and 162,525 unique/nonoverlapping PLRs in 15 cell lines and tissues from ENCODE data, and 1,371,646 ELRs and 447,308 PLRs in 82 cell lines and tissues from NIH Roadmap Epigenomics (with histone modifications of H3K27ac, H3K4me1, H3K4me3, H3K36me3, H3K9me3, and H3K27me3 ChIP-seq data). For the total 4,637,389 repeats annotated by RepeatMasker (Smit et al. 2013–2015) in human (hg38) (we did not analyze simple repeats such as [GTTAGG]n), 1,642,395 (35.42%) repeats were identified as ELRs or PLRs. Corresponding data sets are available at http://www.picb.ac.cn/hanlab/cisTEs. The detected ELRs highly overlap with the typical enhancers defined by ChromHMM (Supplemental Fig. S4A,B). Further comparison between the FOFM-identified ELRs and ChromHMM-enhancer-overlapped TEs (with the requirement that more than half of TEs were overlapped) in GM12878, HeLa, HepG2, and K562 indicate the FOFM detected more reliable unique ELRs by showing higher H3K27ac and EP300 ChIP-seq signal, while many unique ChromHMM-enhancers-overlapped TEs may locate at the enhancer boundaries and thus have lower H3K27ac and H3K4me1 signal (Supplemental Fig. S4C–F).

The majority of TEs are potential regulatory elements

We first examined TE family enrichment for ELRs and PLRs in different cell lines. Based on the results using the FOFM with the ENCODE (using either unique mapping [Fig. 1A] or RH mapping strategy [Supplemental Fig. S5A]) and NIH Roadmap Epigenomics data (consolidated [Fig. 1B] and nonconsolidated [Supplemental Fig. S5B]), most TE families showed varied enhancer activity across cell types, while MIR and L2 were consistently enriched in each cell type. Consistent with previous studies, we found MIR and L2 are enriched for enhancer activities in GM12878 and K562 (Fig. 1A; Huda et al. 2011; Jjingo et al. 2014) and a specific enrichment of the endogenous retroviral sequences ERV1 and ERVL as potential enhancers in pluripotent stem cells (Fig. 1B; Supplemental Fig. S5B; Kunarso et al. 2010; Wang et al. 2014).

To estimate whether we detected all TEs in the genome that could function as CREs, we performed a saturation analysis using cumulative ratios (number of unique sets of ELRs/PLRs compared to all TEs in the genome). Even with the use of 82 cell lines from NIH Roadmap Epigenomics, detection did not reach saturation, which showed that more TEs may function as CREs if using more data and substantially more TEs potentially functioning as enhancers than promoters (Fig. 1C). We further corroborated the saturation analysis results using TEs overlapped within 1kb of the EP300 ChIP-seq peak summits (Fig. 1D). The highly cell-/tissue-specific bindings of EP300 at TEs (55.51%) are consistent with the observation that the majority of TE-derived TFs bindings were cell-type–specific (Sundaram et al. 2014).

As an indication of biological relevance and functional importance of these ELRs, 74.35% of the ELRs are in the vicinity of 10 kb from gene expression quantitative trait loci (eQTLs) (Fig. 1E). In addition, there is a greater number of ELRs with their centers within 10 kb of their nearest eQTL than the number of random background sequences (similar lengths) that are within 10 kb of nearest eQTLs (Fig. 1E), suggesting a functional association of ELRs with gene expression and cellular processes.

Another indication of the functional relevance of these ELRs is that there are more ChIA-PET and Hi-C long-range PETs linking them with TSSs than the background (ELRs' nearby TEs belonging to the same family which were not classified as ELRs or PLRs) using data from GM12878 cells (Fig. 1F,G).

ELRs mark cell identities

ELRs show high tissue- and cell-type–specificity compared to PLRs, as PLRs have very similar presence/absence profiles across samples

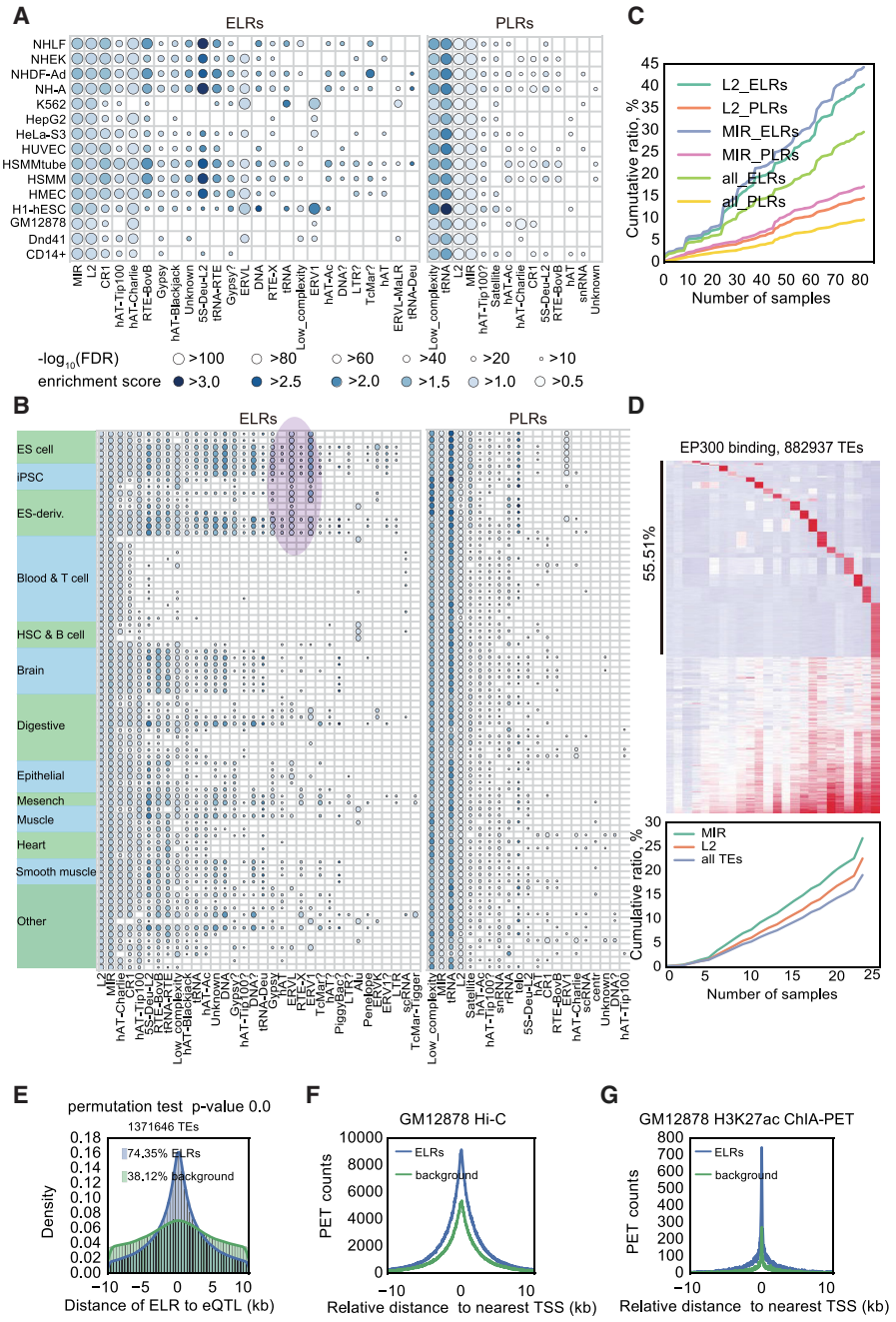

Figure 1. Enhancer- and promoter-like repeat elements (ELRs and PLRs) in human tissues and cell lines. (*A*) TE families enriched for ELR and PLR in different ENCODE cell lines based on histone modification ChIP-seq data using a unique mapping strategy. Bubble size indicates corrected enrichment *P*-value and color marks enrichment score. The enrichment test was performed with a combination of the binomial test and hypergeometric test (Methods). (*B*) TE families enriched for ELRs and PLRs in NIH Roadmap Epigenomics cell lines based on the consolidated histone modification ChIP-seq data. Tissue-specifically enriched TE families such as ERV1 and ERVL in ESCs and iPSCs are marked by lavender ellipses. (*C*) Cumulative ratio of ELRs and PLRs among MIR, L2, and all TEs in human tissues in NIH Roadmap data. (*D*) Saturation estimation of active TEs bound by EP300. *Upper* panel is the heat map of detected active TEs bound by EP300 in human cell lines or tissues collected from GEO; each column is a sample, and each row is a TE; values in the heat map are binary, and 1 (red) means the TE is bound by EP300; *lower* panel is the cumulative ratio of the active TEs among MIR, L2, and all TEs; 55.51% of the EP300-bound TEs are restricted to predominately one tissue or cell. (*E*) Distance distribution of centers of ELRs to nearest eQTLs from GRASP database; 74.35% of the ELR centers have eQTLs in vicinity of 10 kb. The permutated background was the mean value by sampling the same number and same length sequences to the all ELRs 100 times. (*F*,*G*) Hi-C (GSM1551552) PETs and H3K27ac ChIA-PET (GSE59395) PET frequencies linking ELRs and TSSs, compared to the same number of nearest non-ELR and non-PLR TEs in the same families (background).

(Fig. 2A). Quantitatively, the Jensen-Shannon Divergence (JSD) entropy (Cabili et al. 2011) (measurement of tissue specificities) of ELRs is similar to typical enhancers (defined by ChromHMM from NIH Roadmap Epigenomics); both are much higher than PLRs, typical promoters, or the non-TE overlapping fragments in typical promoters or enhancers, indicating TEs contribute to tissue specificities of typical enhancers (Fig. 2B). To investigate whether ELRs can distinguish cell identities, we built a neighbor-joining tree (NJT) using ELRs. The NJT built by ELRs showed an accurately categorized structure based on known lineage groups (cluster purity = 0.812, where cluster purity = 1.0 – [misaligned dendrites/total dendrites]) (Fig. 2C). A previous study has reported that NJT built using ±1.5 kb H3K4me1 densities of long-intergenic noncoding RNAs (lincRNAs) TSSs is the best lineage classification method (Amin et al. 2015), and we find the NJT built by ELRs (cluster purity = 0.812) (Fig. 2C) is better based on cluster purity (cluster purity = 0.681) and the overall structure (Supplemental Fig. S6A). Also, the NJT built by ELRs is better than the NJT built by ±1.5 kb H3K4me1 densities of all lncRNAs (including lincRNAs) TSSs (cluster purity = 0.681) (Supplemental Fig. S6B) and the NJT built by typical enhancer fragments with no TE overlap (cluster purity = 0.754) (Supplemental Fig. S6C).

We further validated the accuracy and generality of ELRs in marking cell identities by using known cell identities of normal blood samples and blood disease samples from the BLUEPRINT (Adams et al. 2012) (cluster purity = 1.000) (Supplemental Fig. S7A,B) and CEEHRC (Bae 2013) data (cluster purity = 0.872) (Supplemental Fig. S7C,D).

ELRs are enriched for motifs and binding sites of tissue-specific master regulators, and are primate-specific

To understand how ELRs mark cell identities, we focused on tissue-specific ELRs (tsELRs) and searched for their upstream regulators and downstream effectors. Using ELRs with JSD > 0.7, we defined 11 clusters of tissue-specific ELRs (Fig. 2D; Supplemental Fig. S9). The nearest genes of tsELRs in each cluster are enriched for tissue-specific functional GO terms (Fig. 2E), suggesting roles of ELRs in regulating tissue-specific functions (Xie et al. 2013; Jjingo et al. 2014). For

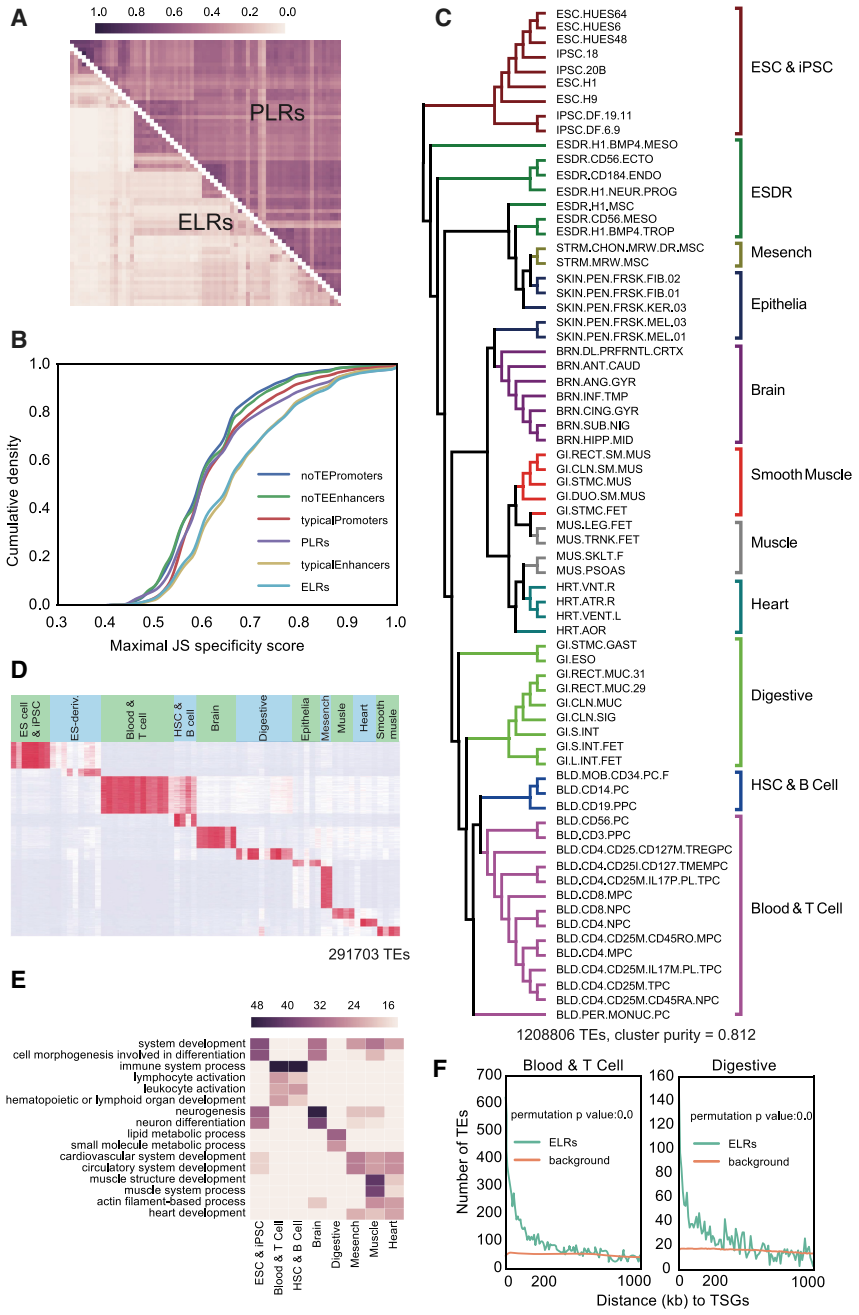

Figure 2. ELRs mark cell identities. (*A*) Correlations of each pair of tissues and cell lines based on their ELRs and PLRs presence/absence (1/0) profiles within the tissues. Cells are grouped by annotation from NIH Roadmap Epigenomics. (*B*) Cumulative density of maximal tissue specificity metric Jensen-Shannon divergence (JSD) for all ELRs, PLRs, typical enhancers/promoters defined by ChromHMM and the fragment sequences from typical enhancers/promoters without TEs overlap that match any tissue-specific pattern. (*C*) Neighbor joining tree (NJT) of the NIH Roadmap Epigenomics samples based on all ELRs' presence/absence profiles, with the ESC and iPSC branch set as the root. (*D*) Heat map of the ELRs (maximal JSD > 0.7) binary matrix, from 1 (red) to 0 (white), across different NIH Roadmap Epigenomics cell lines. (*E*) Heat map of enrichment *P*-values for the top enriched GO terms of each cluster. (*F*) Frequency histogram of absolute distances from each TE to the nearest tissue-specific genes (TSGs) in the group of Blood and T cell and Digestive tissues.

example, the tsELR cluster of immune cells (Blood and T cell, HSC and B cell) is enriched for genes in the "immune system process," and the digestive tissue tsELR cluster (Digestive) enriches for genes in the "lipid metabolic process" (Fig. 2E). Based on the Jensen-Shannon Divergence, we also defined tissue-specific genes (Supplemental Fig. S8A; Supplemental Materials). Tissue-specific ELRs are highly enriched within ~200 kb around the TSSs of tissue-specific genes (permutation FDR = 0.00) (Fig. 2F; Supplemental Fig. S8B–D).

We inferred upstream regulators of the embryonic stem cell (ESC) (ES cell and iPSC) cluster, for which there are rich resources of published ChIP-seq data. Motif analysis for ESC tsELRs showed that ESC master regulators, POU5F1, NANOG, and SOX2 (Boyer et al. 2005; Loh et al. 2006), ranked as the top one, two, and five motifs, respectively (Fig. 3A). Furthermore, most of the mean motif densities for the top 10 motifs in the ESC tsELRs are higher than the ESC-specific typical enhancers or the typical enhancer fragments without overlap with TEs (Fig. 3B). This suggests that these ELRs harbor major binding sites for TFs that are important to ESC identity, which is consistent with previous findings that TEs are a source of genomic regulation (Jacques et al. 2013; Sundaram et al. 2014; Trizzino et al. 2017). We further validated this by using TF binding site data from Cistrome (Fig. 3C, gray bars; Liu et al. 2011) and HUES64 ESC ChIP-seq data (Fig. 3C, green bars; Tsankov et al. 2015); heat maps of these top 10 TFs also show that these tsELRs are enriched at the centers of TF binding sites (Fig. 3D; Supplemental Fig. S10A). The tsELRs also show tissue-specific hypomethylation, and tissue-specifically expressed non-poly(A) and nascent RNA at a low level (Supplemental Fig. S10B–F), similar to enhancer RNAs (eRNAs) (Mousavi et al. 2013; Arner et al. 2015), further indicates that tsELRs could encode lncRNAs (Kapusta et al. 2013).

We next studied whether ELRs are conserved between species. Starting from ESC- and iPSC-specific ELRs, we mapped them to representative species from the topology of the UCSC vertebrate phylogenetic tree (Fig. 3E; Karolchik et al. 2012) according to sequence similarities (Methods). The mapping ratios decrease as the evolutionary distances increase. Generally, the mapping ratios are above 60% in primates, with up to 100% in chimpanzee and gorilla, while less than 1% in chicken, frog, and zebrafish (Fig. 3F). Moreover, different from the UCSC phylogenetic tree or the annotated evolutionary-separation years by TimeTree (Hedges et al. 2015), the tsELRs mapping ratio showed that laurasiatheria are closer to primates than euarchontoglires. The enrichment for master TF motifs in mapped sequences showed

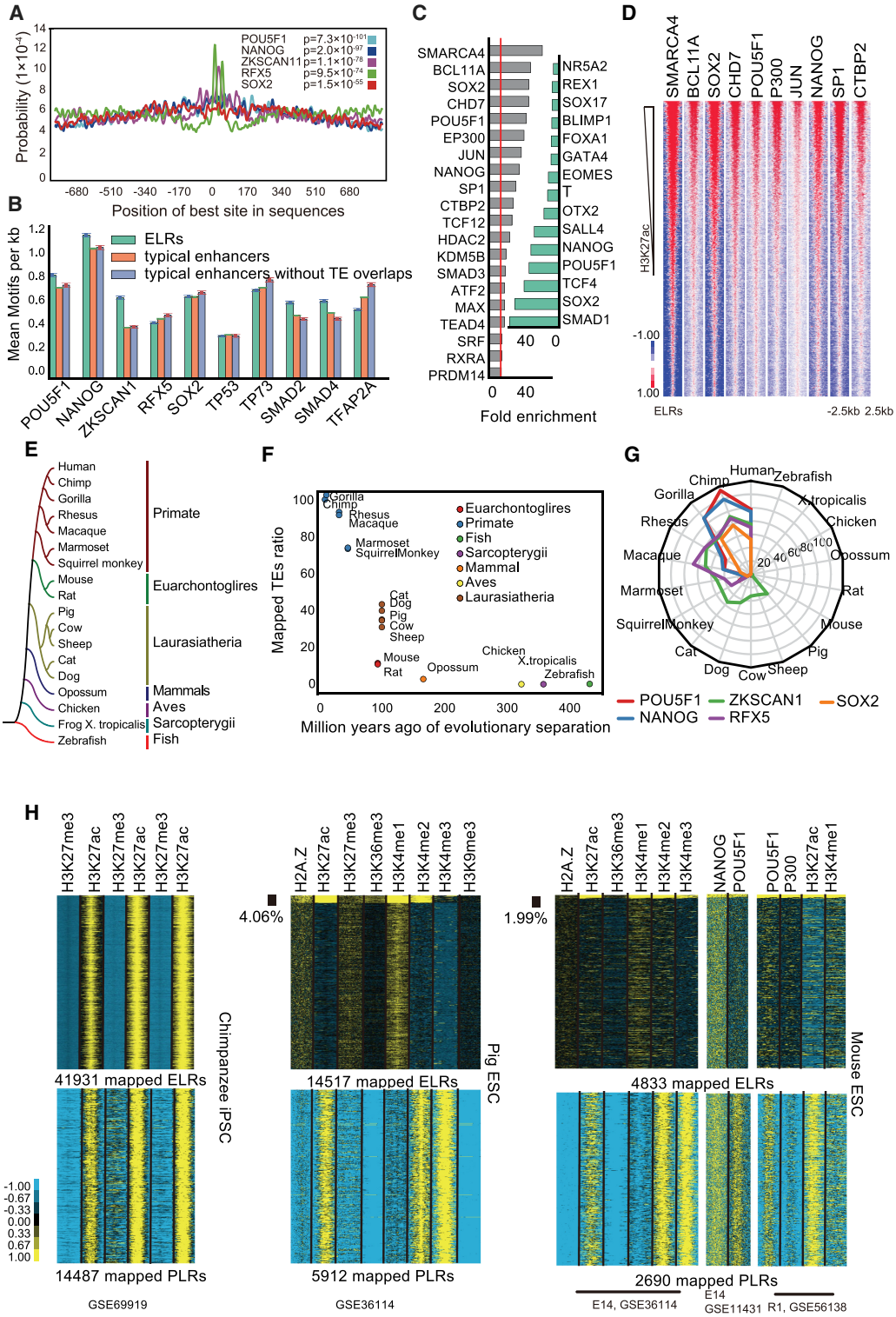

Figure 3. hESC- and iPSC-specific ELRs mark the master TF binding sites. (*A*) Top five enriched motifs detected by CentriMo around centers of ESC-specific ELRs. (*B*) Mean motif densities for the top 10 enriched TF motifs on ESC-specific ELRs, ESC-specific typical enhancers, and ESC-specific typical enhancer fragments that do not overlap with any TE. (*C*) Top enriched TFs bound at the ELRs in the ESC-specific module as determined by ChIP-seq targets, sorted by fold enrichment of TF binding over the background of all TEs. The *left* and *right* graphs are based on hESC and iPSC data from Cistrome, and data from HUES64 (GSE61475), respectively. The red line indicates the mean fold enrichment of all 67 TFs that have overlaps with tsELRs over the background of all TEs. (*D*) Profiles of the top 10 TFs in the *left* graph of panel *D* across −2.5 to +2.5 kb centered around tsELRs. The ELRs are sorted by the mean H3K4me1 intensity. Data sources: SMARCA4 (GSM602297), POU5F1, EP300, and NANGO, CTBP1 (H1-hESC, ENCODE), BCL11A and SP1 (GSE32465), SOX2 (GSM1364026), CHD7 (GSM831027), and JUN (GSM935614). (*E*) Selected representative vertebrates on the topology structure of phylogenetic tree from UCSC. (*F*) The mapping ratios of ELRs in the hESC-specific module in select vertebrate species. Species ages are obtained from TimeTree. (*G*) Radar plot of the $-\log_{10}$ adjusted enrichment *P*-values determined by CentriMo for top enriched motifs within hESC-specific ELRs in select vertebrates. (*H*) Histone marks and TF binding profiles in chimpanzee iPSCs (*left* panel, GSE69919), pig ESCs (*middle* panel, GSE36114), and mouse ESCs (*right* panel, GSE11431, and GSE56138) at −2.5 to +2.5 kb around mapped hESC-specific ELRs (*upper* panel) and PLRs (*bottom* panel). Only <5% and <2% of mapped hESC-specific ELRs show enhancer-like histone modifications in pig and mouse, respectively.

a primate-specific pattern (Fig. 3G). We further examined the histone modification profiles of the mapped ELRs and PLRs and found that most ELRs mapped in chimpanzee iPSC showed enhancer profiles, while only 4.06% of ELRs mapped in pig ESCs did (Fig. 3H) and only 1.99% for mapped ELRs in mouse ESCs did (Fig. 3H). The high conservation of ELRs and PLRs between human and chimpanzee is consistent with a previous study, which showed that the majority of *cis*-regulatory elements that are conserved among primates are a source of genomic regulation (Trizzino et al. 2017). In contrast, the mapped PLRs sequences showed epigenetic features of PLRs in all three species (Fig. 3H), which coincides with a previous study that found that liver promoters are partially or fully conserved across 20 mammalian species (Villar et al. 2015). Collectively, we show human ELRs contribute to primate-specific regulatory sites in that (1) the majority of human tsELR sequences are absent in nonprimate species, (2) the majority of those human tsELR sequences that can be mapped to other nonprimate species do not harbor the cell-identity related TF motifs as in primates, and (3) those mapped TE sequences do not contain histone modification enhancer marks. All these together point out that ELRs are a driving force for primates' specific regulatory innovations, contributing to newly evolved *cis*-regulatory elements (Sundaram et al. 2014; Trizzino et al. 2017).

Interactions between MIR and L2 elements

In contrast to the majority of ELRs, the SINE element MIR (mammalian-wide interspersed repeats) and the LINE element L2 are retrotransposon families enriched for both ELRs and PLRs across all the tissues and cell types we examined (Fig. 1A,B; Supplemental Fig. S5B).

There are ~0.5 million MIR and L2 sequences, respectively, in the human genome, as annotated by RepeatMasker (Smit et al. 2013–2015). According to a previous conserved segment sequences analysis (Silva et al. 2003), MIR and L2 are under strong selective constraint. We wondered whether the conservation of sequences is due to their functions as CREs. Therefore, we investigated mouse ENCODE (Mouse Encode Consortium et al. 2012) data and found that MIR and L2 could also be classified as ELRs and PLRs in mouse (Supplemental Fig. S11A); cumulative analysis also indicates there are more potential mouse ELRs or PLRs if more data are available (Supplemental Fig. S11B). Mouse EP300 bindings also showed an unsaturated detection trend (Supplemental Fig. S11C).

We examined the overlapping or shared mappable ELRs in matched human and mouse tissue/cell lines using the Jaccard Index (JI) (Fig. 4A). MIRs show the highest overlap between human and mouse as measured by the JI, when compared to the other two TE families, L2 and hAT-Charlie—which show generality as ELRs in most tissues examined—and *Alu*, known as primates-specific (Arcot et al. 1995), included as a negative control (Fig. 4A).

Subfamily level enrichment analysis using the ENCODE and NIH Roadmap Epigenomics data also revealed that all subfamilies of MIR (MIR, MIRb, MIRc, MIR3, and MIR1_Amn), and two subfamilies of L2 (L2b and L2c) were highly enriched for both ELRs and PLRs (Supplemental Figs. S12, S13). Consistent with the prevalence of high enrichment of MIR and L2 in ELRs in nearly all tissues and cell lines, among all TE families, the detection saturation levels for MIR and L2 ELRs are the highest, with ~45% and 40% of uniquely mappable MIRs and L2s in the genome identified as ELRs across 82 tissues and cell lines, which was still unsaturated, with the number of MIR ELRs detectable still sharply ascending when more cells/tissues or disease conditions are added (Fig. 1C). This suggests that when more samples are profiled, it may show the majority of MIRs and L2s can serve as ELRs or PLRs in at least one cell type. Indeed, although MIR and L2 serve as ELRs in general, subpopulations of the MIR and L2 elements tend to be tissue-specific. The tissue-specific elements are much more prevalent (53.23% and 52.68% for MIR and L2 ELRs, respectively, requiring JSD > 0.6) (Supplemental Fig. S14A) than tissue-nonspecific MIR and L2 ELRs.

The genomic ratios of MIR and L2 sequences are highly correlated across chordate species at an ~1:1 ratio (Pearson correlation coefficient [PCC] = 0.954) (Fig. 4B). The correlation is much higher than between *Alu* and L1 (PCC = 0.332) (Fig. 4B), two elements that share the same retro-transposition enzyme (Cordaux and Batzer 2009). Though MIR and L2 elements are close in the human genome, with a mean distance of ~1 kb for the nearest pairs, the nearest distances detected for those in ELRs and PLRs are more distant (Supplemental Fig. S14B). This indicates that the reason MIR and L2 are detected as PLRs and ELRs together is not because they are in a very close linear location; thus, we investigated whether they colocalize in a 3D manner. Based on the support and confidence metrics of the association rule mining algorithm Apriori (Agrawal and Srikant 1994), we found a strong association between MIR and L2 between interacting anchors pulled down by RAD21, EP300, H3K4me1-3, and H3K27ac antibodies (Supplemental Fig. S4C). For example, interactions between MIR and L2 are highly enriched with H3K4me1, H3K27ac, and H3K4me3 ChIA-PET interacting anchors both between all MIR and L2 (Fig. 4C, lower left; Supplemental Fig. S14C,D) and between MIR and L2 ELRs (Fig. 4C, lower right; Supplemental Fig. S14E). Although the current 3D genome mapping data do not allow the precise distinction of nearby TEs, at the whole genome level, MIR-L2 interactions are overrepresented compared to random background (generated as the same loop size and same loop number as the true loops but with randomly selected regions as loop anchors, for 1000 times) as shown by their significantly above background support and confidence levels in most of the data set examined, while the support and confidence of MIR-MIR and L2-L2 interactions are lower than between L2-MIR in the majority of the data sets (Supplemental Fig. S14F), suggesting that the interactions between MIR-L2 at the loops level are unlikely due to nearby MIR-MIR or L2-L2 interactions. We further analyzed loops called by cLoops (https://github.com/YaqiangCao/cLoops) using K562 H3K4me1 and H3K27ac ChIA-PET data (GSE59395) (Heidari et al. 2014). Overall, 94% of H3K4me1 and 95% of H3K27ac loops have repeats in their anchors (H3K27ac: 1456/1549; H3K4me1: 3367/3528), and more than 24% of loops have MIR on one anchor and L2 on the other (H3K27ac: 387/1549; H3K4me1: 1125/3538), with 573 total overlapped loops between H3K27ac and H3K4me1, among which 111 (19.4%) loops have MIR on one anchor and L2 on the other. We randomly selected six examples, shown as Supplemental Figure S15, A–F. We also corroborated the interaction between MIR and L2 using paired-end tags from Hi-C data (PETs within 10 kb were removed) (Fig. 4D; Supplemental Fig. S14I–N) and ChIA-PET data (Supplemental Fig. S14G,H).

There are 272 TFs with enriched motifs on MIR ELRs (Centrimo log_adj_*P*-value > 4), 253 TFs on L2 ELRs (Centrimo log_adj_*P*-value > 4), and 119 TFs are shared between the two, based on a total of 641 motifs recorded in the MEME motif database (HOCOMOCOv10_HUMAN_mono_meme) (Bailey et al. 2006). The high enrichment for TF motifs on MIR and L2 ELRs indicates that they provide a rich source of TF binding sites and therefore are potential regulatory elements. MIR and L2 ELRs are

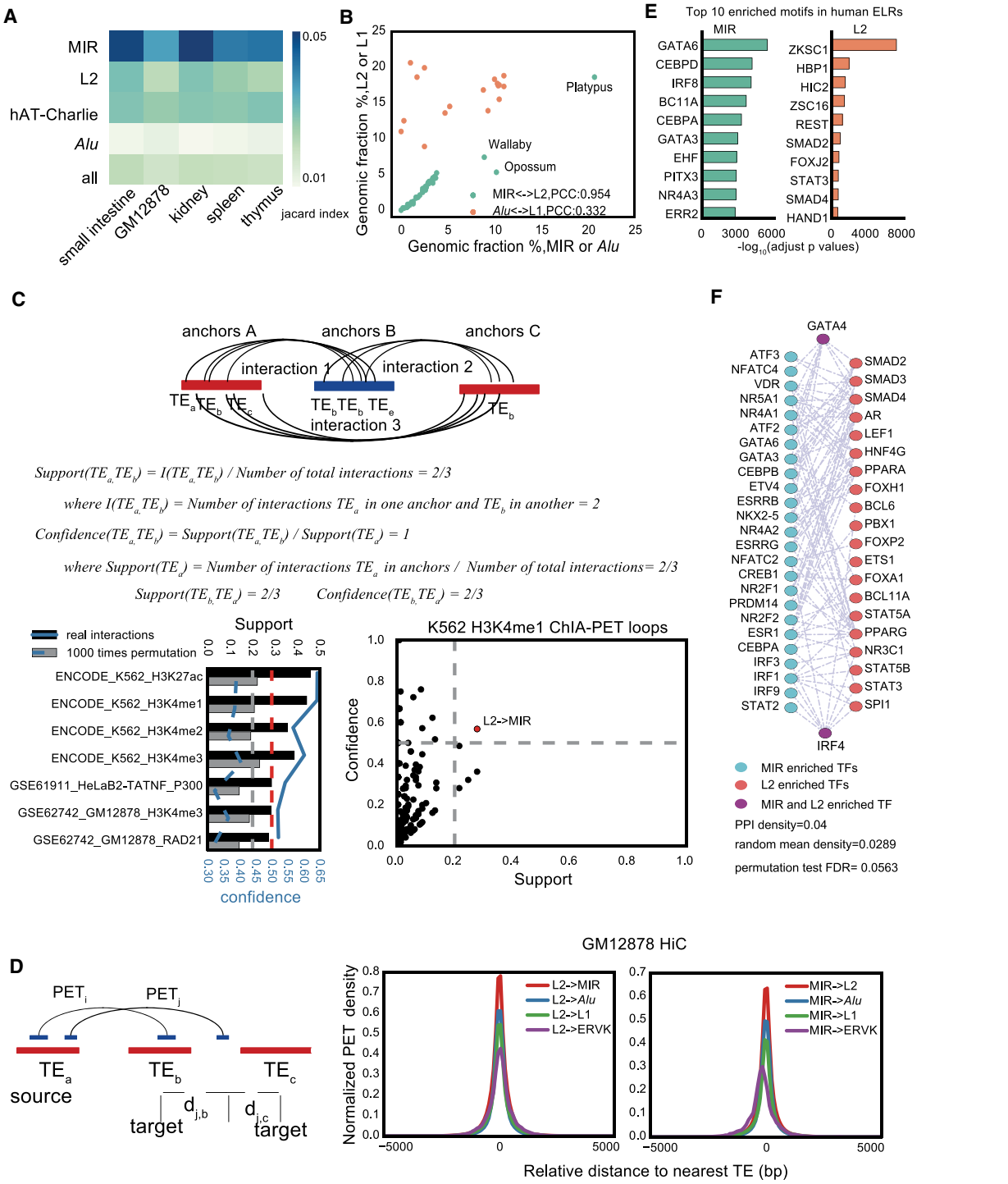

Figure 4. Association between MIR and L2. (*A*) Heat map of Jaccard index between the same type of human and mouse tissues or cell lines for human and mouse overlapping ELRs. (*B*) Correlated genomic ratio between MIR and L2, or *Alu* and L1 across chordate species. Three outlier points (all Australian species) are labeled. (PCC) Pearson's correlation coefficient. (*C*) Support and confidence for association between L2 to MIR between ChIA-PET loop anchors; only significant samples were shown. *Lower left* panel, the gray bars and dashed lines indicate support and confidence background expectation for L2 to MIR interactions in the same number of randomly selected regions as ChIA-PET interaction anchors. *Lower right* panel, the H3K4me1 ChIA-PET interactions from L2 to MIR (both are ELRs) in K562 cells, which is the case in the *lower left* panel. Black (real loops) and gray (random background) bars measure the support; blue line (real loops) and blue dashed line (random background) indicate the confidence; gray and red dashed line indicate the support (0.2) and confidence (0.5) cutoffs. Red dots mark those that have significantly higher support and confidence compared to the all possible TE family pairs background. (*D*) Density of distant L2-interacting tags that fall into MIR, *Alu*, L1, ERVK, or another L2 based on the dense 1-kb resolution in situ Hi-C data of GM12878 (GSM1551552). ERVK is included as a negative control. Density of distantly L2 (or MIR)-interacting tags that fall into MIR (or L2), *Alu*, L1, and ERVK. (*E*) Top 10 enriched motifs on MIR and L2 ELRs in human cells. (*F*) Protein-protein interaction network (STRING v10.0) among top 50 TFs that have enriched motifs on MIR and L2 ELRs.

bound by many important but different TFs in their top 10 motif-enriched TFs (Fig. 4E), with the top 50 motif-enriched TFs on MIR and L2 elements sharing significantly more protein-protein interactions (PPI) (STRING [v10.0]) (Szklarczyk et al. 2014) (evidence score > 900; TFs that do not have links in the PPI were not shown) than random TF background (permutation test FDR = 0.0281) (Fig. 4F). Some of the top TF motifs are conserved in mouse, such as IRF8 motif within MIR, and the HBP1 motif within L2 (Fig. 4E; Supplemental Fig. S11D).

To confirm the enhancer activity of MIR and L2 ELRs, we selected and synthesized 15 L2 and 15 MIR active ELR sequences for further testing, which are either common between HeLa and HepG2 cells, HeLa-specific, or HepG2-specific according to the H3K27ac, H3K4me1 profile, and EP300 binding (Fig. 5A). In HeLa cells, eight L2 ELRs and nine MIR ELRs show significantly higher enhancer activity than empty vector in the luciferase reporter assay by a fold change > 2 (Fig. 5B). Though there was significant activity for one HepG2-specific MIR and one L2 ELR in HeLa cells, this may be due a lack of a repressive chromatin environment in the reporter construct. The recently improved STARR-seq (self-transcribing active regulatory region sequencing), a high-throughput parallel method similar to luciferase reporter assays that overcomes several technical issues that may lead to unreliable measurement of enhancer due to plasmid transfection (by utilizing only the bacterial plasmid origin of replication [ORI] as the core promoter to prevent dual promoter-induced false positives, and by adding inhibitors to IFN-I–inducing kinases to prevent IFN-I response-induced false positives), has been used to assess enhancer activities at the whole-genome scale (Muerdter et al. 2017). Therefore, besides the small number of validated ELRs by luciferase

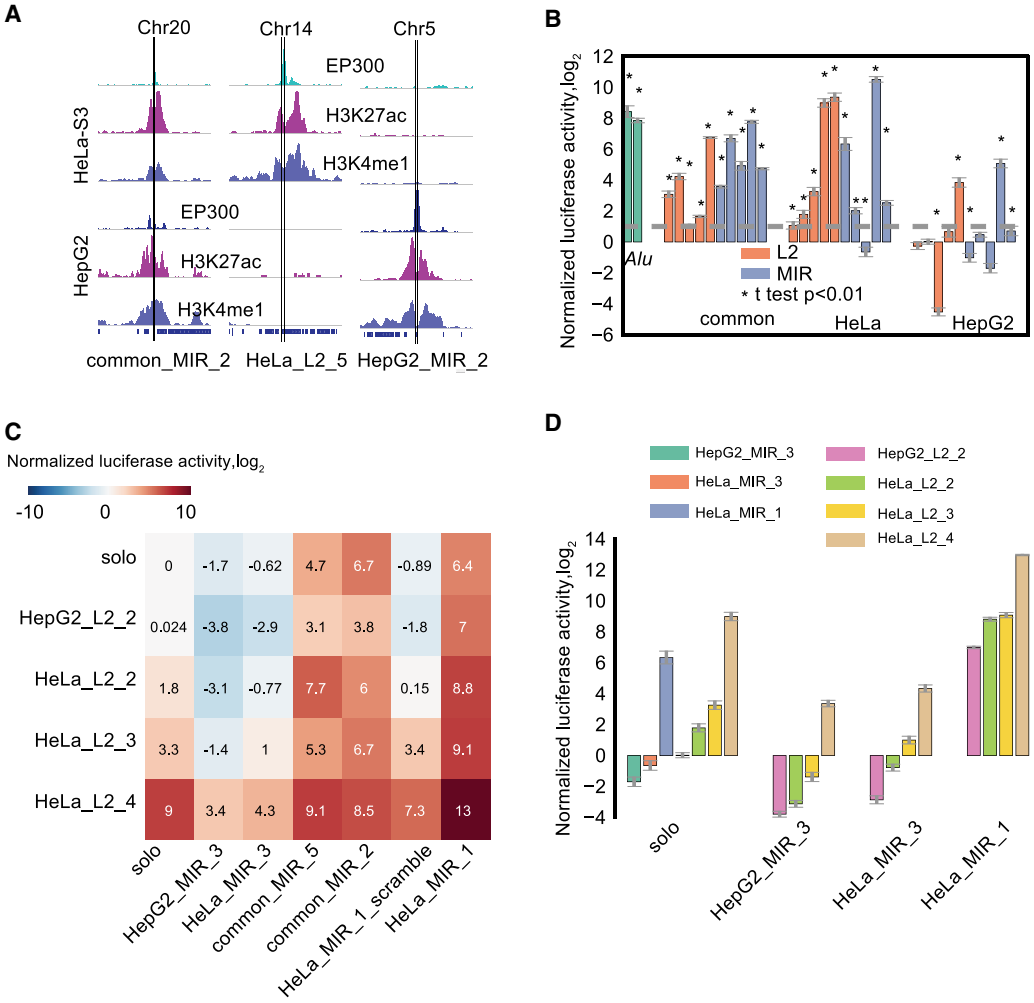

Figure 5. Experimental validation of the enhancer activity and interaction between MIR and L2. (*A*) H3K27ac, H3K4me1, and EP300 profiles for three example cases of MIR and L2 ELRs used for enhancer reporter assay. (*B*) Experimental validation of the enhancer activity using the luciferase reporter assay in HeLa cells. Fifteen MIR and 15 L2 ELRs were synthesized and subcloned into the luciferase reporter vector. The *y*-axis shows the $\log_2$ transformed normalized firefly versus *Renilla* luciferase activity compared with empty vector (fold change). Two *Alu* sequences showing the highest enhancer activity from Su et al. (2014) were used as a positive control. Each measurement was based on three biological replicates. (*) *t*-test *P*-value < 0.01 when compared to empty vector. The gray dashed line indicates the activity fold change of 1 compared to empty vector. (*C*) Enhancer activity for combinations of MIR and L2 using the luciferase reporter assay in HeLa cells. Mean $\log_2$-transformed activities normalized to empty vectors of three biological replicates are shown in the heat map. Each row is an L2 selected from *B* sorted by the activity in HeLa cells; each column is a MIR, and "solo" shows the effect of a single sequence before combination. HepG2_MIR_3 and HeLa_MIR_1_scrambled are negative controls for MIR, HepG2_L2_2 is a negative control for L2. (*D*) Enhancer activity measured by luciferase reporter assay in HeLa cells for the repressor and booster MIRs (labeled *below* the bars) on L2 elements.

reporter assays, as an additional evidence to support the global enhancer activities of our identified ELRs from ChIP-seq data, we further analyzed the improved STARR-seq (Muerdter et al. 2017). The genome-wide enhancer activities of ELRs in HeLa cells were confirmed via the improved STARR-seq both by the much higher mean signal of STARR-seq over input on ELRs and by the heat map of STARR-seq signal around ELRs (Supplemental Fig. S16A). In addition, our independently validated HeLa_MIR_1 (Supplemental Fig. S16B) and HeLa_L2_5 ELRs (Supplemental Fig. S16C) overlap with two STARR-seq peaks.

To test the interactions between L2 and MIR ELRs, we generated 20 pairwise combinations of the individual MIR and L2 to drive the luciferase reporter gene expression according to our validated elements (Fig. 5B). Compared to a scrambled MIR sequence, a MIR ELR (HeLa_MIR_1) augmented the enhancer activity of all L2 ELRs tested in a multiplicative manner (before $\log_2$ transformation), two other enhancer-like MIRs in HeLa (common_MIR_5 and common_MIR_2) augmented the enhancer activities of L2 ELRs in an additive way (before $\log_2$ transformation), and a HepG2-specific MIR ELR (HepG2 MIR_3) suppressed the enhancer activity of all the HeLa cell enhancers (Fig. 5C,D). However, though HeLa_MIR_3 was classified as an enhancer and has high H3K27ac, H3K4me1, and EP300 ChIP-seq signal, it actually functions as a repressor according to the luciferase reporter assay result (Fig. 5B–D). The L2 ELRs tested here did not strongly influence activities of MIR ELRs. These results indicate that some MIR ELRs augment, while some repress L2 ELRs' enhancer activities.

Discussion

Here, we studied the roles of TEs in gene regulatory networks as a repertoire of potential enhancers and promoters to show that this phenomenon goes beyond any particular TE family—it is instead widespread for many TE families, especially MIR and L2. Our analysis also shows that ELRs in general tend to interact more with promoters than flanking regions, pointing to an important role of the ELRs in the organization of the 3D genomic structure.

We found that ELRs precisely mark cell identities and show high tissue specificity and primate specificity, which raises many questions. Why are they tissue-specific? Why do they appear in the vicinity of tissue-specifically expressed genes? Are tissue-specific master TFs that have enriched binding sites on these ELRs involved in their selection?

Nearly 40% of L2 and MIR repeat elements in human show epigenetic profiles of enhancers. The use of 82 human cell types resulted in unsaturated detection of these elements; thus, with more tissues and cell lines profiled, especially those of diseases or cancers, more ELRs can be identified. This leads to two conjectures: (1) Many TEs, such as L2 and MIR TEs, might function as enhancers in one or more tissues; and (2) different elements of the same repeat family may evolve different TF binding sites and serve as different tissue-specific enhancers in different tissues. How such selection occurs remains an intriguing question.

An earlier report identified a MIR element serving as an enhancer booster (Smith et al. 2008). Our results show that this might be a more general phenomenon, where MIR elements interact extensively with L2 in the 3D genome and MIR ELRs potentially act as either general repressors or boosters toward L2 ELR enhancer activity. This is consistent with the observation that genomic ratios of MIR and L2 are highly correlated across different genomes. However, how such ratios are maintained during evolution is still an open question. It would be interesting to see whether long distance genomic interactions mediated by common binding TFs and physically interacting TFs on L2 and MIR, or even noncoding transcripts derived from the two TE families, play a role in such evolutionary selection.

In most studies of disease-associated mutations, mutations in TE regions are often ignored—for example, in the mapping of mutations in acute myeloid leukemia (Papaemmanuil et al. 2016). However, our observation that eQTLs are enriched in the vicinity of ELRs suggests that mutations/variants in such TE regions may have functional consequences and can no longer be neglected. Our maps of ELRs in the majority of tissues and a number of cell lines provide a catalog of the ELRs to be examined for disease association and gene expression regulation, extensive annotation of functional variants and mutations, as well as tracing cell lineages and their master regulatory TFs, and identifying new roles of TEs in 3D genomic structures.

Methods

Genome reference sequences and annotations

Soft-masked assembly sequences of hg38 and mm10 were downloaded from the UCSC Genome Browser (Karolchik et al. 2012). Genome annotation files were downloaded from GENCODE (human: v21, mouse: vM4) (Harrow et al. 2012). Repeat annotations annotated by RepeatMasker (Smit et al. 2013–2015) for all species used in this paper were downloaded from the UCSC Genome Browser according to the used genome version. LiftOver chain files were downloaded from the UCSC Genome Browser. Based on GENCODE annotations, repeats were classified into 5′ UTR, exon, intron, 3′ UTR, proximal upstream, proximal downstream, distal upstream, distal downstream, and intergenic, which is the same as previous studies (Faulkner et al. 2009; Su et al. 2014). Proximal upstream was defined as (−10 kb, 100 bp) in relation to the 5′ end of a gene, while proximal downstream was (−100 bp, 10 kb) to the 3′ end of the gene. Distal upstream and downstream were defined as (−100 kb, 10 kb) to a gene's TSS and TTS, respectively. Intergenic refers to any repetitive element located more than 100 kb from the nearest gene. Considering that a repeat could be classified as exon as well as proximal upstream, the order of priority was defined as: 5′ UTR, exon, intron, 3′ UTR, proximal upstream, proximal downstream, distal upstream, distal downstream, and intergenic; a repeat was only assigned to one genomic region.

Data sources and data processing of ENCODE ChIP-seq and RNA-seq data

ENCODE (Boyle et al. 2014), Cistrome (Liu et al. 2011), mouse ENCODE, NIH Roadmap Epigenomics, Blueprint, and CEEHRC data were downloaded from respective websites and processed using standard tools. Due to the mappability issue for assigning short reads to TEs (Derrien et al. 2012), we generated both uniquely mapped and random hit mapped reads using the same strategy implemented in a previous study (Su et al. 2014). Briefly, for ChIP-seq data, replicates were merged first and then mapped to the human genome (hg38) or mouse genome (mm10) by Bowtie (v1.1.0) (Langmead et al. 2009) allowing up to two mismatches. Both unique map (-k 1 –m 1) and random hit map (-k 1 –m 100) were generated for comparison. After mapping, redundant reads were removed. More details are described in the Supplemental Materials.

Mappability correction and normalization of ChIP-seq data

While mappability correction was not done for preprocessed NIH Roadmap Epigenomics "consolidated" and "unconsolidated" data, or random hit mapped ENCODE and mouse ENCODE data, for the uniquely mapped ENCODE and mouse ENCODE ChIP-seq data, the following methods of mappability correction, normalization against input, and quantification were used. We provide a graph to illustrate some of the variables used for the following (Supplemental Fig. S1A).

Mapped reads were first extended to the fragment length *fl* in the 3′ direction, as each read actually represents a sequencing fragment. The fragment length can be obtained experimentally or computationally; however, to make different types of histone modification comparable, we used $fl = 150$, which is also the default parameter of the "iteres" package (Xie et al. 2013).

The mappability is defined as $m_p = 1/f_k$, p denotes the genomic location of the nucleotide, and f_k is the number of locations where the k-mer started from p could map to the genome using defined a Bowtie parameter. The reads length of most of ENCODE and mouse ENCODE ChIP-seq data selected is 36 bp, so $k = 36$ was used. The whole genome mappability score was generated by gem-mappability (Derrien et al. 2012).

We first performed the correction of signal loses due to low mappability using the following formula both for input and ChIP data:

$$\hat{d}_i = \begin{cases} \min\left(\frac{d_i}{m_i},\ \max(d_i, d_{i,j})\right), & m_i < 1 \\ d_i, & m_i = 1 \end{cases} \quad | \ d_i \rangle 0 \tag{1}$$

$$d_{i,j} = \text{median}\left(\{d_j\}_{j \in [p-w, p+w]}\right).$$

For the ith genomic region with the same read counts, we denote the read counts as d_i; m_i is the mean mappability for the region. If $d_i > 0$ & $m_i < 1$, then the correction was carried out. For the ith genomic region, $d_{i,j}$ is the median count of its nearby upstream and downstream w bins with the same length as the ith region. In analysis, $w = 5$ was used.

We then normalize the corrected ChIP signal to input as

$$\tilde{d}_i = \max(0, \hat{d}_i^{chip} - R\hat{d}_i^{input}), \tag{2}$$

where R is the normalization factor for chip vs. input. For different histone modifications with the same input, R is different. Here, minus was used rather than division because there are millions of repeats in the genome; otherwise, TEs would need +1 to avoid division by zero, which would bias the total counts. We estimated the normalization factor R in a similar way to a previous study (Liang and Keles 2012), except setting the searching window vector as [5000, 10,000, 20,000, 50,000, 100,000].

We defined the FPM (fragments per million) to quantify each histone modification on individual TEs. FPM is similar to TPM (Wagner et al. 2012) in mathematical nature and can be seen as the coverage ratio of the total coverage.

$$\text{FPM}_i = \frac{10^6 \times n_i \times fl}{L_i \times F}, \tag{3}$$

$$\text{F} = \sum_i \frac{n_i \times fl}{L_i},$$

where n_i is the read counts for the ith repeat, *fl* is the fragment length defined above, and L_i is the length for the ith repeat.

The length of nucleotides covered by mapped reads can be represented in two ways:

$$\sum_j d_{i,j} \cdot l_{i,j} = n_i \times fl, \tag{4}$$

where n_i is the read counts for the ith repeat, *fl* is the fragment length, and in the repeat, there are several regions in which each region has the equal read counts $d_{i,j}$, and the region's length is $l_{i,j}$.

As we obtained the mappability-corrected and input-normalized signal $\tilde{d}_{i,j}$ from Equation (2), then Equation (3) could be converted by Equation (4) to the following as the final estimation of the signal for the individual TE:

$$\text{FPM}_i = \frac{10^6 \times \sum_j \tilde{d}_{i,j} \cdot l_{i,j}}{L_i \times F}, \tag{5}$$

$$\text{F} = \sum_i \frac{\sum_j \tilde{d}_{i,j} \cdot l_{i,j}}{L_i}.$$

Forest of forest model detecting regulatory transposable elements

The golden standard positive (GSP) set for enhancer-like repeats was defined as distal or intergenic repeats bound by EP300 and overlapped with DHSs, with H3K27ac levels higher than that of the background sets. In the collected ENCODE data, EP300 binding sites were only available in GM12878, H1-hESC, HeLa-S3, HepG2, and K562. The GSP set for promoter-like repeats were defined as proximal upstream or 5′ UTR repeats with significant H3K4me3 peaks, overlapping with DHSs and >1 TF binding site. The golden negative sets, termed as background repeats (BRs), were defined as DHS-overlapped repeats, without any TF binding or any significant histone modification peaks. For ELRs, we used the GSP PLRs and BRs as negative controls to train the model. For PLRs, we used the GSP ELRs and BRs as negative controls. Random forest (Liaw and Wiener 2002) implemented in scikit-learn (v0.16.1) (Pedregosa et al. 2011), named ExtraTreesClassifier (Geurts et al. 2006), was used for classification. Classification was done in two steps based on random forest, which we termed the forest of forest model. In the first step, we trained a random forest model for each of the five cell lines that have EP300 data using a curated positive and negative set for the cell line. Models trained in one cell line showed good performance in other cells, so the models trained in the five cell lines can be used in other cell lines which did not have EP300 data and we can integrate the output from the five models to achieve a higher performance. In the second step, for the training set in each of the five cell lines, we first obtained the possibility matrix using the five models trained in the first step. Then, using the possibility matrix, we trained a random forest again and reported the final binary value. Random forest only contains one parameter, which controls the number of decision trees used, and was decided by iterations that reached steady performance. AUC and ZOL were calculated by the functions of auc and zero_one_loss from the sklearn (v0.16.1) (Pedregosa et al. 2011) to evaluate model performance. The FOFM trained in ENCODE data was used to classify ELRs and PLRs using NIH Roadmap Epigenomics and mouse ENCODE data.

Transposable element family and subfamily enrichment analysis

Hypergeometric test and binomial test *P*-values were combined by Stouffer's Z-score method (Stouffer 1949; Darlington and Hayes 2000), and the FDR was calculated by the combined *P*-values. The hypergeometric test, binomial test, and FDR functions implemented in the Orange Bioinformatics Toolbox (Demšar et al. 2013) were used, and the Stouffer's *Z*-score method implemented in

Scipy (https://www.scipy.org/) was used to combine the two *P*-values. FDR < 1×10^{-100} was assigned 1×10^{-100}, and FDR < 1×10^{-10} was defined as significant for TE families and FDR < 1×10^{-20} for TE subfamilies. Fold enrichment over background was calculated as

$$\mathrm{FE} = \frac{k}{m} / \frac{n}{N},$$

where k is the number of specific TE family/subfamily ELRs/RLRs in the input list, m is the number of TEs in the input list, n is the number of the specific families in the genome, and N is the total number of TEs in the genome.

Saturation estimation for ELRs, PLRs, and active TEs

The saturation estimation was performed by iteratively adding the number of unique set of ELRs or PLRs by samples compared to the total number of TEs in the genome. Active TEs were defined as within 1 kb upstream of or downstream from the EP300 peak summits. All TEs were assessed using FPM described above against respective input data. Here, no correction for mappability for the TF ChIP-seq data was done according to the observation made by a previous study (Harmanci et al. 2014).

Neighbor joining tree analysis

The neighbor joining trees (Saitou and Nei 1987) were built using MEGA (v6.0) (Tamura et al. 2013). Euclidean distances were precomputed. For the NJT built by ELRs, we used the binary matrix whose columns are different samples and rows are TEs and the value is set to 1 if a repeat is classified as an ELR and otherwise 0, to calculate the distance. Cluster purity was calculated as 1.0 – [misaligned dendrites/total dendrites], and misaligned was counted as the samples not aligned to the branch where the majority of samples from the same group were.

Motif analysis

FIMO (v4.10.0) (Grant et al. 2011) was used to identify motifs' location in the genome using default parameters. Motif density was calculated relative to TE centers and normalized by the segment size. CentriMo (v4.10.0) (Bailey and Machanick 2012) was used to identify the motifs showing significant preference at the center of a set of TE sequences. The HOCOMOCOv10_HUMAN_mono motif data sets (Kulakovskiy et al. 2013) curated by MEME (Bailey et al. 2006) were used in both FIMO and CentriMo.

Fold enrichment of TF binding over background on tsELRs

Fold enrichment of TF binding peaks on tsELRs over background was calculated as

$$\mathrm{FE} = \frac{k}{m} / \frac{n}{N},$$

where k is the number of tsELRs overlapped with peaks for a specific TF, m is the number of TEs overlapped with peaks for the TF, n is the number of tsELRs, and N is the total number of TEs.

TEs mapped to other species

We selected representative vertebrates on the topology structure of the phylogenetic tree from UCSC, namely human, chimp, gorilla, rhesus, macaque, marmoset, squirrel monkey, mouse, rat, pig, cow, sheep, cat, dog, opossum, chicken, frog (*Xenopus tropicalis*), and zebrafish. We first mapped human TE sequences to other species by bnMapper (Denas et al. 2015) with key parameters –f BED4 –gap 20 –threshold 0.1, the same as those used in the conservation analysis of DHSs between human and mouse (Vierstra et al. 2014). For the mapped TEs, the nonredundant TEs that could overlap with TEs in the target species were kept, which result in the same family TEs or highly similar families.

Association of TE pairs between ChIA-PET interaction anchors

The processed interaction anchors of ChIA-PET data were from ENCODE or GEO as annotated. For ENCODE data sets, replicates were merged. For inter-anchors association analysis, we calculated support and confidence for all possible pairs of TE families in pairs of interaction anchors using the same formula of support and confidence in the Apriori algorithm (Agrawal and Srikant 1994) and requiring support >0.2 and confidence >0.5.

Estimation of paired-end tags (PETs) level density of interacting TEs

Paired-end tags (PETs) densities of interacting TEs in dense in situ Hi-C data (Rao et al. 2014) for HMEC, K562, and GM12878 cells were estimated using the method described in a previous study (Su et al. 2014). The raw data were preprocessed by HiCUP (Wingett et al. 2015) (v0.5.4) to obtain intra-chromosomal PETs. The ChIA-PET data were preprocessed by Mango (Phanstiel et al. 2015) to obtain intra-chromosomal PETs; other analyses are the same as Hi-C data analysis. Only PETs with distance >10 kb were used to avoid self-ligation PETs. If PETs have one end located in a TE, then the distances of the other end to the nearest TSS or TE's center were recorded. The distances were grouped into 100 bins, and the PETs in each bin were counted. The binned PET counts were then normalized for the same source TE by dividing the copy number of target TE in the genome.

Plasmids and luciferase reporter assay

The sequences of *Alu*, L2, MIR, and scrambled MIR were synthesized as a single or L2 and MIR combined element, and cloned into the firefly luciferase reporter vector pGL4.23[luc2/minP] (Promega) between the KpnI and HindIII sites by Shanghai Majorbio. All clones were validated by sequencing. HeLa cells were cotransfected with the firefly luciferase reporter construct and the internal control pRL-TK *Renilla* luciferase vector (Promega) at a ratio of 40:1 (reporter vector:control vector) using Lipofectamine 3000 (Life Technologies) in triplicate. Forty-eight hours after transfection, the cells were lysed and firefly and *Renilla* luciferase activities were measured using the Dual-Luciferase Reporter Assay System (Promega) and Synergy H1 Microplate Reader (BioTek) according to the manufacturer's protocols.

Data access

Essential codes to reproduce our result and training sets are in the Supplemental Material of Supplemental_Code_Data.tar.gz. All data and code are available at: http://www.picb.ac.cn/hanlab/cisTEs.

Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China (91749205, 91329302, and 31210103916) to J.-D.J.H., and 31371188 to G.W.; China Ministry of Science and Technology (2015CB964803 and 2016YFE0108700) and Chinese Academy of Sciences (CAS) (XDA01010303 and YZ201243) and Max Planck fellowship to J.-D.J.H. G.W. acknowledges support from the Youth Innovation Promotion Association of the CAS, and the SA-SIBS Scholarship

Program. This study makes use of data generated by the Blueprint Consortium.

Author contributions: Y.C. and J.-D.J.H. designed the analysis; Y.C. and G.C. performed the analyses; Y.C., G.W., J.M., and J.-D.J.H. designed the experiments; G.W. and X.Z. carried out experiments. All authors contributed to data interpretation and wrote the paper.

References

Adams D, Altucci L, Antonarakis SE, Ballesteros J, Beck S, Bird A, Bock C, Boehm B, Campo E, Caricasole A. 2012. BLUEPRINT to decode the epigenetic signature written in blood. *Nat Biotechnol* **30:** 224–226. doi:10.1038/nbt.2153

Agrawal R, Srikant R. 1994. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th international conference on very large data bases*, pp. 487–499. Morgan Kaufmann Publishers, Inc., Burlington, MA.

Amin V, Harris RA, Onuchic V, Jackson AR, Charnecki T, Paithankar S, Lakshmi Subramanian S, Riehle K, Coarfa C, Milosavljevic A. 2015. Epigenomic footprints across 111 reference epigenomes reveal tissue-specific epigenetic regulation of lincRNAs. *Nat Commun* **6:** 6370. doi:10.1038/ncomms7370

Arcot SS, Wang Z, Weber JL, Deininger PL, Batzer MA. 1995. *Alu* repeats: a source for the genesis of primate microsatellites. *Genomics* **29:** 136–144. doi:10.1006/geno.1995.1224

Arner E, Daub CO, Vitting-Seerup K, Andersson R, Lilje B, Drabløs F, Lennartsson A, Rönnerblad M, Hrydziuszko O, Vitezic M, et al. 2015. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* **347:** 1010–1014. doi:10.1126/science.1259418

Bae JB. 2013. Perspectives of international human epigenome consortium. *Genomics Inform* **11:** 7–14. doi:10.5808/GI.2013.11.1.7

Bailey TL, Machanick P. 2012. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res* **40:** e128. doi:10.1093/nar/gks433

Bailey TL, Williams N, Misleh C, Li WW. 2006. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res* **34:** W369–W373. doi:10.1093/nar/gkl198

Biémont C, Vieira C. 2006. Genetics: junk DNA as an evolutionary force. *Nature* **443:** 521–524. doi:10.1038/443521a

Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG, et al. 2005. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122:** 947–956. doi:10.1016/j.cell.2005.08.020

Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, Cheng Y, Gardner K, Hillier LW, Janette J, Jiang L, et al. 2014. Comparative analysis of regulatory information and circuits across distant species. *Nature* **512:** 453–456. doi:10.1038/nature13668

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25:** 1915–1927. doi:10.1101/gad.17446611

Chénais B, Caruso A, Hiard S, Casse N. 2012. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. *Gene* **509:** 7–15. doi:10.1016/j.gene.2012.07.042

Cheung MS, Down TA, Latorre I, Ahringer J. 2011. Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic Acids Res* **39:** e103. doi:10.1093/nar/gkr425

Chung D, Kuan PF, Li B, Sanalkumar R, Liang K, Bresnick EH, Dewey C, Keleif in text S. 2011. Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-seq data. *PLoS Comput Biol* **7:** e1002111. doi:10.1371/journal.pcbi. 1002111

Chuong EB, Elde NC, Feschotte C. 2016. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet* **18:** 71–86. doi:10.1038/nrg.2016.139

Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10:** 691–703. doi:10.1038/nrg2640

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci* **107:** 21931–21936. doi:10.1073/pnas.1016071107

Darlington RB, Hayes AF. 2000. Combining independent *p* values: extensions of the Stouffer and binomial methods. *Psychol Methods* **5:** 496. doi:10.1037/1082-989X.5.4.496

Day DS, Luquette LJ, Park PJ, Kharchenko PV. 2010. Estimating enrichment of repetitive elements from high-throughput sequence data. *Genome Biol* **11:** R69. doi:10.1186/gb-2010-11-6-r69

de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet* **7:** e1002384. doi:10.1371/journal.pgen.1002384

Demšar J, Curk T, Erjavec A, Gorup Č, Hočevar T, Milutinovič M, Možina M, Polajnar M, Toplak M, Starič A. 2013. Orange: data mining toolbox in Python. *J Mach Learn Res* **14:** 2349–2353.

Denas O, Sandstrom R, Cheng Y, Beal K, Herrero J, Hardison RC, Taylor J. 2015. Genome-wide comparative analysis reveals human-mouse regulatory landscape and evolution. *BMC Genomics* **16:** 87. doi:10.1186/s12864-015-1245-6

Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, Guigó R, Ribeca P. 2012. Fast computation and applications of genome mappability. *PLoS One* **7:** e30377. doi:10.1371/journal.pone.0030377

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74. doi:10.1038/nature11247

Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9:** 215–216. doi:10.1038/nmeth.1906

Erwin JA, Marchetto MC, Gage FH. 2014. Mobile DNA elements in the generation of diversity and complexity in the brain. *Nat Rev Neurosci* **15:** 497–506. doi:10.1038/nrn3730

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, et al. 2009. The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* **41:** 563–571. doi:10.1038/ng.368

Geurts P, Ernst D, Wehenkel L. 2006. Extremely randomized trees. *Mach Learn* **63:** 3–42. doi:10.1007/s10994-006-6226-1

Goke J, Ng HH. 2016. CTRL+INSERT: retrotransposons and their contribution to regulation and innovation of the transcriptome. *EMBO Rep* **17:** 1131–1144. doi:10.15252/embr.201642743

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27:** 1017–1018. doi:10.1093/bioinformatics/btr064

Harmanci A, Rozowsky J, Gerstein M. 2014. MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework. *Genome Biol* **15:** 474. doi:10.1186/s13059-014-0474-3

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22:** 1760–1774. doi:10.1101/gr.135350.111

Hedges SB, Marin J, Suleski M, Paymer M, Kumar S. 2015. Tree of life reveals clock-like speciation and diversification. *Mol Biol Evol* **32:** 835–845. doi:10.1093/molbev/msv037

Heidari N, Phanstiel DH, He C, Grubert F, Jahanbanian F, Kasowski M, Zhang MQ, Snyder MP. 2014. Genome-wide map of regulatory interactions in the human genome. *Genome Res* **24:** 1905–1917. doi:10.1101/gr.176586.114

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39:** 311–318. doi:10.1038/ng1966

Huda A, Tyagi E, Mariño-Ramírez L, Bowen NJ, Jjingo D, Jordan IK. 2011. Prediction of transposable element derived enhancers using chromatin modification profiles. *PLoS One* **6:** e27513. doi:10.1371/journal.pone.0027513

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921. doi:10.1038/35057062

Jacques PE, Jeyakani J, Bourque G. 2013. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genet* **9:** e1003504. doi:10.1371/journal.pgen.1003504

Jjingo D, Conley AB, Wang J, Marino-Ramirez L, Lunyak VV, Jordan IK. 2014. Mammalian-wide interspersed repeat (MIR)-derived enhancers and the regulation of human gene expression. *Mob DNA* **5:** 14. doi:10.1186/1759-8753-5-14

Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C. 2013. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet* **9:** e1003470. doi:10.1371/journal.pgen.1003470

Karmodiya K, Krebs AR, Oulad-Abdelghani M, Kimura H, Tora L. 2012. H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. *BMC Genomics* **13:** 424. doi:10.1186/1471-2164-13-424

Karolchik D, Hinrichs AS, Kent WJ. 2012. The UCSC Genome Browser. *Curr Protoc Bioinformatics* **Chapter 1:** Unit1.4. doi:10.1002/0471250953.bi0104s40

Kulakovskiy IV, Medvedeva YA, Schaefer U, Kasianov AS, Vorontsov IE, Bajic VB, Makeev VJ. 2013. HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res* **41:** D195–D202. doi:10.1093/nar/gks1089

Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, Bourque G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* **42:** 631–634. doi:10.1038/ng.600

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25. doi:10.1186/gb-2009-10-3-r25

Lee H, Schatz MC. 2012. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics* **28:** 2097–2105. doi:10.1093/bioinformatics/bts330

Liang K, Keles S. 2012. Normalization of ChIP-seq data with control. *BMC Bioinformatics* **13:** 199. doi:10.1186/1471-2105-13-199

Liaw A, Wiener M. 2002. Classification and regression by randomForest. *R News* **2/3:** 18–22.

Liu T, Ortiz JA, Taing L, Meyer CA, Lee B, Zhang Y, Shin H, Wong SS, Ma J, Lei Y. 2011. Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol* **12:** R83. doi:10.1186/gb-2011-12-8-r83

Loh YH, Wu Q, Chew JL, Vega VB, Zhang W, Chen X, Bourque G, George J, Leong B, Liu J, et al. 2006. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet* **38:** 431–440. doi:10.1038/ng1760

Mills RE, Bennett EA, Iskow RC, Devine SE. 2007. Which transposable elements are active in the human genome? *Trends Genet* **23:** 183–191. doi:10.1016/j.tig.2007.02.006

Mousavi K, Zare H, Dell'Orso S, Grontved L, Gutierrez-Cruz G, Derfoul A, Hager Gordon L, Sartorelli V. 2013. eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci. *Mol Cell* **51:** 606–617. doi:10.1016/j.molcel.2013.07.022

Mouse Encode Consortium, Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, Groudine M, Bender M, Kaul R, et al. 2012. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol* **13:** 418. doi:10.1186/gb-2012-13-8-418

Muerdter F, Boryń ŁM, Woodfin AR, Neumayr C, Rath M, Zabidi MA, Pagani M, Haberle V, Kazmar T, Catarino RR, et al. 2017. Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat Methods* **15:** 141. doi:10.1038/nmeth.4534

Papaemmanuil E, Gerstung M, Bullinger L, Gaidzik VI, Paschka P, Roberts ND, Potter NE, Heuser M, Thol F, Bolli N, et al. 2016. Genomic classification and prognosis in acute myeloid leukemia. *N Engl J Med* **374:** 2209–2221. doi:10.1056/NEJMoa1516192

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. 2011. Scikit-learn: machine learning in Python. *J Mach Learn Res* **12:** 2825–2830.

Phanstiel DH, Boyle AP, Heidari N, Snyder MP. 2015. Mango: a bias-correcting ChIA-PET analysis pipeline. *Bioinformatics* **31:** 3092–3098. doi:10.1093/bioinformatics/btv336

Rajagopal N, Xie W, Li Y, Wagner U, Wang W, Stamatoyannopoulos J, Ernst J, Kellis M, Ren B. 2013. RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS Comput Biol* **9:** e1002968. doi:10.1371/journal.pcbi.1002968

Rao Suhas SP, Huntley Miriam H, Durand Neva C, Stamenova Elena K, Bochkov Ivan D, Robinson James T, Sanborn Adrian L, Machol I, Omer Arina D, Lander Eric S, et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159:** 1665–1680. doi:10.1016/j.cell.2014.11.021

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518:** 317–330. doi:10.1038/nature14248

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4:** 406–425. doi:10.1093/oxfordjournals.molbev.a040454

Seberg O, Petersen G. 2009. A unified classification system for eukaryotic transposable elements should reflect their phylogeny. *Nat Rev Genet* **10:** 276. doi:10.1038/nrg2165-c3

Silva J, Shabalina S, Harris D, Spouge J, Kondrashov A. 2003. Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes. *Genet Res* **82:** 1–18. doi:10.1017/S0016672303006268

Smit AFA, Hubley R, Green P. 2013–2015. RepeatMasker Open-4.0. http://www.repeatmasker.org.

Smith AM, Sanchez MJ, Follows GA, Kinston S, Donaldson IJ, Green AR, Gottgens B. 2008. A novel mode of enhancer evolution: The *Tal1* stem cell enhancer recruited a MIR element to specifically boost its activity. *Genome Res* **18:** 1422–1432. doi:10.1101/gr.077008.108

Stouffer SA. 1949. *Studies in social psychology in World War II: the American soldier: adjustment during army life*. Princeton University Press, Princeton, NJ.

Su M, Han D, Boyd-Kirkup J, Yu X, Han JD. 2014. Evolution of *Alu* elements toward enhancers. *Cell Rep* **7:** 376–385. doi:10.1016/j.celrep.2014.03.011

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. doi:10.1101/gr.168872.113

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP. 2014. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* **43:** D447–D452. doi:10.1093/nar/gku1003

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* **30:** 2725–2729. doi:10.1093/molbev/mst197

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. doi:10.1101/gr.218149.116

Tsankov AM, Gu H, Akopian V, Ziller MJ, Donaghey J, Amit I, Gnirke A, Meissner A. 2015. Transcription factor binding dynamics during human ES cell differentiation. *Nature* **518:** 344–349. doi:10.1038/nature14233

Vierstra J, Rynes E, Sandstrom R, Zhang M, Canfield T, Hansen RS, Stehling-Sun S, Sabo PJ, Byron R, Humbert R, et al. 2014. Mouse regulatory DNA landscapes reveal global principles of *cis*-regulatory evolution. *Science* **346:** 1007–1012. doi:10.1126/science.1246426

Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT, Jasinska AJ, et al. 2015. Enhancer evolution across 20 mammalian species. *Cell* **160:** 554–566. doi:10.1016/j.cell.2015.01.006

Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457:** 854. doi:10.1038/nature07730

Wagner GP, Kin K, Lynch VJ. 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci* **131:** 281–285. doi:10.1007/s12064-012-0162-3

Wang Z, Zang C, Cui K, Schones DE, Barski A, Peng W, Zhao K. 2009. Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell* **138:** 1019–1031. doi:10.1016/j.cell.2009.06.049

Wang J, Huda A, Lunyak VV, Jordan IK. 2010. A Gibbs sampling strategy applied to the mapping of ambiguous short-sequence tags. *Bioinformatics* **26:** 2501–2508. doi:10.1093/bioinformatics/btq460

Wang J, Xie G, Singh M, Ghanbarian AT, Rasko T, Szvetnik A, Cai H, Besser D, Prigione A, Fuchs NV, et al. 2014. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature* **516:** 405–409. doi:10.1038/nature13804

Wingett S, Ewels P, Furlan-Magaril M, Nagano T, Schoenfelder S, Fraser P, Andrews S. 2015. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res* **4:** 1310. doi:10.12688/f1000research.7334.1

Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, Zhou X, Lee HJ, Maire CL, Ligon KL, et al. 2013. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nat Genet* **45:** 836–841. doi:10.1038/ng.2649

Received February 6, 2018; accepted in revised form November 12, 2018.