Supplementary Materials of

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

Yaqiang Cao 1 *, Guoyu Chen 1 *, Gang Wu 1*, Xiaoli Zhang 1*, 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, 320 Yue Yang Road, Shanghai, 200031, China.

2 School of Life Science and Technology, ShanghaiTech University, Shanghai, 201210, China

* These authors contributed equally to this work

Corresponding Author: Dr. Jing-Dong J. Han, CAS-MPG Partner Institute for Computational Biology, Shanghai, 200031, China, (Phone): +86-21-54920458, (Fax): +86-21-54920451, E-mail: jdhan@picb.ac.cn.

Key words: transposable elements, enhancers, promoters, MIR, L2, primate specific, master transcription factors, cell identities
Supplementary Methods

Processing of ENCODE ChIP-seq and RNA-seq data

Transcription factor binding sites and DNase I hypersensitive sites

Processing of mouse ENCODE ChIP-seq and RNA-seq data

Processing of NIH Roadmap Epigenomics ChIP-seq and RNA-seq data

Processing of Blueprint and CEEHRC data

Mean ChIP-seq profiles, enriched GO terms analysis and genome browser visualization

Enrichment of tsELRs to nearby tissue specific genes

Sequences of MIR and L2 for enhancer luciferase reporter assay

Supplementary Figures

Supplementary Figure S1 – Supplementary Figure S16
METHODS

Processing of ENCODE ChIP-seq and RNA-seq data

We analyzed selected data from ENCODE (Boyle et al. 2014) in following cell lines: Dnd41, GM12878, H1-hESC, HMEC, HSMM, HSMMtube, HUVEC, HeLa-S3, HepG2, K562, NH-A, NHDF-Ad, Monocytes-CD14+_RO01746, NHEK, NHLF. The reason we chose these cell lines is there are rich and comparable histone modification ChIP-seq data, corresponding polyA and non-polyA RNA-seq data and transcription factor binding data. Raw histone modification ChIP-seq data of H3K27me3, H3K9me3, H4K20me1, H2A.Z, H3K27ac, H3K36me3, H3K4me1, H3K4me2, H3K4me3 and H3K9ac were downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/ . Replicates were merged first and then mapped to the genome (hg38) by Bowtie (Langmead et al. 2009) (v1.1.0) allowing up to 2 mismatches. Both unique map (-k 1 –m 1) and random hit one map (-k 1 –m 100) were generated for comparison. After mapping, redundant reads were removed. Peak calling was carried out by MACS2 (Zhang et al. 2008) (v2.1.0.20140616) using default parameters except –shift was set to 73.

RNA-seq data were obtained from GSE30567 (Djebali et al. 2012). STAR (Dobin et al. 2013) (v2.4.0d) was used to map the reads to the genome, also both in unique way and multiple hit (–outFilterMultimapNmax 1000, ). The quantitative result of each individual repeat was calculated by iteres (Xie et al. 2013).

Transcription factor binding sites and DNase I hypersensitive sites

Total processed 508 bed files for 134 transcription factors (TF) were retrieved from http://encodedcc.stanford.edu/modENCODE_VS_ENCODE/Regulation/Human/peakCalls/uniformPk/ on 2015-01-24. We also retrieved data (550 bed files, 237 TFs) from ReMap (Griffon et al. 2015) on 2015-07-24. Processed DNase I hypersensitive sites (DHS) were downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUw
For hESC (include iPSC) and HSC specific TFs, the peaks were downloaded from Cistrome (Liu et al. 2011) on 2016-03-21 and 2016-03-30. Replicates were merged into union sets; peaks for same TF but from different studies were also merged into union sets. The coordinates of the TFBSs and DHSs were converted from original genome version to hg38 using liftOver with the default parameters. The TEs overlapped with these TFBSs and DHSs were obtained by an in-house script.

Processing of mouse ENCODE ChIP-seq and RNA-seq data

Raw data of histone modification ChIP-seq data for CH12 and MEL were downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeLicrHistone/. And the corresponding non-polyA RNA-seq data were obtained from http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeSydhRnaSeq/ and polyA RNA-seq data from http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodePsuRnaSeq/. Processed TFBSs were downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeSydHTfbs/ and DHS data were downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeUwDnase/. The coordinates of the TFBS and DHS were converted from original genome version mm9 to mm10 using liftOver. Raw ChIP-seq and RNA-seq data were processed in the same way as human data.

Processing of NIH Roadmap Epigenomics ChIP-seq and RNA-seq data

NIH Roadmap Epigenomics consolidated data was obtained from http://egg2.wustl.edu/roadmap/data/byFileType/alignments/consolidated/, and non-consolidated data was obtained from http://egg2.wustl.edu/roadmap/data/byFileType/alignments/consolidated_nosubsampling/.

Processing of Blueprint and CEEHRC data
Processed peaks of BLUEPRINT 2015 release histone modification data were obtained from ftp://ftp.ebi.ac.uk/pub/databases/blueprint/ at 2016 July. Processed peaks of CEEHRC were obtained from ftp://epigenomesportal.ca/public_data/donor/ at 2016 July. With H3K4me1 peaks and without neither H3K27me3 nor H3K4me3 peaks were required to define TE enhancers.

**Mean ChIP-seq profiles, enriched GO terms analysis and genome browser visualization**

The command annotatePeak.pl in HOMER (Heinz et al. 2010) was used to generate the normalized matrix of ChIP-seq data for repeats, with following options: -size 10000 –hist 200 –ghist. The smoothed average signal for the matrix was used to plot the average profile and the matrix was visualized by Treeview (Eisen et al. 1998) to show heatmaps. The findGO.pl implemented in HOMER (Heinz et al. 2010) was used to test the enriched GO terms. Terms from biological process, p-value < 1e-5 and at least each 5 genes involved in a term were required for enriched terms. If a GO enrichment test outputs too many results, we showed the top 20s. For the visualization of ChIP-seq data in genome browser, we used makeMultiWigHub.pl in HOMER to generate the individual tracks and the overlay ones.

**Enrichment of tsELRs to nearby tissue specific genes**

The expression matrix for 57 samples was obtained from NIH Roadmap Epigenomics (57epigenomes.RPKM.pc.txt, http://egg2.wustl.edu/roadmap/data/byDataType/rna/expression/), samples not in our groups defined above were removed, and we only kept tissues with more than one sample. Tissue specific genes (TSGs) were obtained by an entropy-based method. Briefly, Jensen-Shannon Divergence (JSD) was used to measure the specificity of genes expression pattern and pre-defined tissue specific patterns, similar method to (Cabili et al. 2011). Pre-defined tissue specific pattern was a binary vector as samples in target tissue are marked 1 and others as 0. Finally, a JSD cutoff > 0.6 was used to define the tissue specific genes. There were overlapped genes for different tissues.
The distance enrichment of tsELRs to TSGs’ TSSs was evaluated by the similar methods to (Chuong et al. 2016). Briefly, the distances were grouped by 10 kb bin sizes. The expected background was drawn by the mean values of sampling (10,000 times) equal number of neither enhancer nor promoter TEs.

**Sequences of MIR and L2 for enhancer luciferase reporter assay**

> common_L2_1; chr4|38160986|38161342|+|L2c|Distal.downstream
tgtacaagctcttctcctttgcctaacatgaccccccttcagctgccctgggaagacctttactgtattccagccc
cattgatcctcttcggaacactgtcagtatggtaacaaagcttaactctctgctgtgtcccat
ggcaacttgtaacaatgtcagatactcatcatctcctcgtgttataataataaatctatccatcaattcatatg
gtactagatcatagggtcaggagggacggactgtgtgtccatccagctataaattgttatctccagcacc
cacagaaggattgatacagagcatacccaatctcaatgaatgataatgataaaaa

>common_L2_2; chr2|26012264|26012454|+|L2a|Distal.upstream
cccaggttgagttcaatcttatctattgatattttttttttttctttttatgtcattccccctcatcagcgt
cgacctctgtgagataagagctcaaggaaggctgctgacctatggcgtaaatgg
cactcaataaatattgtttgaactgatgtgatgaatgaaatatatatatattgtttgaatgaat

>common_L2_3; chr1|248828933|248829039|+|L2|Proximal.upstream
cctccacttcctcctccccattcaccctctactagctgctgctgctgcaattcataggaattcaggttcatgctctatgcccacatc

>common_L2_4;chr17|59853385|59853484|+|L2c|Distal.upstream
tggcctcctagtgccgagggccaggtctactattccctccctctactattccctctctgttcagttcatttttgtttgaat

>common_L2_5; chr9|100429406|100429575|-|L2b|intron
tttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttt
>common_MIR_1; chr6|47135774|47135901|-|MIRb|Distal.upstream
aatatttatcaaatatcaccctctgtgcaagtgatattttacaaacatttatattaactcccccaacccctactattt
acagatgaggaactttggagatgctagctatttgcttagttac

>common_MIR_2; chr20|38094447|38094544|+|MIR3|Proximal.downstream
tcactatgaccttggaagctcaactcagccttgggaacctcagtgttttttcacaggaatgataggg
tgattacgtaacccctgaagtcc

>common_MIR_3; chr8|40159095|40159261|+|MIR3|Proximal.upstream
agcatcgtgtgagcaaaacaccgacctggaattccttggggacctcagctgcttcttacaggaaatgagtaggg
ttgattacgtaacccctgaagtcc

>common_MIR_4; chr8|10799946|10800070|-|MIRb|intron
taaggctcagaaagttatcctgccaaaggtcatatgactcacaagggcagagccgagactgcgatc
caggtctgcgtgacagctatcccggactctttctactacaccacaccattctt

>common_MIR_5; chr10|72320812|72320939|+|MIRC|Distal.downstream
tgagtcactggagtaggctacttccctctctcagtctccacatctgtgcaatagggataagggctggctgctct
ctagtctctggtgacagggatagtcagataacggttggaaggccgc

>HeLa_L2_1; chr5|39459240|39459424|+|L2b|Distal.upstream
atattgccaacctagttagctttctttctgtgcaacaacactatcaacccatcattccatcatcacaagcatacatc
atcactgcatagttctttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttt
ga\text{t\texttt{ttggatagcctgcttcaaaactcagttatcaca\texttt{aaaactttgcatcttcctccatggtgcac4ccttctc}}
\text{cttcccagtaaaacctccagtgcatctctggactcttctctctctctgtctatctgagttgcaatctttctctc}
\text{catggtcaatcttctgagaagtcttgataggtgtctccactttcttgcctcctattcactctttcaaccttc}
\text{tctggcagtcagtcatttcatagataatatctcacgtcatcctctataatgcagatgtttgag}
\text{gcaatgctattttatctatgttacagaggaaggaggttcagagagctttacgtaacccttcaaaatctacgc}
\text{ccaaataagacaataaagatttgagaagttttgaattctctggtgactcaggtccagggtcattgcttcccac}
\text{ctct}
\text{gagacacagatcagaggcaatctggtctgtatgaactcactttggtggaatcatttctctctc}
\text{ggtttcaggttcctcattgcaacagtaatgatgga}
\text{taacccctgaagggattttagttcaccctgacagatgtggaatttagagcccttagagagatttaaaat}
\text{agaaccctctagactaagtgacttaattcaggtctctgtatttcaagccatatctttctcttgagttgcattgc}
\text{cctt}
\text{acaatatttatcagtaatccctgaatgtcagggcattagctcaagtcacactccatatcttacagttcc}
\text{acagcaaccttctgaggtaggctatctgaaccacatcacaaggttaaatcagtaagagcttaagtg}
\text{cctttccaaaggttccacagattgtccagctggccagataattattcaggcagtctagatccagagccttt}
\text{aatcactgtac}
\text{ga}\text{t\texttt{ttggatagcctgcttcaaaactcagttatcaca\texttt{aaaactttgcatcttcctccatggtgcac4ccttctc}}
\text{cttcccagtaaaacctccagtgcatctctggactcttctctctctctgtctatctgagttgcaatctttctctc}
\text{catggtcaatcttctgagaagtcttgataggtgtctccactttcttgcctcctattcactctttcaaccttc}
\text{tctggcagtcagtcatttcatagataatatctcacgtcatcctctataatgcagatgtttgag}
\text{gcaatgctattttatctatgttacagaggaaggaggttcagagagctttacgtaacccttcaaaatctacgc}
\text{ccaaataagacaataaagatttgagaagttttgaattctctggtgactcaggtccagggtcattgcttcccac}
\text{ctct}
\text{gagacacagatcagaggcaatctggtctgtatgaactcactttggtggaatcatttctctctc}
\text{ggtttcaggttcctcattgcaacagtaatgatgga}
\text{taacccctgaagggattttagttcaccctgacagatgtggaatttagagcccttagagagatttaaaat}
\text{agaaccctctagactaagtgacttaattcaggtctctgtatttcaagccatatctttctcttgagttgcattgc}
\text{cctt}
\text{acaatatttatcagtaatccctgaatgtcagggcattagctcaagtcacactccatatcttacagttcc}
\text{acagcaaccttctgaggtaggctatctgaaccacatcacaaggttaaatcagtaagagcttaagtg}
\text{cctttccaaaggttccacagattgtccagctggccagataattattcaggcagtctagatccagagccttt}
\text{aatcactgtac}
\text{ga}\text{t\texttt{ttggatagcctgcttcaaaactcagttatcaca\texttt{aaaactttgcatcttcctccatggtgcac4ccttctc}}
\text{cttcccagtaaaacctccagtgcatctctggactcttctctctctctgtctatctgagttgcaatctttctctc}
\text{catggtcaatcttctgagaagtcttgataggtgtctccactttcttgcctcctattcactctttcaaccttc}
\text{tctggcagtcagtcatttcatagataatatctcacgtcatcctctataatgcagatgtttgag}
\text{gcaatgctattttatctatgttacagaggaaggaggttcagagagctttacgtaacccttcaaaatctacgc}
\text{ccaaataagacaataaagatttgagaagttttgaattctctggtgactcaggtccagggtcattgcttcccac}
\text{ctct}
\text{gagacacagatcagaggcaatctggtctgtatgaactcactttggtggaatcatttctctctc}
\text{ggtttcaggttcctcattgcaacagtaatgatgga}
\text{taacccctgaagggattttagttcaccctgacagatgtggaatttagagcccttagagagatttaaaat}
\text{agaaccctctagactaagtgacttaattcaggtctctgtatttcaagccatatctttctcttgagttgcattgc}
\text{cctt}
\text{acaatatttatcagtaatccctgaatgtcagggcattagctcaagtcacactccatatcttacagttcc}
\text{acagcaaccttctgaggtaggctatctgaaccacatcacaaggttaaatcagtaagagcttaagtg}
\text{cctttccaaaggttccacagattgtccagctggccagataattattcaggcagtctagatccagagccttt}
\text{aatcactgtac}
agagaagtcagctgtgtaatgggaagagaatctcaaagtcatcagctcaggtgagctttcatctctttatctgtgaga
catggggagattactatgtgcttctcatgaccagttatcctcctagtggccctcattggcttcctttggtc
atgtaatcacattgacattgtaactgttggtcataatcggtctcactggaatttgaagttttgttcatagagggg
gatgtctttcatctctgtatccagctgacgccctgtgtgctatatagctataagtaat

>HepG2_L2_1;chr6|1268377|1268586|+|L2c|Distal.upstream
gctggtgttttaaagttgagctctttcctacgagcccttcacagagactctttctctggctttggttc
atgtaatcagctgtgtaatgggaagagaatctcaaagtcatcagctcaggtgagctttcatctctttatctgtgaga
catggggagattactatgtgcttctcatgaccagttatcctcctagtggccctcattggcttcctttggtc
atgtaatcacattgacattgtaactgttggtcataatcggtctcactggaatttgaagttttgttcatagagggg
gatgtctttcatctctgtatccagctgacgccctgtgtgctatatagctataagtaat

>HepG2_L2_2;chr16|10647811|10647937|-|L2a|intron
tcataatgtcagctgcacacaaatgtctggtcagacctacttagtgccaaagcactgtctagctgatccga
gtaaccacacacagctggaaaccctgtctctgtggagtggtgagctgacattctctgg

>HepG2_L2_3;chr11|111800621|111800845|+|L2b|intron
cccactgtcactctctgtgcaacacccctgtactctctcatcaacacataacacatggtaattaactcatgtgatt
tgtggtgatattgtaaggctagcctcaccattgtatcgaaggttcaaggaagagcacgcagctgtgtctact
gttatattccctgtctcatctatgtatctattgtcctcgttaaatatttctataacttga

>HepG2_L2_4;chr6|130815306|130815438|+|L2a|Distal.upstream
gctggtgtctgtctctcctggaatgttaaatacactctcataaggataaggaccttgcatttataaaggtatgttca
cagccccccagatcaggcctctgggcacatgttaagccttcctgtaataatatttctataacttga

>HepG2_L2_5;chr9|6590987|6591039|-|L2a|intron
tcataatgtcagctgcacacaaatgtctggtcagacctacttagtgccaaagcactgtctagctgatccga
gtaaccacacacagctggaaaccctgtctctgtggagtggtgagctgacattctctgg

>HepG2_MIR_1;chr5|55934179|55934315|+|MIR|Proximal.downstream
agaatgcggactttggatcaggttgagtttgaacctcactgtgtgtaaccacatgaccggcagttata
tcaccctctgaggctcattgtaatgtgtgtctgatattacccataagtatttcagtt

>HepG2_MIR_2;chr5|55934179|55934315|+|MIR|Proximal.downstream
agaatgcggactttggatcaggttgagtttgaacctcactgtgtgtaaccacatgaccggcagttata
tcaccctctgaggctcattgtaatgtgtgtctgatattacccataagtatttcagtt

>HepG2_MIR_3;chr6|155486858|155487027|+|MIR3|Intergenic
gttgagagaacattagtagtagtagagagagcgccccctgcatcttacctctaattctataatgctatgt
gataatctactgaccccactcagttgccccaaagttacaataaaaggtgtagtagtaaattctttgaaggtc
gctatatctgtatgttctgtaa

> HepG2_MIR_4; chr16|8987535|8987661|+|MIRc|Distal.upstream
ttaggcacactgggataagttcaggttagccactccccagggccccagcctgggtaggctgtggtctcttt
tgaacttcactgltttgttttggtaacacggaggtgtcacagtagctctgttct

> HepG2_MIR_5; chr4|17559340|17559451|+|MIR|Distal.upstream
acttaaggttaaaggtagagaccccttaggttcaacaggccagggtcuaatcctgtctgtgaaactgta
gcttggcctttgggaagatattttgtaacctgtc

> HeLa_MIR_1_scrambled
ggcttagatagagacacagcattgtgtgatgcacatctgtgtcattccatagacctatgaatccttttagcagtgt
tattactgtgcattctatagtacagacatactagtaatgagatagatccagataaaacactgggcccct
aattgtgaaattttctcttcctccgctgccccatttaacagattggtggtggcgttagtttagttgtgatgacc
Supplemental Figure S1. Mappability correction and quantification against control

(A) Illustration of variables used in the mappability correction and normalization against control. (B) Example of H3K4me1 in GM12878 after mappability correction and normalization against the input. X-axis is reads count in input and y-axis is reads count in ChIP data, both are after mappability correction. The points are the density of hexbin plot. (C) Correlation between mappability and quantitative units. X-axis is different histone variant and modification. Y-axis is the correlation between mappability and quantitative units. The RPKM values were obtained by iteres and signal in chip minus that in input.
Supplemental Figure S2. An integrated framework for cis-regulatory TE detection

(A) The first layer of enhancer-like repeats (ELRs) classification model, tuned key parameter of a random forest model in cell lines with curated gold standard positive and negative sets. The x-axis displays the tree number, the key parameter of the random forest model. The y-axis shows the mean value of Area Under Curve (AUC) in Receiver Operating Characteristic (ROC) through 10-fold cross validation (CV). (B) Evaluation of feature importance for the first layer ELRs classification model. All features were quantified histone modification markers. The error bar indicates standard deviation. (C)
Performance of the first layer of ELR classification model for crossing prediction. X-axis labels the cell line model used for training, and the error bars mark the AUC standard deviation evaluated in other cell lines through 10-fold CV. (D) Tuned key parameter of the random forest model in the second layer of ELR classification model. The x-axis shows the number of trees, and the left and right y-axes are mean AUC and mean zero-one-loss (ZOL) through corresponding 10-fold CV. The black line marks the AUC and the red line marks the ZOL. Feature importance evaluation was done for each cell line, and the second layer input is the output possibility of first layer models. The error bar indicates standard deviation. (E) Performance of the forest of forest model (FOFM) for ELR and promoter-like repeats (PLR) classification. Unique mapping and random hit (RH) mapping strategies are highly overlapped. (F) Boxplot of performance comparison between FOF and multiple forest model (MF). Left and right panel boxplots are distributions of 10-fold CV AUC and ZOL, respectively. Student’s t-test was used for testing the difference between both methods, a difference cutoff as p-value < 0.01 was considered significant and marked by black dots.
**Supplemental Figure S3. Integrated framework for PLRs detection**

(A) The first layer of PLRs classification model, tuned key parameter of random forest in each cell line with curated positive and negative sets. The x-axis displays the tree number, the key parameter of random forest model. The y-axis shows the mean value of AUC through 10-fold CV. (B) Evaluation of feature importance for the first layer PLRs classification model. (C) Performance of the first layer of PLRs classification model for crossing prediction. X-axis labels the cell line model used for training, and the error bars mark the AUC evaluated in other cell lines through 10-fold CV. (D) Color legends for cell lines used in above plot A, B and C. (E) Tuned key parameter of random forest model in the second layer of PLRs classification model. The x-axis shows the number of tree, and left y-axis and right y-axis are mean AUC and mean zero-one-loss (ZOL) through 10-fold CV. The black line marks the AUC and the red line marks the ZOL. Feature importance evaluation was done for each cell, and the second layer input is the output possibility of first layer models.
Supplemental Figure S4. Comparison of ELRs detected by FOFM and ChromHMM

(A,B) Overlap ratio between ELRs and ChromHMM states (ENCODE (G) and NIH Roadmap Epigenomics (H)). ChromHMM states were downloaded from http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChromModels/coreMarks/jointModel/final and the description of each state can be found in http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#core_15state. (C-F) Mean ChIP-seq profiles of H3K27ac, H3K4me1, H3K4me2, H3K4me3 and EP300 on enhancer-like repeats defined by FOFM and ChromHMM-enhancer-overlapped TEs (with requirement that more than half of TEs were overlapped).
Supplemental Figure S5. Widespread MIR and L2 form cis-regulatory elements in human cells
(A) Enriched TE families of ELRs and PLRs in collected ENCODE cell lines using the random hit mapping strategy. Bubble size indicates corrected enrichment p-value and color marks enrichment score. The enrichment test was carried by a combination of binomial test and hypergeometric test. (B) Enriched TE families of ELRs and PLRs in NIH Roadmap Epigenomics cell lines using the unconsolidated data.
Supplemental Figure S6. Neighbor joining trees of lincRNA, IncRNA and typical enhancers
(A) NJT of H3K4me1 profiles in upstream and downstream 1.5 kb of lincRNA TSSs. (B) NJT of H3K4me1 profiles in upstream and downstream 1.5 kb of lncRNA TSSs. (C) NJT of binary matrix for NonTE typical enhancers, data obtained from Roadmap

http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final/all.dense.browserFiles.tgz
Supplemental Figure S7. ELRs mark cell identities in Blueprint and CEEHRC data

(A) Principle component analysis (PCA) plot of ELRs in Blueprint normal and disease blood samples. (B) NJT of ELRs in Blueprint normal and disease blood samples. (C) PCA plot of ELRs in CEEHRC samples. (D) NJT of ELRs in CEEHRC samples.
Supplemental Figure S8. tsELRs are associated with TSGs’ expression

(A) Heatmap of redundant tissue specific genes (TSGs) selected by JSD method. (B-D) Frequency histogram of absolute distances from each TE to the nearest TSG in the group of ESC, Digestive and Epithelial.
Supplemental Figure S9. Heatmap of the ELRs binary matrix
Supplemental Figure S10. Profiles of hESC and iPSC specific ELRs

(A) Mean profiles of CHD7, EZH2, SUZ12 and input on ELRs in H1-hESCs (GSE32509). EZH2 and SUZ12 are not found to have enriched bindings in tsELRs, and therefore were used as negative controls. (B) Mean profiles of whole genome-wide bisulfite sequencing around ELRs, PLRs and background in H1-hESC (ENCODE accession ENCFF770YJW and ENCFF263KSB) (C) Mean profiles of GRO-seq around ELRs, PLRs and background in H1-hESCs (GSE41009). (D) Non-polyA and polyA RNA expression of ELRs in H1-hESC whole cells, cytosolic and nuclear fractions (GSE26284). (E) Expression of ELRs, PLRs and background non-polyA RNA in H1-hESC nuclei (GSE26284). Annotated p-values were obtained by Mann-Whitney rank test. (F) Box plot of ELRs’ non-polyA RNA nuclear expression across multiple cell types (GSE26284).
Supplemental Figure S11. MIR and L2 form *cis*-regulatory elements in mouse cells generally

(A) Enriched TE families of ELRs and PLRs in Mouse ENCODE cell lines using the unique mapping and random hit (RH) mapping strategy. Bubble size indicates corrected enrichment p-value and color marks enrichment score. The enrichment test was carried by a combination of binomial test and hypergeometric test. (B) Cumulative ratio of enhancer-like and promoter-like
MIR, L2 and all TEs in Mouse ENCODE data. (C) Saturation estimation of active TEs bound by EP300. Upper panel is the heatmap of detected active TEs bound by EP300 for mouse data, lower panel is cumulative ratio for the active TEs. (D) Top 10 enriched motifs for enhancer-like MIR and L2 in mouse cells.
Supplemental Figure S12. Enriched TE subfamilies in ENCODE data
### Table 1: Consolidate Hospital Sampling

| Hospital A | Hospital B | Hospital C | Hospital D | Hospital E |
|------------|------------|------------|------------|------------|
| 50%        | 30%        | 20%        | 10%        | 5%         |

### Diagram 1: Hospital Location

- Hospital A: North
- Hospital B: East
- Hospital C: South
- Hospital D: West
- Hospital E: Center

**Note:** The table and diagram represent sample distribution across five hospitals in different geographical locations.
Supplemental Figure S13. Enriched TE subfamilies in NIH Roadmap consolidated data
Supplemental Figure S14. MIR and L2 are highly associated in 3D genome

(A) Most MIR (left panel) and L2 (right panel) ELRs are tissue specific. (B) Distance distribution of MIR to nearest L2 in any genomic region (all), PLRs or ELRs. (C-D) H3K27ac (C) and H3K4me1 (D) ChIA-PET loops from L2 to MIR in K562 cell. Red dots mark those have significantly higher support and confidence compared to the all possible TE family pairs background. (E) Support and confidence for association between L2 and MIR between ChIA-PET interaction anchors. All repeats are ELRs. (F) Support and confidence for association between L2 to MIR (left panel), MIR to MIR (middle panel) and L2 to L2 (right panel) for collected loops from GEO and ENCODE. Black (real loops) and gray (random background) bars measure the support, blue line (real loops) and blue dash line (random background) indicate the confidence, gray and red dash line indicate the support (0.2) and confidence (0.5) cutoffs. Random background was generated as the same loop size and same loop number as the true loops, but with randomly selected regions as loop anchors. 1000 times permutation was used to draw the mean random background levels. (G-H) Density of distant L2-interacting tags that fall into
MIR, Alu, L1 or another L2 based on ChIA-PET PETs level data, PETs within 10kb were removed to avoid self-ligation PETs. (I-J) Source end is MIR or L2, target end is selected MIR, L2, Alu and L1. All repeats are ELRs. Only PETs with distance larger than 10kb are used to avoid self-ligation bias. (K-N) Source end is MIR or L2, target end is MIR, L2, Alu, L1 and ERVK. ERVK is a negative control while others are highly involved in interactions. Only PETs with distance larger than 10kb are used to avoid self-ligation bias.
Supplemental Figure S15. Examples of interacting MIR and L2 from K562 H3K27ac and H3K4me1 ChIA-PET loops

The loops were obtained from cLoops (v0.9, https://github.com/YaqiangCao/cLoops_supplementaryData/tree/master/SupplementaryData/loops/ChIA-PET). Only overlapped loops of H3K27ac and H3K4me1 that contain MIR interacting L2 were randomly selected.
Supplemental Figure S16. STARR-seq support enhancer activities of ELRs in HeLa cell

(A) Mean coverage (left) and coverage heatmaps (right) of STARR-seq and input signal in HeLa around ELRs centers. Data from GSE100432 (Muerdter et al. 2017). (B) Example of STARR-seq, H3K27ac and EP300 ChIP-seq signal
on enhancer like MIR (B) and L2 (C). These two ELRs were selected for enhancer luciferase reporter experiments in our manuscript.

Reference
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.

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.

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

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-e30377.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F et al. 2012. Landscape of transcription in human cells. *Nature* **489**: 101-108.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 14863-14868.

Griffon A, Barbier Q, Dalino J, van Helden J, Spicuglia S, Ballester B. 2015. Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. *Nucleic Acids Res* **43**: e27.

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. *Molecular cell* **38**: 576-589.
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.

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

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 biology* 12: 1.

Wagner GP, Kin K, Lynch VJ. 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences* 131: 281-285.

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.

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.