Figure S1

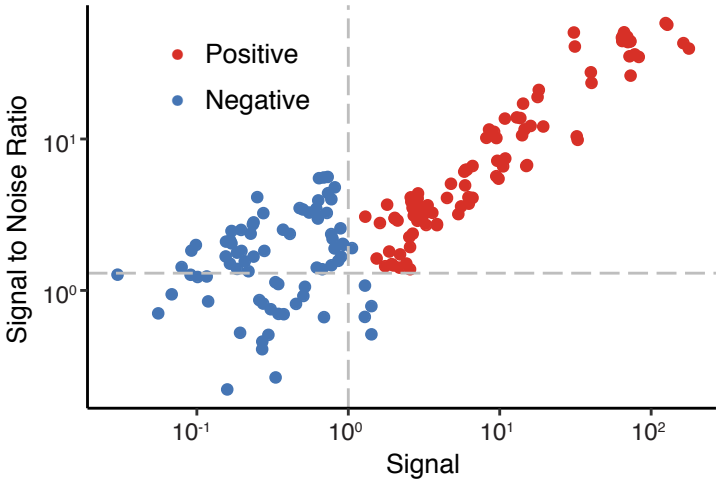

Positive
Negative
Signal to Noise Ratio
10^1
10^0
10^{-1}
10^0
10^1
10^2
Signal

Figure S2

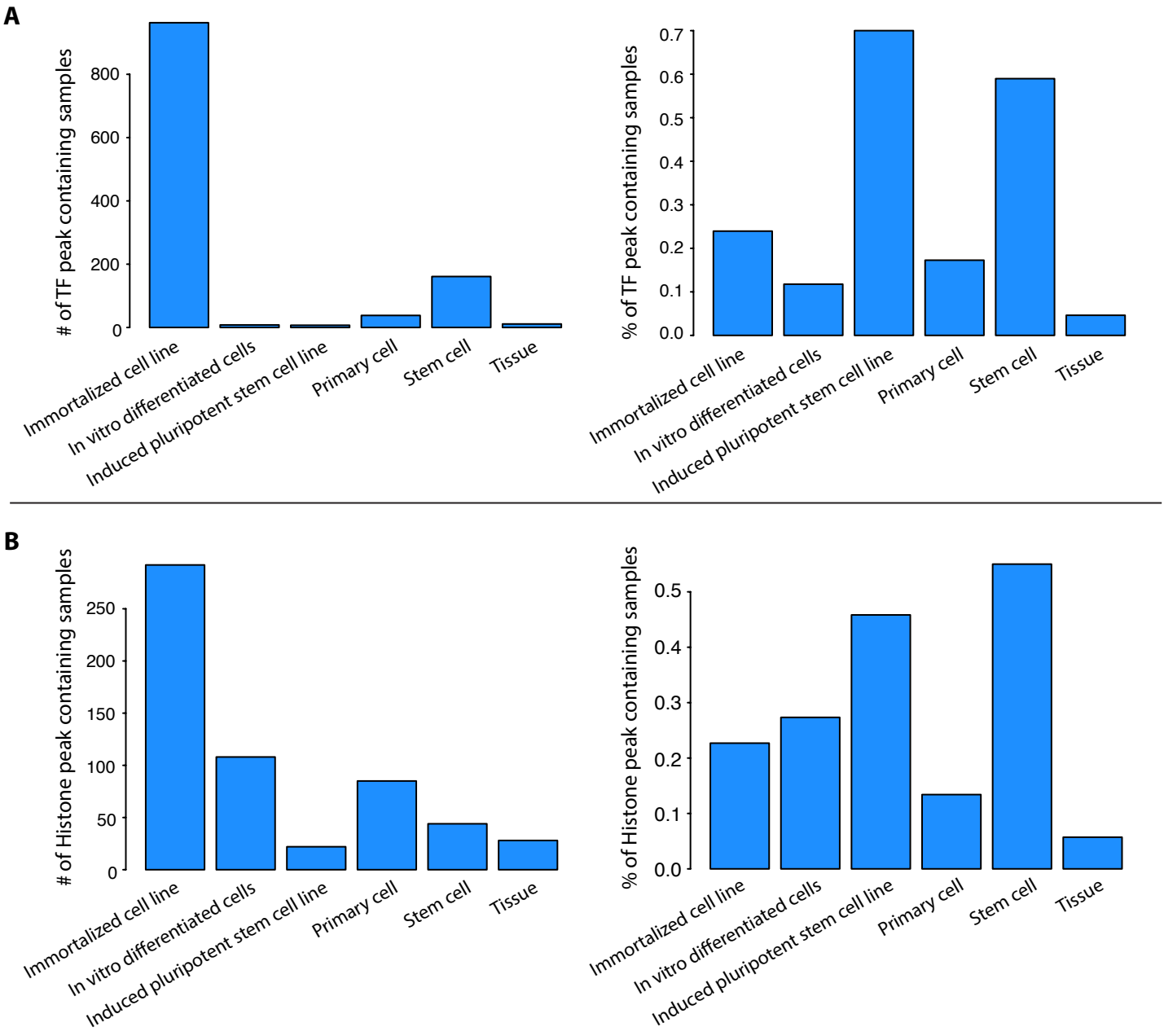

A
# of TF peak containing samples
0
200
400
600
800
Immortalized cell line
In vitro differentiated cells
Induced pluripotent stem cell line
Primary cell
Stem cell
Tissue
% of TF peak containing samples
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Immortalized cell line
In vitro differentiated cells
Induced pluripotent stem cell line
Primary cell
Stem cell
Tissue
B
# of Histone peak containing samples
0
50
100
150
200
250
Immortalized cell line
In vitro differentiated cells
Induced pluripotent stem cell line
Primary cell
Stem cell
Tissue
% of Histone peak containing samples
0.0
0.1
0.2
0.3
0.4
0.5
Immortalized cell line
In vitro differentiated cells
Induced pluripotent stem cell line
Primary cell
Stem cell
Tissue

Figure S3

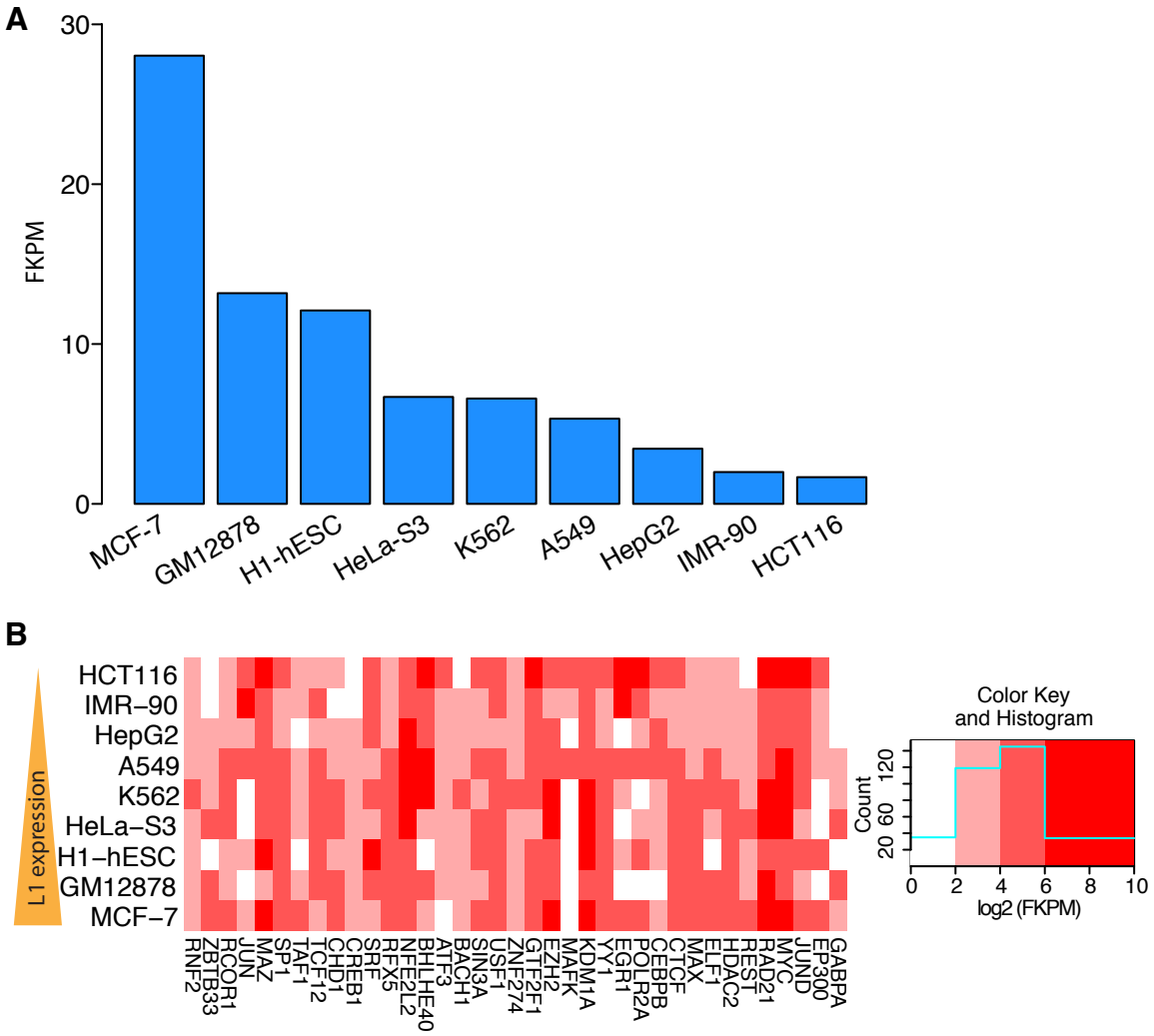

A
30
20
10
0
FKPM
MCF-7
GM12878
H1-hESC
HeLa-S3
K562
A549
HepG2
IMR-90
HCT116
B
HCT116
IMR−90
HepG2
A549
K562
HeLa−S3
H1−hESC
GM12878
MCF−7
L1 expression
RNF2
ZBTB33
RCOR1
JUN
MAZ
SP1
TAF1
TCF12
CHD1
CREB1
SRF
RFX5
NFE2L2
BHLHE40
ATF3
BACH1
SIN3A
USF1
ZNF274
GTF2F1
EZH2
MAFK
KDM1A
YY1
EGR1
POLR2A
CEBPB
CTCF
MAX
ELF1
HDAC2
REST
RAD21
MYC
JUND
EP300
GABPA
Color Key
and Histogram
Count
20
60
120
0
2
4
6
8
10
log2 (FKPM)

Figure S4

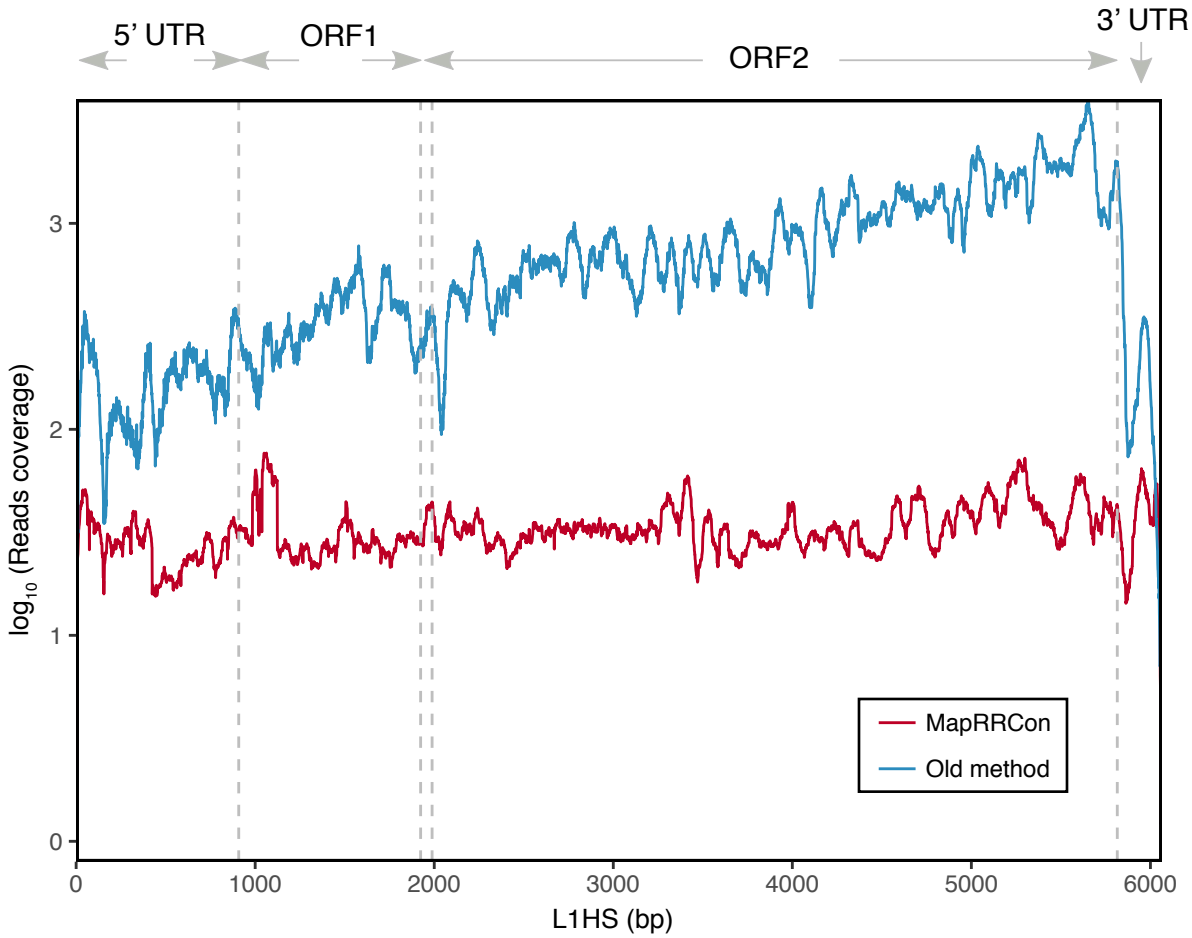

5' UTR
ORF1
ORF2
3' UTR
log10 (Reads coverage)
0
1
2
3
0
1000
2000
3000
4000
5000
6000
L1HS (bp)
MapRRCon
Old method

Figure S5

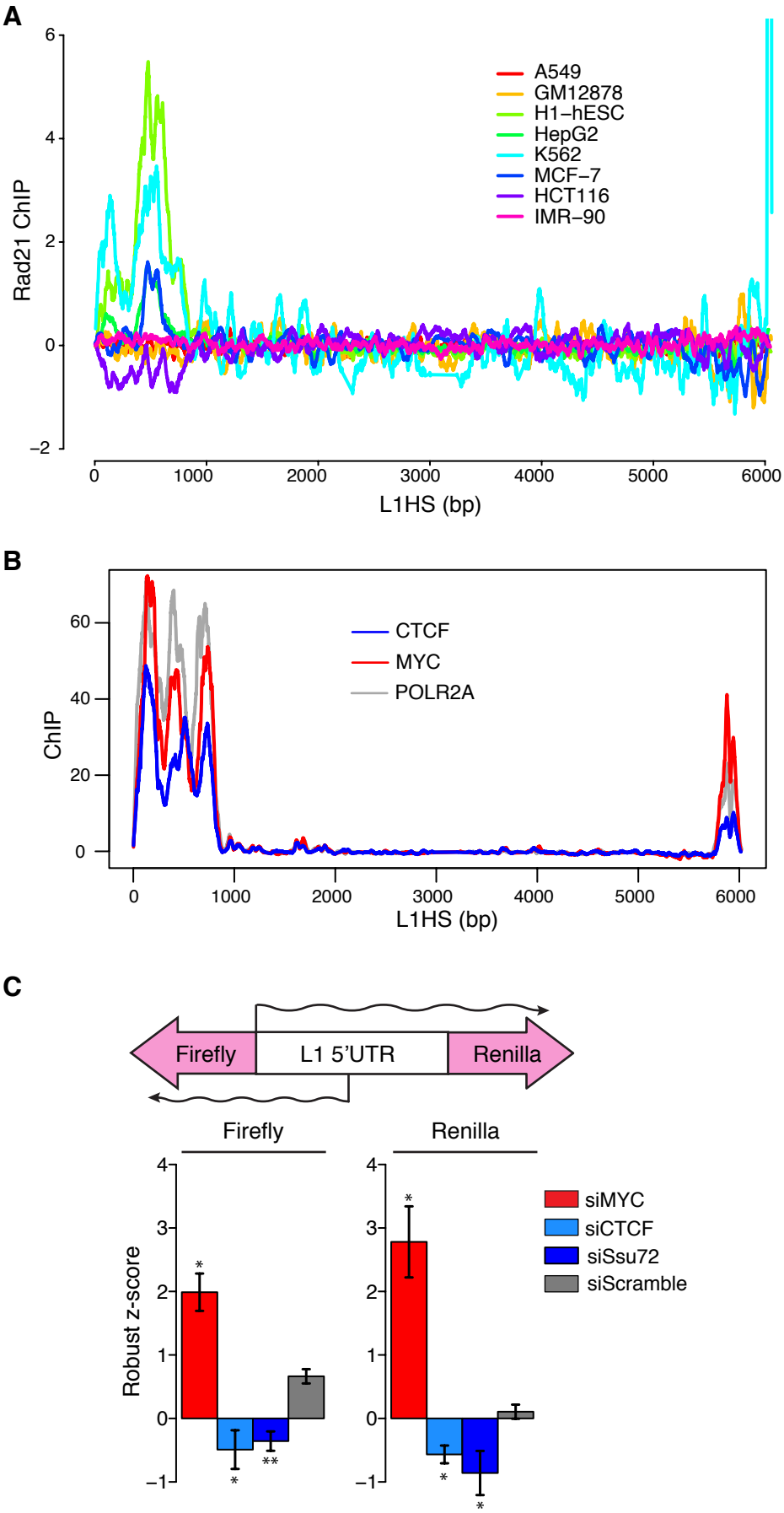

A
Rad21 ChIP
L1HS (bp)
A549
GM12878
H1−hESC
HepG2
K562
MCF−7
HCT116
IMR−90
B
ChIP
L1HS (bp)
CTCF
MYC
POLR2A
C
Firefly
L1 5'UTR
Renilla
Firefly
Renilla
Robust z-score
siMYC
siCTCF
siSsu72
siScramble

Figure S5

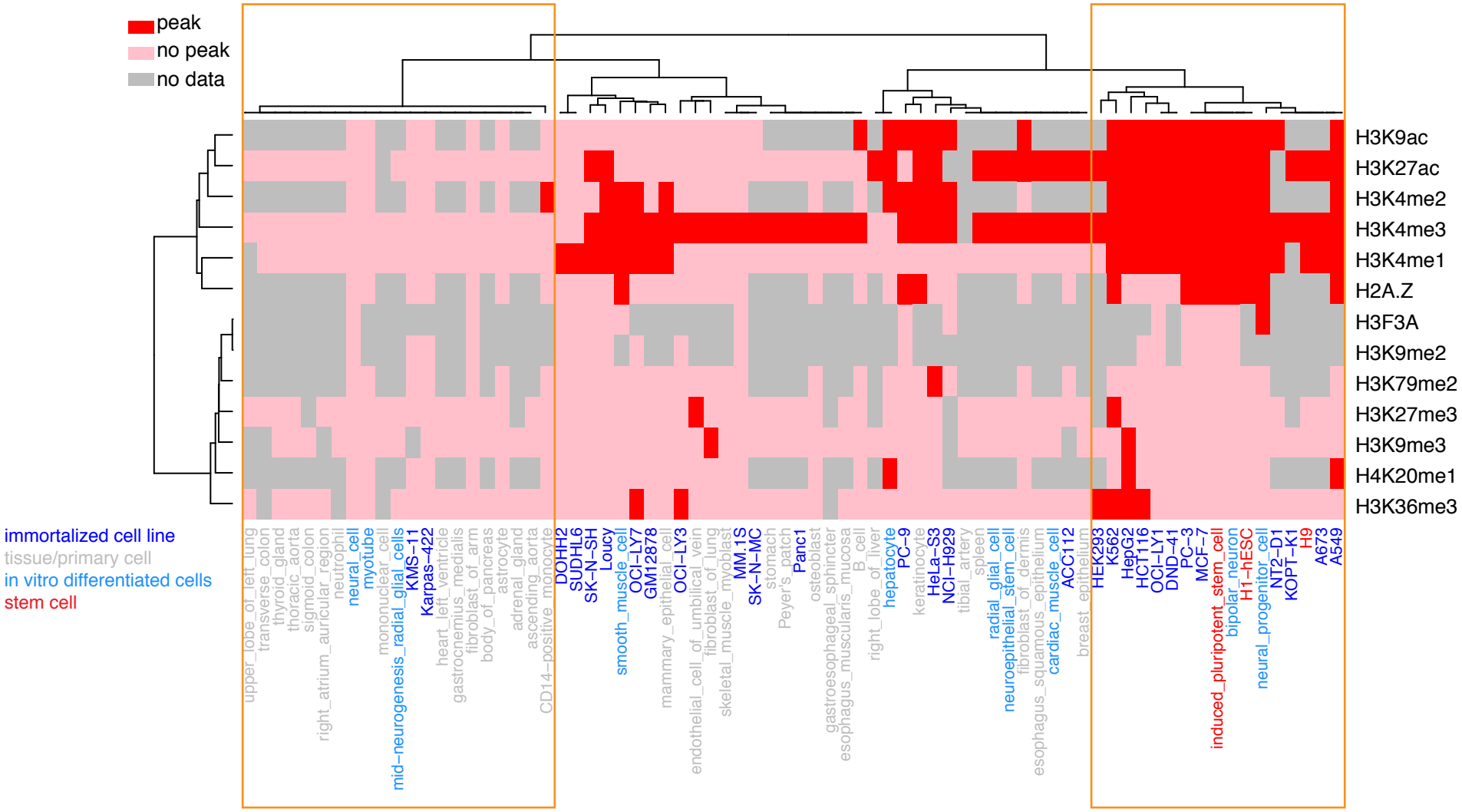

peak
no peak
no data
H3K9ac
H3K27ac
H3K4me2
H3K4me3
H3K4me1
H2A.Z
H3F3A
H3K9me2
H3K79me2
H3K27me3
H3K9me3
H4K20me1
H3K36me3
immortalized cell line
tissue/primary cell
in vitro differentiated cells
stem cell
upper_lobe_of_left_lung
transverse_colon
thyroid_gland
thoracic_aorta
sigmoid_colon
right_atrium_auricular_region
neutrophil
neural_cell
myotube
mononuclear_cell
mid−neurogenesis_radial_glial_cells
KMS−11
Karpas−422
heart_left_ventricle
gastrocnemius_medialis
fibroblast_of_arm
body_of_pancreas
astrocyte
adrenal_gland
ascending_aorta
CD14−positive_monocyte
DOHH2
SUDHL6
SK−N−SH
Loucy
smooth_muscle_cell
OCI−LY7
GM12878
mammary_epithelial_cell
OCI−LY3
endothelial_cell_of_umbilical_vein
fibroblast_of_lung
skeletal_muscle_myoblast
MM.1S
SK−N−MC
stomach
Peyer's_patch
Panc1
osteoblast
gastroesophageal_sphincter
esophagus_muscularis_mucosa
B_cell
right_lobe_of_liver
hepatocyte
PC−9
keratinocyte
HeLa−S3
NCI−H929
tibial_artery
spleen
radial_glial_cell
neuroepithelial_stem_cell
fibroblast_of_dermis
esophagus_squamous_epithelium
cardiac_muscle_cell
ACC112
breast_epithelium
HEK293
K562
HepG2
HCT116
OCI−LY1
DND−41
PC−3
MCF−7
induced_pluripotent_stem_cell
bipolar_neuron
H1−hESC
neural_progenitor_cell
NT2−D1
KOPT−K1
H9
A673
A549

Figure S7

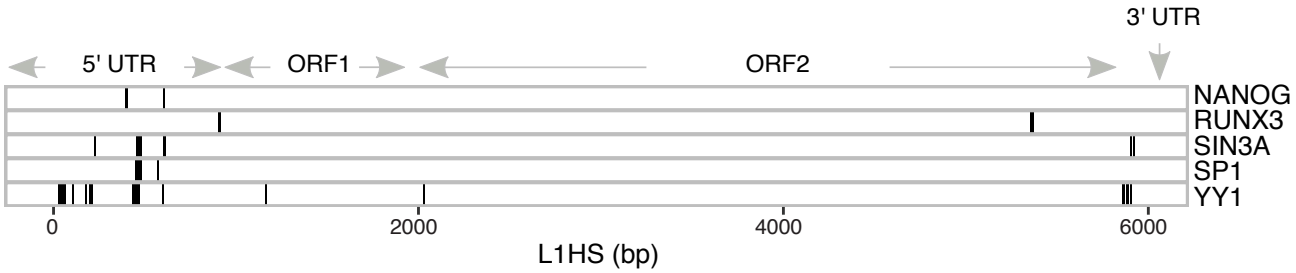
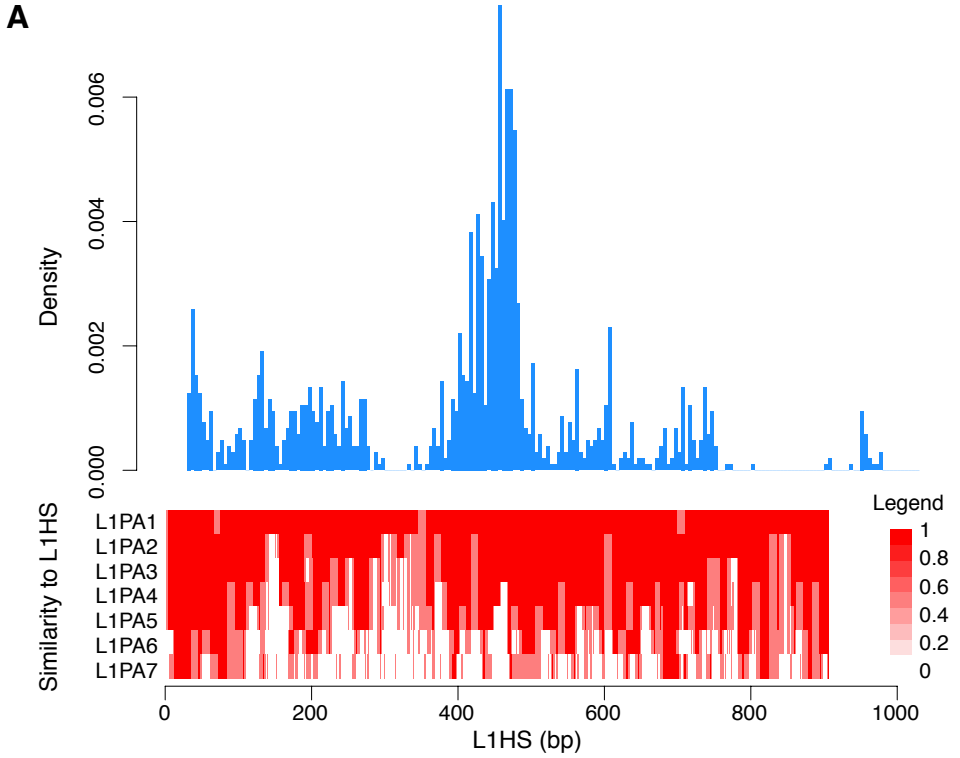

3' UTR
5' UTR
ORF1
ORF2
NANOG
RUNX3
SIN3A
SP1
YY1
0
2000
4000
6000
L1HS (bp)

Figure S8


A
Density
0.000
0.002
0.004
0.006
Similarity to L1HS
L1PA1
L1PA2
L1PA3
L1PA4
L1PA5
L1PA6
L1PA7
Legend
1
0.8
0.6
0.4
0.2
0
0
200
400
600
800
1000
L1HS (bp)

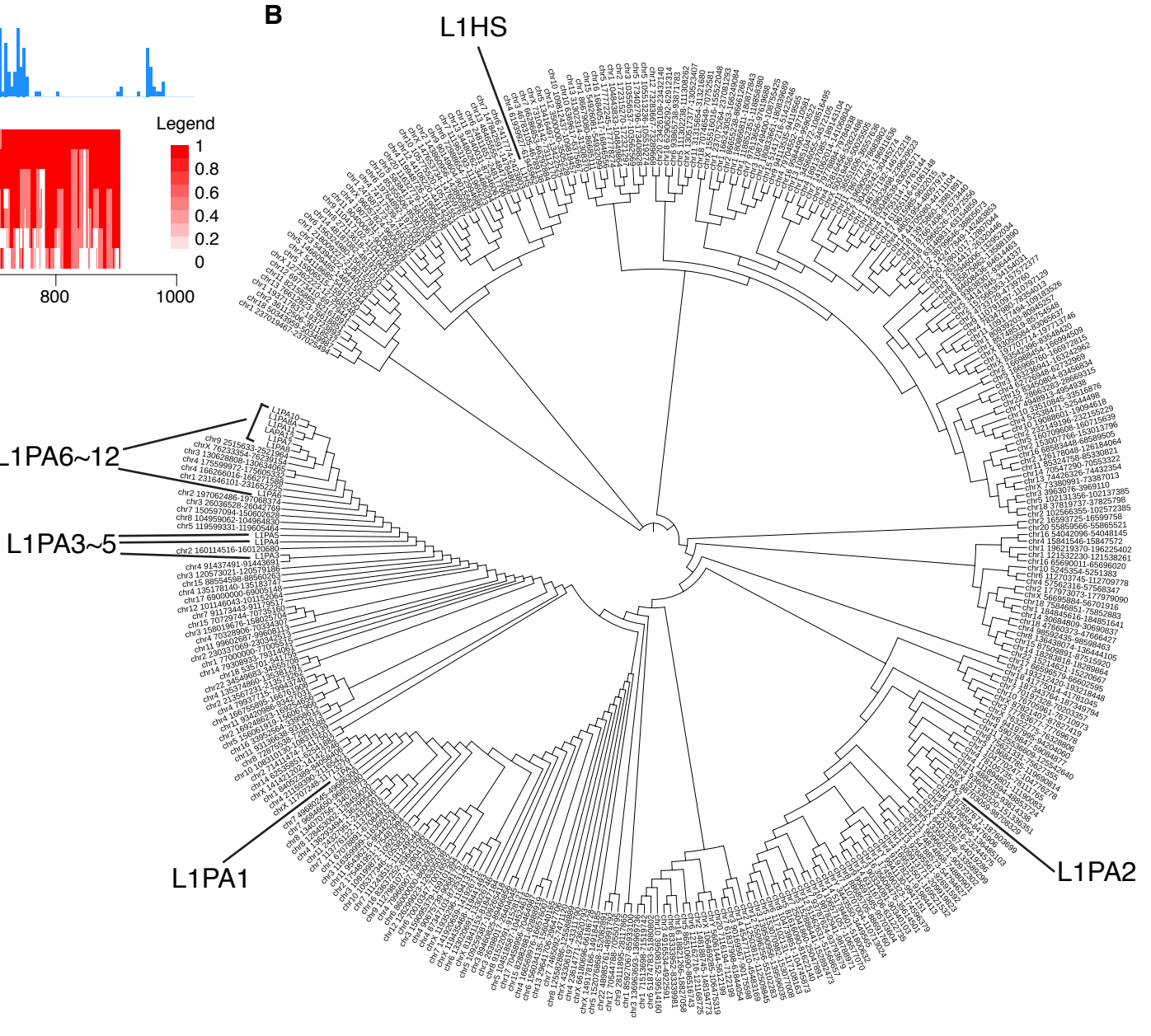
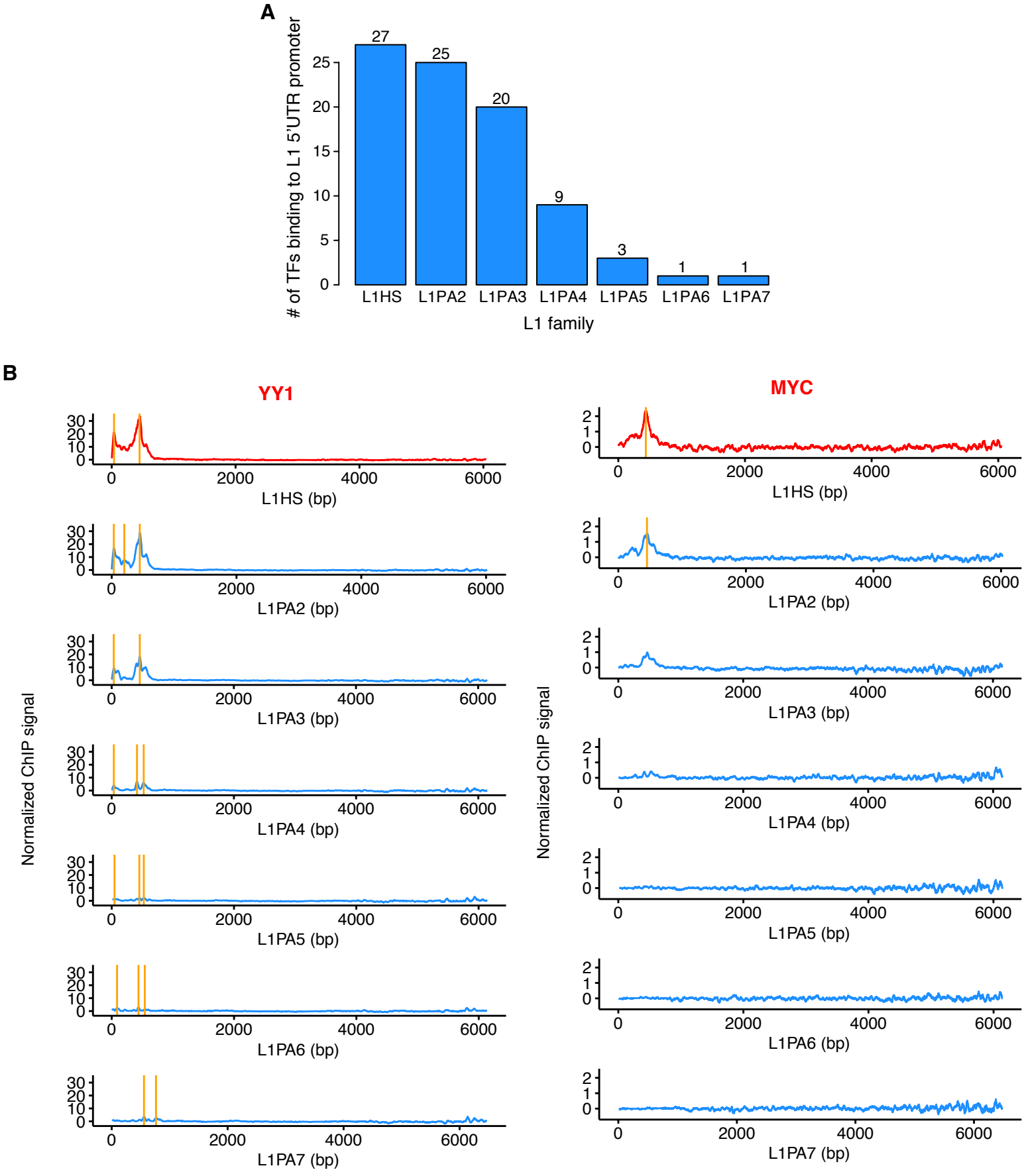

B
L1HS
L1PA6~12
L1PA3~5
L1PA1
L1PA2

Figure S9


A
# of TFs binding to L1 5'UTR promoter
27
25
20
9
3
1
1
L1HS
L1PA2
L1PA3
L1PA4
L1PA5
L1PA6
L1PA7
L1 family
B
YY1
MYC
Normalized ChIP signal
L1HS (bp)
L1PA2 (bp)
L1PA3 (bp)
L1PA4 (bp)
L1PA5 (bp)
L1PA6 (bp)
L1PA7 (bp)

SI figure legends

Figure S1. Training datasets of the Peak Calling Algorithm
We optimized two parameters of the algorithm to fit our datasets. We used 91 positive and 77 negative datasets to train the algorithm. After plotting them based on intensity and signal-to-noise ratio, we found that these two datasets clustered into two populations, separated by two thresholds (normalized intensity >1; signal-to-noise ratio >1.3).

Figure S2. Overview of the TF/histone mark peaks in different biosamples.
AB. We plot the number of samples (without removing replicates) that contain peaks at L1. The right panel shows the percentage of peak-containing samples in each of the categorized groups.

Figure S3. Expression level of L1 and TFs in various cell lines.
A. L1 expression level in 9 ENCODE cell lines quantified by RNA-seq data. Three bars for each cell lines represents FKPM (Fragments Per Kilobase of transcript per Million mapped reads) values L1. **B.** Expression (log2 FPKM) of a selected number of TFs (same set in Figure 2D) in 9 cell lines.

Figure S4. No 3' bias on L1HS is observed for RNA-seq data in 77 breast tumor samples.
The distributions of mapped reads are plotted along the L1HS, and each line represents one tumor sample. The boundaries of 5'UTR, ORF1, ORF2 and 3'UTR are shown with gray dashed lines. The lack of a 3' bias is consistent with most of the reads coming from full length transcripts and furthermore shows that the level of genomic DNA contamination in the RNA prep is very low.

Figure S5. CTCF may mediate L1 gene loops.
A. ChIP profiles of Rad21 in various cell lines on L1HS (replicates were averaged to form a single line). **B.** Colocalization of CTCF, c-Myc and RNA polymerase II ChIP-seq signals in MCF-7 cells. **C.** siRNA knockdown experiments of Ssu72. See Figure 4D legends for details.

Figure S6. Histone marks that enrich at L1 5'UTR in various samples.
We present the binding profiles of 13 histone marks in various samples with colors indicating whether peaks are detected at L1 5'UTR promoter. The samples are grouped into 4 categories shown by different colors on the x axis. Both dimensions are ranked by hierarchical clustering as described in Figure 1A. Two distinct clusters are marked with orange boxes: (1) The left box showed a cluster of samples that have no histone marks at L1 5'UTR promoter. (2) The right box showed a cluster of samples that have enriched active histone marks at L1 5'UTR promoter.

Figure S7. Peaks identified for known TFs.
Peak locations identified in this study of 5 published TFs, previously shown to bind and regulate L1 transcription, were plotted along L1HS.

Figure S8. Sequence similarity between L1 consensus sequences.
A. We calculated the pairwise sequence similarity between each of the L1PA1-xx consensus sequences to L1HS on a 10-bp sliding window with a step of 1-bp. For each window of the L1HS consensus sequence, we searched for sequences in L1PA with either a perfect match or 1-bp

mismatch. We then assigned the score 1 if there was a perfect matched sequence or as 0.5 if the best match had a 1-bp mismatch in each window. The score was set to 0 if the best matched sequence had more than 1-bp mismatches. The similarity scores are color coded (shown in red) and aligned with the peak density histogram (shown in blue). **B.** 332 Repeakmasker annotated L1HS insertion sequences are compared to L1 consensus sequences (L1HS and L1PAs) via phylogenetic analysis. Multiple sequence alignment was performed using Clustal Omega(1) and the tree was generated by ITOL (2).

Figure S9. TF binding profiles on L1PA families.
A. We ran MapRRCon on a subset of hESC datasets (182 datasets, 36 TFs: ATF3, BACH1, BRCA1, CHD1, CHD2, CTCF, EGR1, EP300, FOSL1, GABPA, GTF2F1, JUN, JUND, KDM1A, KDM6A, MAFK, MYC, NRF1, PHF8, POLR2A, RAD21, REST, RFX5, RXRA, SAP30, SIN3A, SIX5, SP1, SRF, TAF1, TAF7, TBP, USF1, USF2, YY1, ZNF143) using each one of the following consensus sequences and its corresponding insertion locations annotated by Repeatmasker. After combining the replicates, we plotted the number of TFs that show peaks at 5'UTR at each of the consensus sequences. **B**. Binding profiles of two TFs on each of the L1 consensus sequence are shown as examples. Orange lines indicate locations of peaks called by our peak-calling algorithm.

Reference

1. Sievers F, et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7(1):539–539.

2. Letunic I, Bork P (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucl Acids Res* 44(W1):W242–5.