Supplementary materials for "Chromatin environment, transcriptional regulation and splicing distinguish lncRNAs and mRNAs"

Marta Melé, Kaia Mattioli, William Mallard, David M Shechner, Chiara Gerhardinger and John Rinn

Index

Supplementary methods

Nuclear fractionation of mouse ES cells and RNA sequencing.

We isolated cytoplasmic and nuclear fractions essentially as described (Hacisuleyman et al. 2014) for mESC cells. We grew V6.5 mESC cells (Novus Biologicals NBP1-41162) on gelatinized 15-cm dishes, as described previously (Hacisuleyman et al. 2014). We harvested them at approximately 80% confluency by trypsinization, quenched them with growth media, pelleted them at 200×g for five minutes in a swinging-bucket centrifuge, and washed with ice-cold PBS. We pelleted and washed cells twice more, after which we estimated the packed cell pellet volume ("cv") by eye. We resuspended pellets with five cv's of ice-cold Cyto Extract Buffer(+) (20 mM Tris, pH 7.6, 0.1 mM EDTA, 2 mM $MgCl_2$, supplemented with 0.5 U/mL RNAseOUT (Life Technologies) and 1x EDTA-free Proteinase Inhibitor Cocktail (Roche)), and allowed them to swell by incubation at room temperature for two minutes, and on ice for ten minutes more. We lysed the cells by dropwise addition of CHAPS to 0.6% final, gentle mixing, and two passages through a syringe equipped with a 20G needle. We then clarified the lysates by centrifugation at 500×g for five minutes in a tabletop microcentrifuge at 4°C, producing nuclear pellets and cytoplasmic supernatants. We retrieved seventy percent of the supernatant samples, and further clarified them by centrifugation at 1000×g in a tabletop microcentrifuge at 4°C for five minutes, and stored samples at –80°C prior to use. We washed the nuclear pellets twice by gentle resuspension into five cv's of Nuclear Wash Buffer(+) (Cyto Extract Buffer, supplemented to 0.6% CHAPS and with inhibitors, as above), followed by centrifugation at 500×g at 4°C for five minutes. We gently reuspended the washed nuclei into two cv's of Nuclei Resuspension Buffer(+) (10 mM Tris, pH 7.5, 150 mM NaCl, 0.15% (v/v) NP–40, supplemented with inhibitors, as above), layered them onto a cushion of five cv's Sucrose Buffer(+) (10 mM Tris, pH 7.5, 150

mM NaCl, 24% (w/v) Sucrose, plus inhibitors, as above), and pelleted them at 14,000 rpm for 10 minutes in a tabletop microcentrifuge at 4°C. We resuspended the resulting nuclear pellets into two cv's of ice-cold PBS(+) (PBS, supplemented to 1 mM EDTA, and with inhibitors, as above) and pelleted them at 500×g for five minutes. The resulting pellets contained the final purified nuclei. We isolated RNA from cytoplasmic and nuclear fractions using Trizol LS and Trizol reagents (Thermo Scientific), respectively, following the manufacturer's protocols. Following extraction, we isopropanol precipitated aqueous RNA using GlyoBlue as coprecipitent (Thermo Scientific). We washed RNA pellets once with 70% ethanol, we resuspended them into 100 mL water, and further purified them using RNeasy micro spin columns (QIAGen), with on-column DNase treatment. We measured integrity and concentration of eluted RNA using an Agilent 2100 model Bioanalyzer. We prepared poly(A)+ mRNA-seq libraries from 500 ng of each RNA sample, using the TruSeq RNA sample preparation kit, v2 (Illumina) as described (Goff et al. 2015). We checked that known nuclear genes enriched in the nucleus were present in the corresponding nuclear fraction and vice versa (Supplemental Fig S41).

_K_-mer analyses:

To perform discriminative *k*-mer enrichment analyses, we used the program Jellyfish (Marçais and Kingsford 2011). For each *k*-mer, we then calculated the log2 fold ratio between sequences of interest: either *k*-mers enriched in the downstream sense direction of a TSS as compared to the upstream antisense direction or *k*-mers enriched in stable versus unstable transcripts (details in supplementary methods)

Sense/antisense TSS: To analyze *k*-mers enriched in the downstream sense direction of a TSS as compared to the upstream antisense direction, we used 1 kb downstream on the sense strand as well as 1 kb upstream on the antisense strand of all annotated lincRNA and mRNA TSSs. We then counted hexamers present in sense and antisense sequences and calculated the log2 fold ratio between these two groups. We used hexamers similarly to what has been used in similar analyses (Almada et al. 2013).

Stability: To analyze *k*-mers enriched in stable versus unstable transcripts, we ranked lincRNAs and mRNAs based on their experimentally derived half lives (see below) and selected those lincRNAs and mRNAs within the first and last quartile of the ranking. We selected 500 nt from either their 5´ or 3´ ends (lincRNAs) or UTRs (mRNAs) of the longest transcript. We then counted 7-mers present in stable and unstable transcripts in each group and calculated the log2 fold ratio. We considered a *k*-mer significantly enriched or depleted using a binomial test and adjusted p-values using an FDR of 0.05. To find whether certain repetitive elements were enriched or depleted in stable versus unstable transcripts, we compared the half life of those transcripts with at least one exon that overlapped a specific repeat family annotated by repeat masker (http://www.repeatmasker.org) versus those that did not.

ESE density calculations:

As ESEs are concentrated near exon-intron boundaries, we then mapped ESEs in the 100bp of exons adjacent to exon-intron boundaries using custom Python scripts. We calculated density of ESE motifs per gene (# successfully mapped motifs / # bp) as well as mean ESE density per nucleotide (# successfully mapped motifs / # sequences / # possible motifs). Intergenic regions were defined to be randomly selected 200nt long regions that do not overlap any type of GENCODE annotation +/- 5kb, including genes, small non-coding RNAs, pseudogenes, antisense transcripts, etc. To control for GC content, we used the R package matchIt with default settings (Ho et al. 2011) to match GC content of 200bp long exon-intron boundaries (100bp of exon and 100bp of intron) in lincRNAs with mRNAs and

intergenic regions. To control for repetitive elements, we removed any exon-intron boundaries that overlapped any repeat annotated by repeat masker (http://www.repeatmasker.org).

CLIP seq data analysis

All reads were mapped against human hg19 using TopHat. In those cases where the adaptor sequences were available, we removed them from the reads using trimmgalore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). When the adaptor sequences were not described in the publication, we followed the same strategy as in (Kelley et al. 2014) in which we iteratively trim and re-map reads that do not align. In all cases, we removed duplicates using Samtools (Li et al. 2009). Peaks were called using a method that parameterizes a null model using inferred isoform abundances for each gene (Kelley et al. 2014). To increase power, reads from all replicates were analyzed together as in (Kelley et al. 2014).

We calculated peak density by creating a set of merged lincRNA and mRNA transcript regions including a set of gene loci, a set of exons, and a set of all non-redundant annotated 3´ and 5´ splice site regions of 100bp centered in the exon-intron junction. To explore the binding patterns of HuR, we divided all exons into first, middle and last. First exons were those that overlapped the transcription start site; last exons were those that overlapped the transcription end site and middle were all remaining ones. We discarded single exon transcripts that spanned the whole gene locus. We merged all intervals before calculating density to avoid counting the same genomic region more than once.

RNA stability assay.

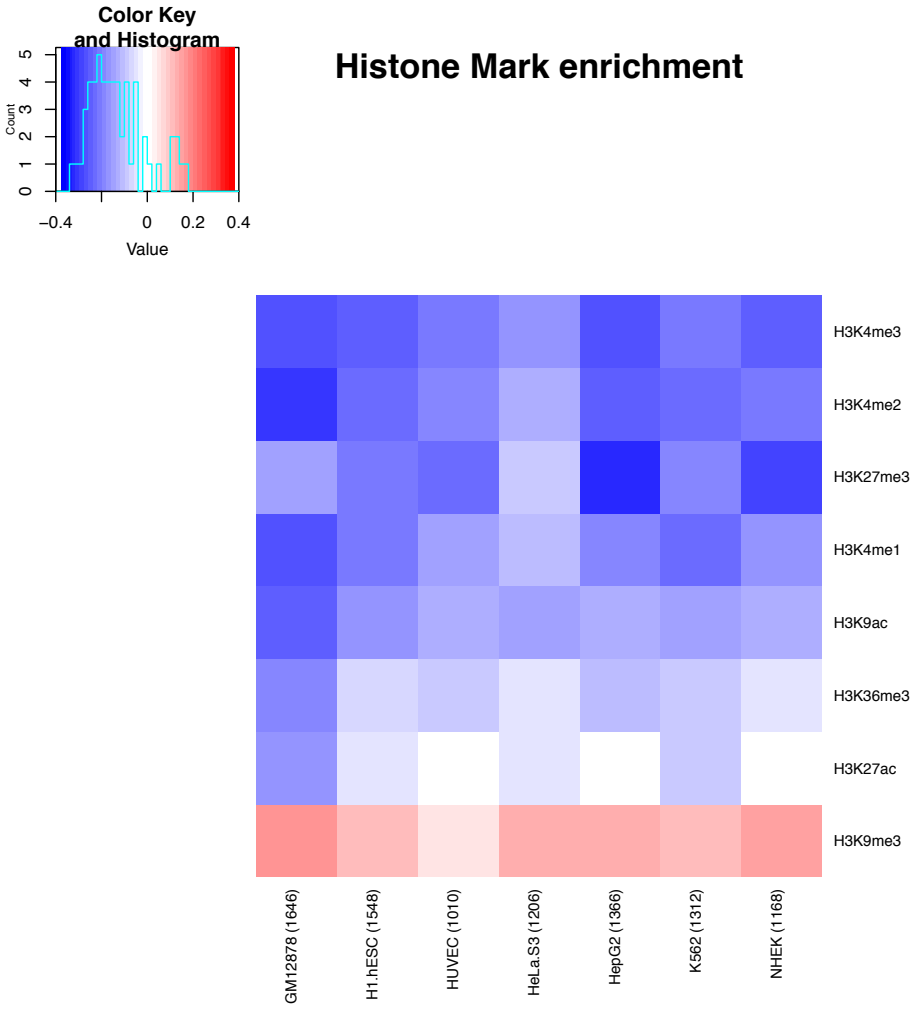
We cultured K-562 cells (ATCC) in RPMI 1640 supplemented with 10% FBS and 1X antibiotic-antimycotic (all from Life Technologies) in 75 cm^2 flasks. Once the cells reached a density of $0.7x10^6$ we added actinomycin D (Sigma) to a final concentration of 5 μg/ml of culture. We collected aliquots (2 ml) at time 0, 30 min, 2, 4 and 8 hours, pelleted the cells by centrifugation at 700 rpm for 5 min and immediately lysed them in TRIzol reagent (Life Technology) for subsequent RNA isolation. We collected smaller aliquots (0.2 ml) at each time point for cell counting and morphological evaluation. We maintained HUES9 cells (HSCI iPS Core, Harvard University) on ~15,000cells/cm^2 irradiated murine embryonic fibroblasts (MEFS, Global Stem) in KO DMEM supplemented with 20% Knockout Serum Replacement, 2 mM GlutaMax, 1% MEM Non-Essential Amino Acid, 10 ng/ml bFGF (all from Life Technologies) and 55 μM β-mercaptoethanol. We passaged the cells every 4-5 days using Collagenase IV (1 mg/ml, Life Technologies). Before starting the experiment, we adapted the HUES9 cells to feeder-free conditions by culturing them in Geltrex (Life Technologies) coated plates in mTeSR1 medium (Stem Cell Technologies) for two passages (EDTA 0.5mM). We then plated feeder-free cells as clumps on 6-well plates and we cultured to ~70% confluency prior to adding actinomycin D to a final concentration of 5 μg/ml medium. At time 0, 30 min, 2, 4 and 8 hours, we removed the medium and immediately lysed the cells in TRIzol reagent for subsequent RNA isolation. We used a parallel 6-well plate for morphological evaluation and cell counting. We performed triplicate actinomycin D experiments for both K-562 and HUES9 cells.

RNA isolation and RT-qPCR. We isolated total RNA from the TRIzol aqueous-phase by spin-column purification (RNeasy Mini kit, Qiagen) following manufacturers instructions as described (Sauvageau et al. 2013). We reverse transcribed 0.5 μg (HUES9) or 1 ug (K562) of total RNA for each sample using SuperScript III Reverse Transcriptase and random hexamers (Life Technologies). We performed qPCR with 4μl of cDNA, diluted to the

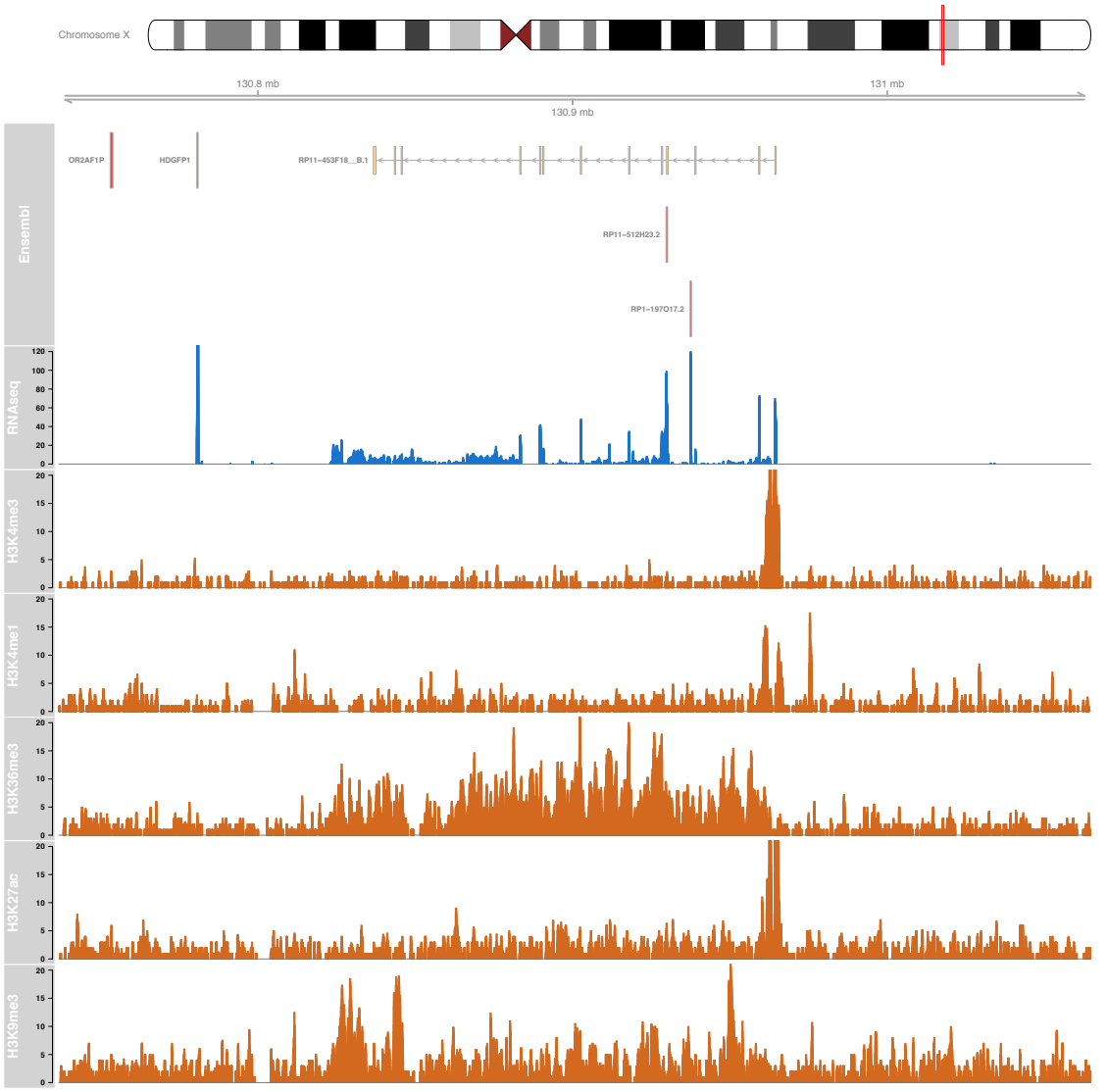
equivalent of 5 ng/µl of input RNA, using human MYC and GAPDH primers (IDT) and FastStart Universal SYBR Green Master–Rox (Roche) on an ABI7900HT real-time PCR (Applied Biosystems). We determined relative expression of MYC at each time point by the comparative Ct method using GAPDH as the endogenous control for input normalization and the ΔCt of the 0h time point within each replicate as the calibrator for ΔΔCt calculation. RT-qPCR values for c-myc and GAPDH are in Supplemental Fig S42.

RNA-Seq library preparation and sequencing. We performed RNA-Seq for the 0, 30 min, 2, 4, and 8h time points of 3 replicate RNA stability experiments of each cell type. We constructed poly(A)+ mRNA-seq libraries (TruSeq RNA Sample Preparation Kit v2, Illumina) as previously described (Goff et al. 2015) using 500 ng of total RNA as input and a 10-cycle PCR enrichment. We sequenced the indexed libraries in pools of 5 on the Illumina HiSeq 2500 platform using the rapid-full flow cell, paired-end, 75 bp read-length sequencing protocol (Genomic Platform, Broad Institute).

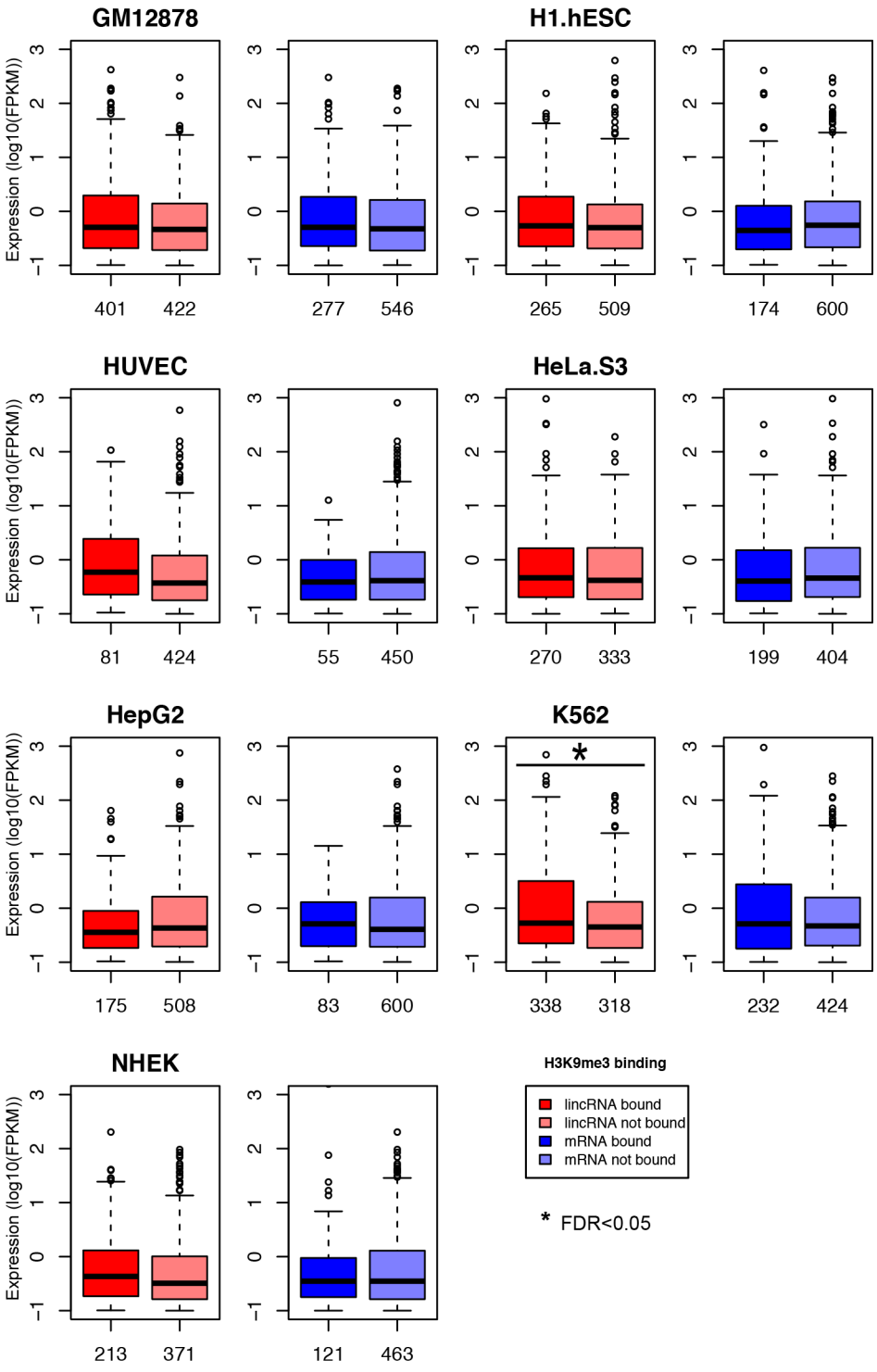
Supplementary figures



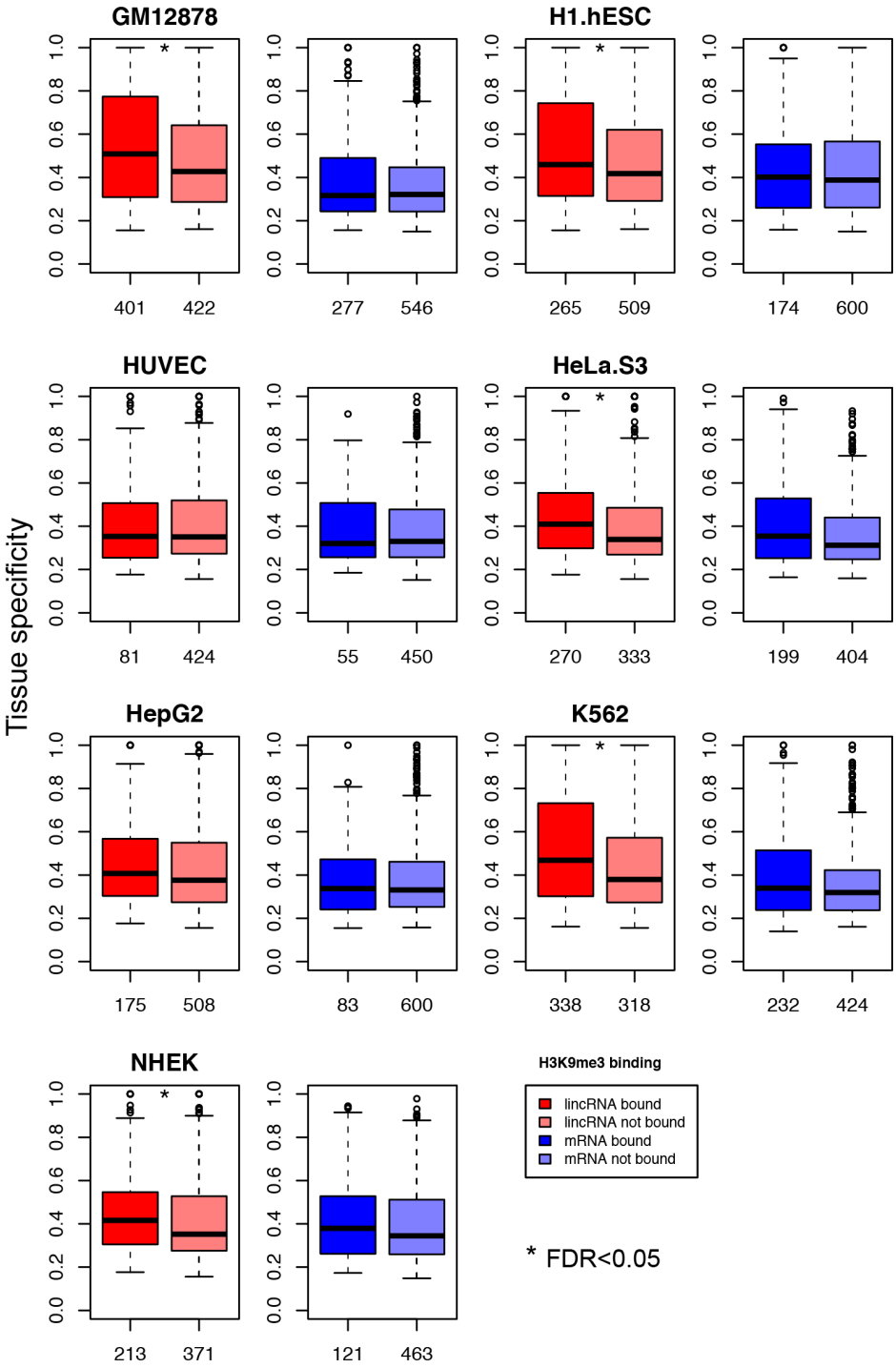
Supplemental Fig S1. Fisher effect size differences between lincRNAs and mRNAs promoters comparing presence absence of histone marks when using the narrow definition of promoter (-2Kb/+1Kb from the TSS). The total number of genes analyzed ranged is indicated in parenthesis and corresponded to all promoters of lincRNAs expressed at more than 0.1 FPKM in a specific cell line and the same number of expression matched mRNAs. Blue corresponds to larger values in mRNAs and red to larger values in lincRNAs. H3K9me3 is the only histone mark enriched in active lincRNA promoters compared to active mRNA promoters of the same expression levels.



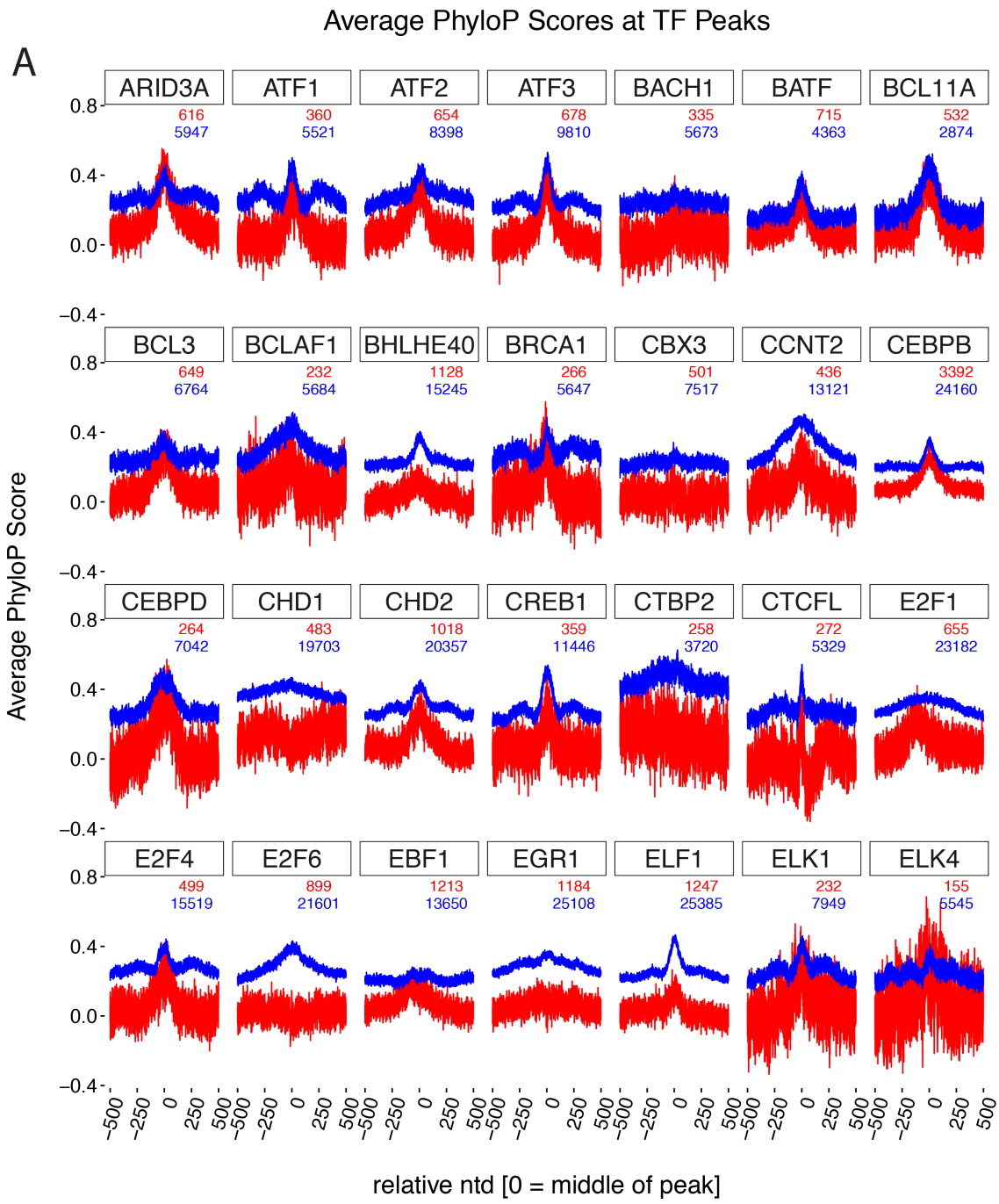
Supplemental Fig S2. Coverage distribution for ChIP-seq of five histone marks (orange) across the FIRRE lincRNA locus in H1-ESCs. H3K9me3 together with H3K36me3 is present across the FIRRE locus.

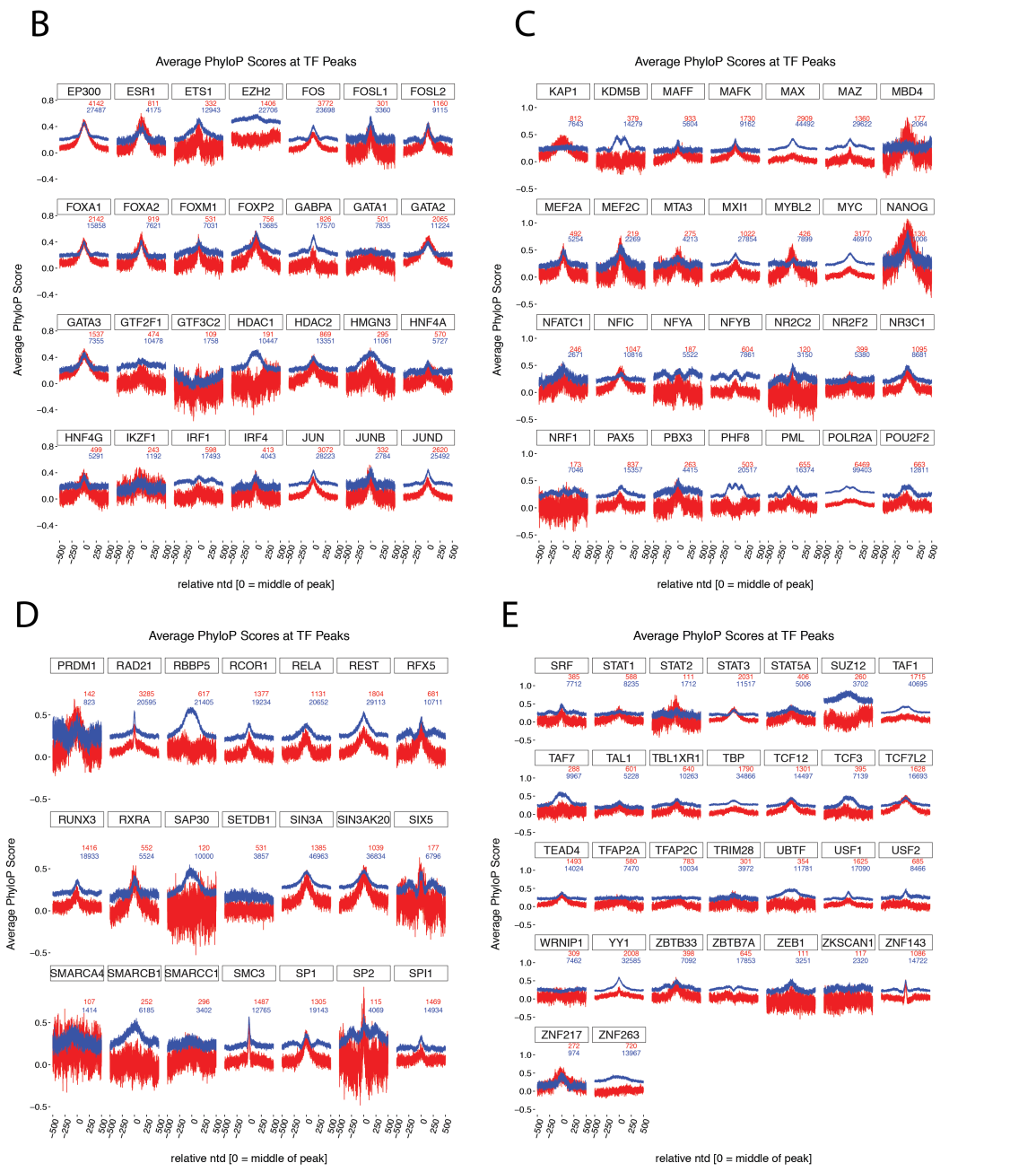


Supplemental Fig S3. Expression differences between expression matched lincRNAs and mRNAs expressed at higher than 0.1 FPKM comparing those with or without H3K9me3 in their promoter. Labels below each boxplot indicate the number of genes in that group. In the case of lincRNAs, only one cell line, K562, was significant (Wilcoxon test; FDR<0.05). In the case mRNAs, none of the cell lines were significant (Wilcoxon test; FDR<0.05).

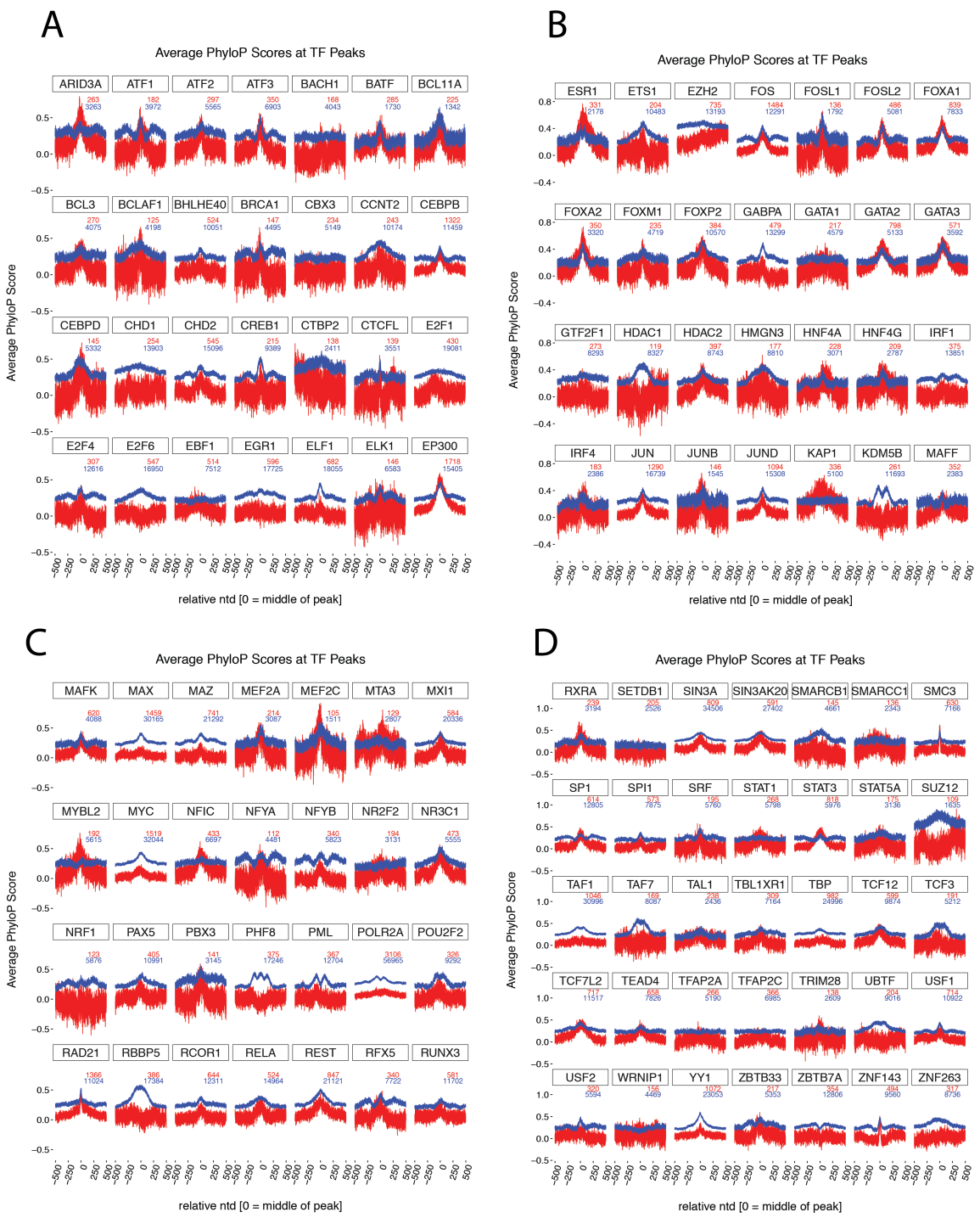


Supplemental Fig S4. Tissue specificity differences between expression matched lincRNAs and mRNAs expressed at higher than 0.1 FPKM comparing those with or without H3K9me3 in their promoter. Tissue specificity was calculated using an independent data set of RNA-seq for 20 human tissues. Labels below each boxplot indicate the number of genes in that group. In all cell lines except for one, lincRNAs with H3K9me3 in their promoter were significantly more tissue specific than those without (Wilcoxon test; FDR<0.05). Significant comparisons are indicated with an asterisk. None of the mRNA comparisons were significant.

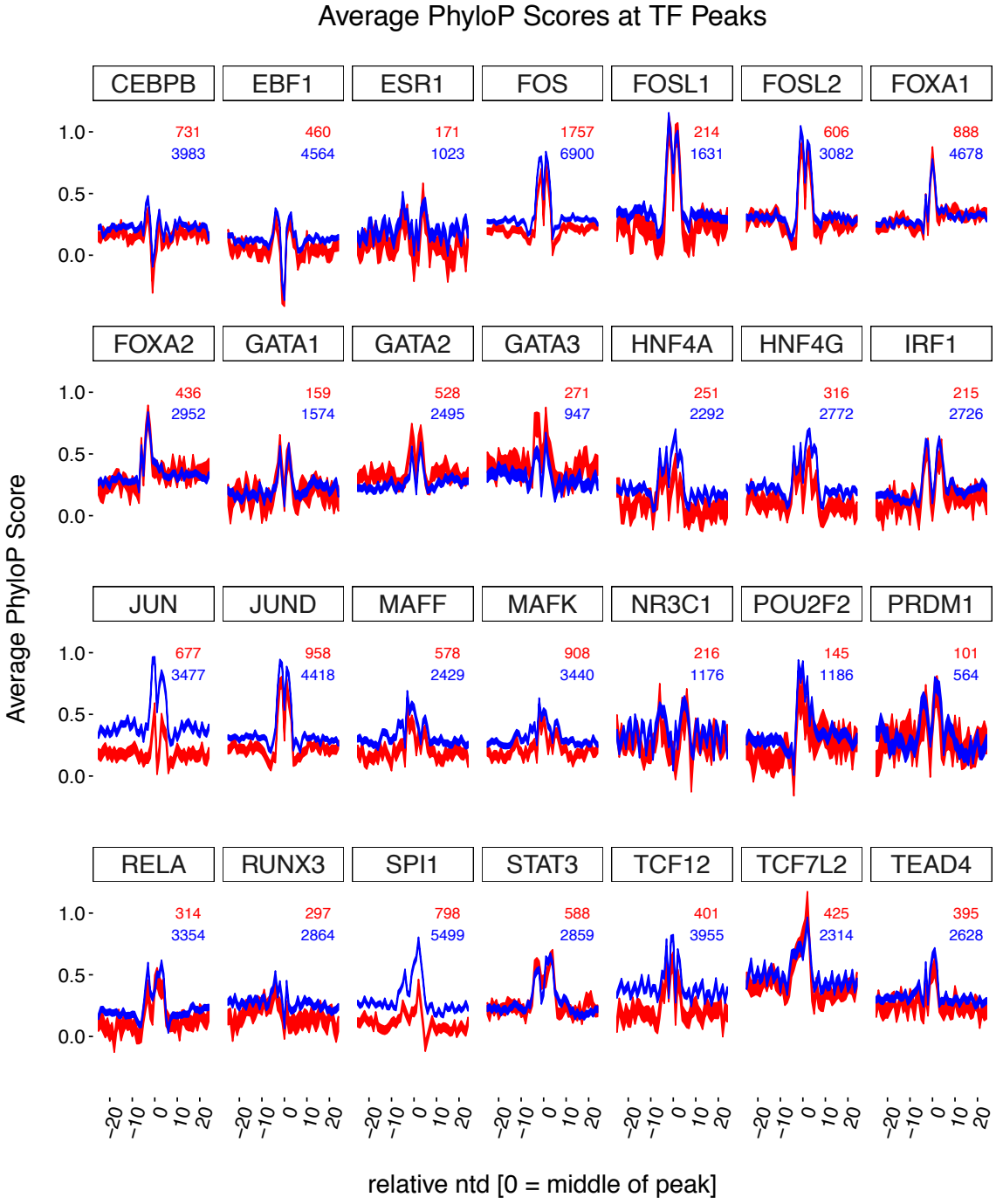

A
Average PhyloP Scores at TF Peaks
ARID3A
616
5947
ATF1
360
5521
ATF2
654
8398
ATF3
678
9810
BACH1
335
5673
BATF
715
4363
BCL11A
532
2874
BCL3
649
6764
BCLAF1
232
5684
BHLHE40
1128
15245
BRCA1
266
5647
CBX3
501
7517
CCNT2
436
13121
CEBPB
3392
24160
CEBPD
264
7042
CHD1
483
19703
CHD2
1018
20357
CREB1
359
11446
CTBP2
258
3720
CTCFL
272
5329
E2F1
655
23182
E2F4
499
15519
E2F6
899
21601
EBF1
1213
13650
EGR1
1184
25108
ELF1
1247
25385
ELK1
232
7949
ELK4
155
5545
0.8
0.4
0.0
−0.4
−500
−250
0
250
500
Average PhyloP Score
relative ntd [0 = middle of peak]

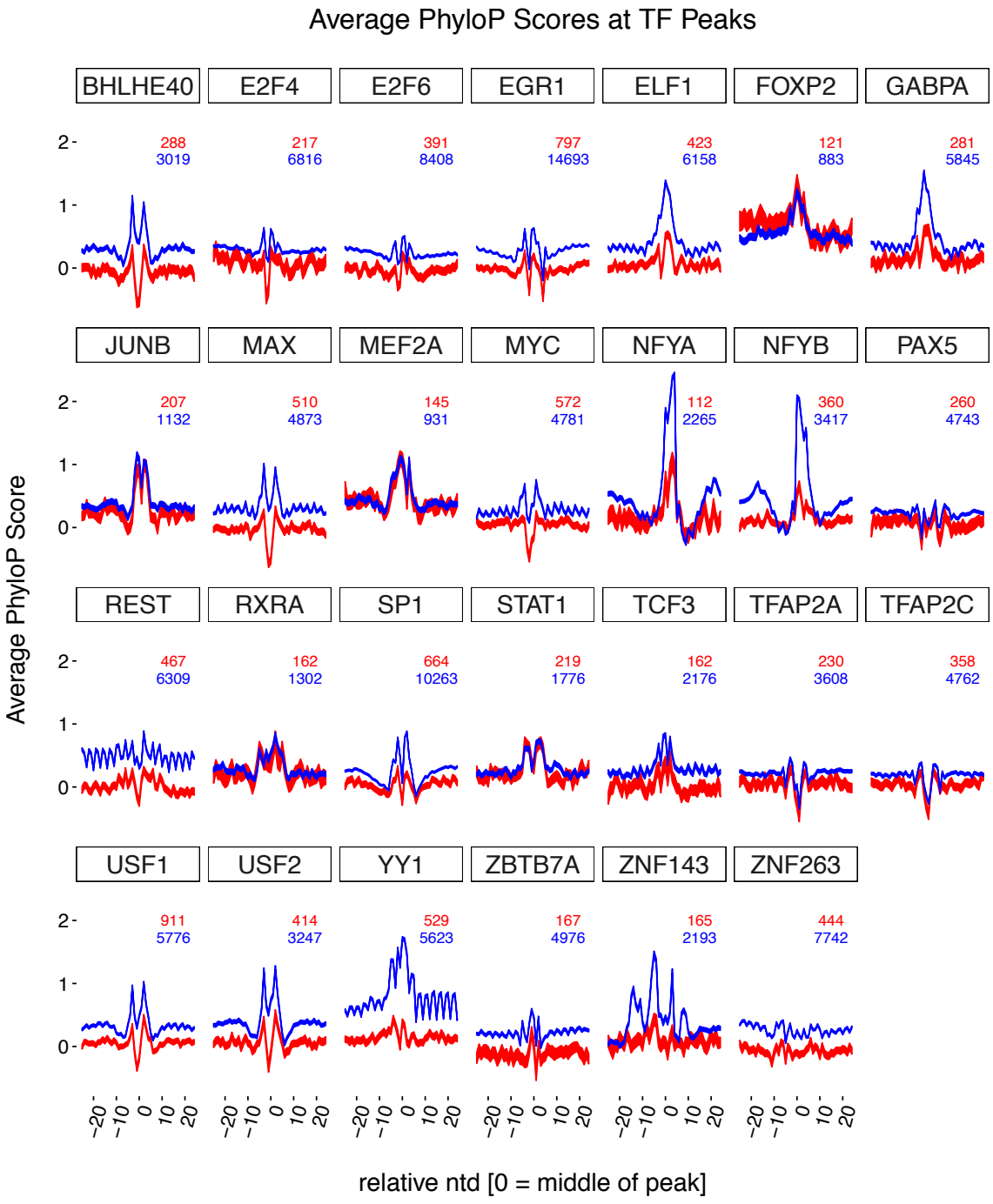


Supplemental Fig S5. Average conservation levels of transcription factor (TF) binding sites (TFBS) for all TFs intersecting lincRNA (red) and mRNA (blue) promoters (defined as minus and plus 5Kb from the TSS). We only plotted those TFBS that had more than 100 binding sites. For each TF, all ChIP-seq peaks intersecting a lincRNA or an mRNA promoter region where centered to the peak maxima and then we calculated average conservation score per nucleotide across sites. Extension of the vertical lines represents standard error. **A**. zoomed in version of the first 28 TF conservation plots. **B-E.** The remaining TF conservation plots grouped alphabetically in four groups. In general, mRNA peaks are larger than those for lincRNAs. However, a few TFs such as GATA2, MBD4, KAP1, ESR1 show slightly larger conservation in lincRNAs than in mRNAs.

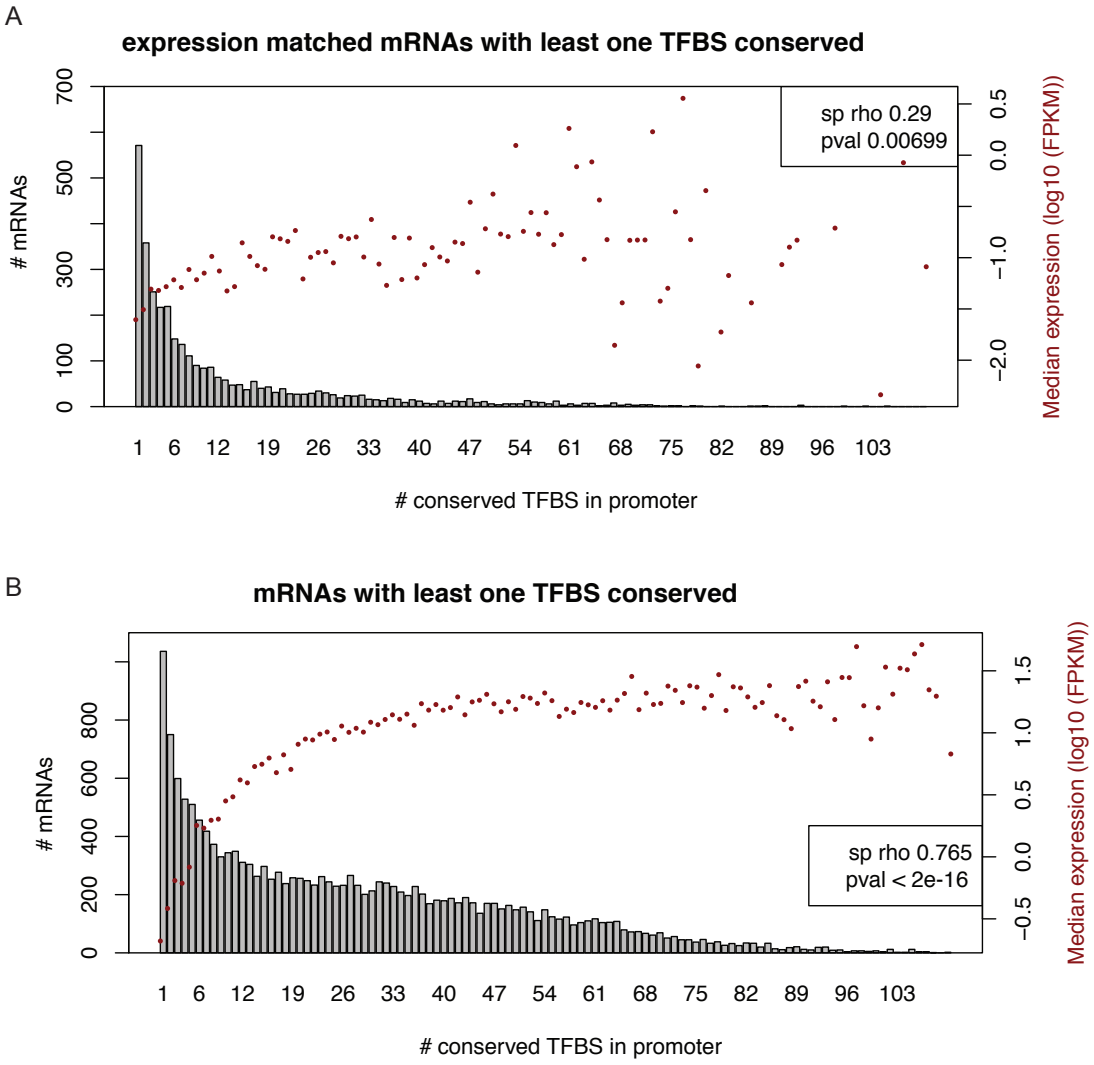


Supplemental Fig S6. Average conservation levels of transcription factor (TF) binding sites (TFBS) for all TFs intersecting lincRNA (red) and mRNA (blue) promoters (defined as minus 2Kb and plus 1Kb from the TSS) with more than 100 binding sites. For each TF, all ChIP-seq peaks intersecting a lincRNA or an mRNA promoter region where centered to the peak maxima and average conservation score per nucleotide was calculated across sites. Extension of the vertical lines represents standard error. In general, mRNA peaks are larger than those for lincRNAs. However, a few TFs such as GATA2 and KAP1 show slightly larger conservation in lincRNAs than in mRNAs.

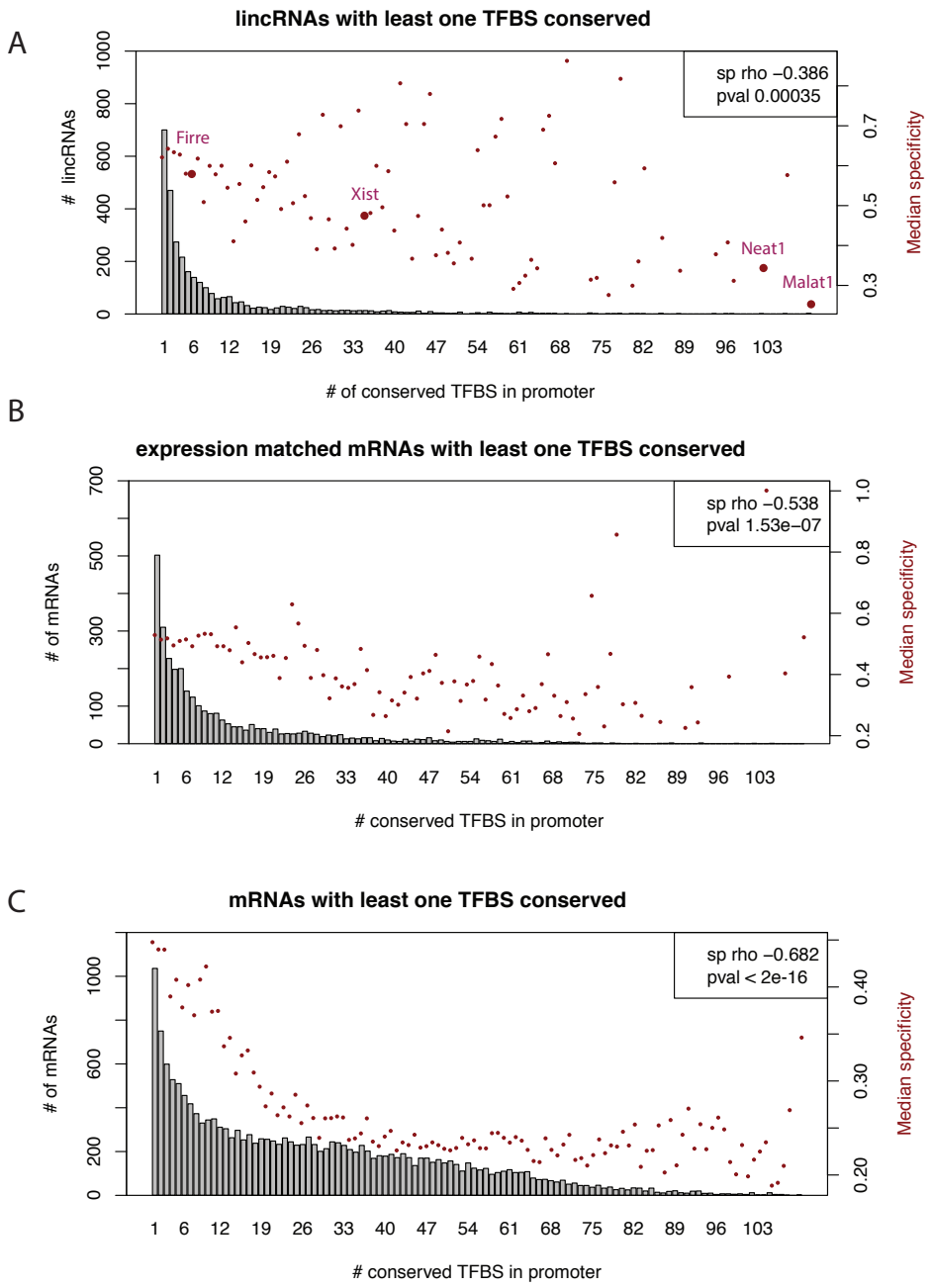

Average PhyloP Scores at TF Peaks
CEBPB
EBF1
ESR1
FOS
FOSL1
FOSL2
FOXA1
731
3983
460
4564
171
1023
1757
6900
214
1631
606
3082
888
4678
FOXA2
GATA1
GATA2
GATA3
HNF4A
HNF4G
IRF1
436
2952
159
1574
528
2495
271
947
251
2292
316
2772
215
2726
JUN
JUND
MAFF
MAFK
NR3C1
POU2F2
PRDM1
677
3477
958
4418
578
2429
908
3440
216
1176
145
1186
101
564
RELA
RUNX3
SPI1
STAT3
TCF12
TCF7L2
TEAD4
314
3354
297
2864
798
5499
588
2859
401
3955
425
2314
395
2628
1.0
0.5
0.0
-20
-10
0
10
20
Average PhyloP Score
relative ntd [0 = middle of peak]



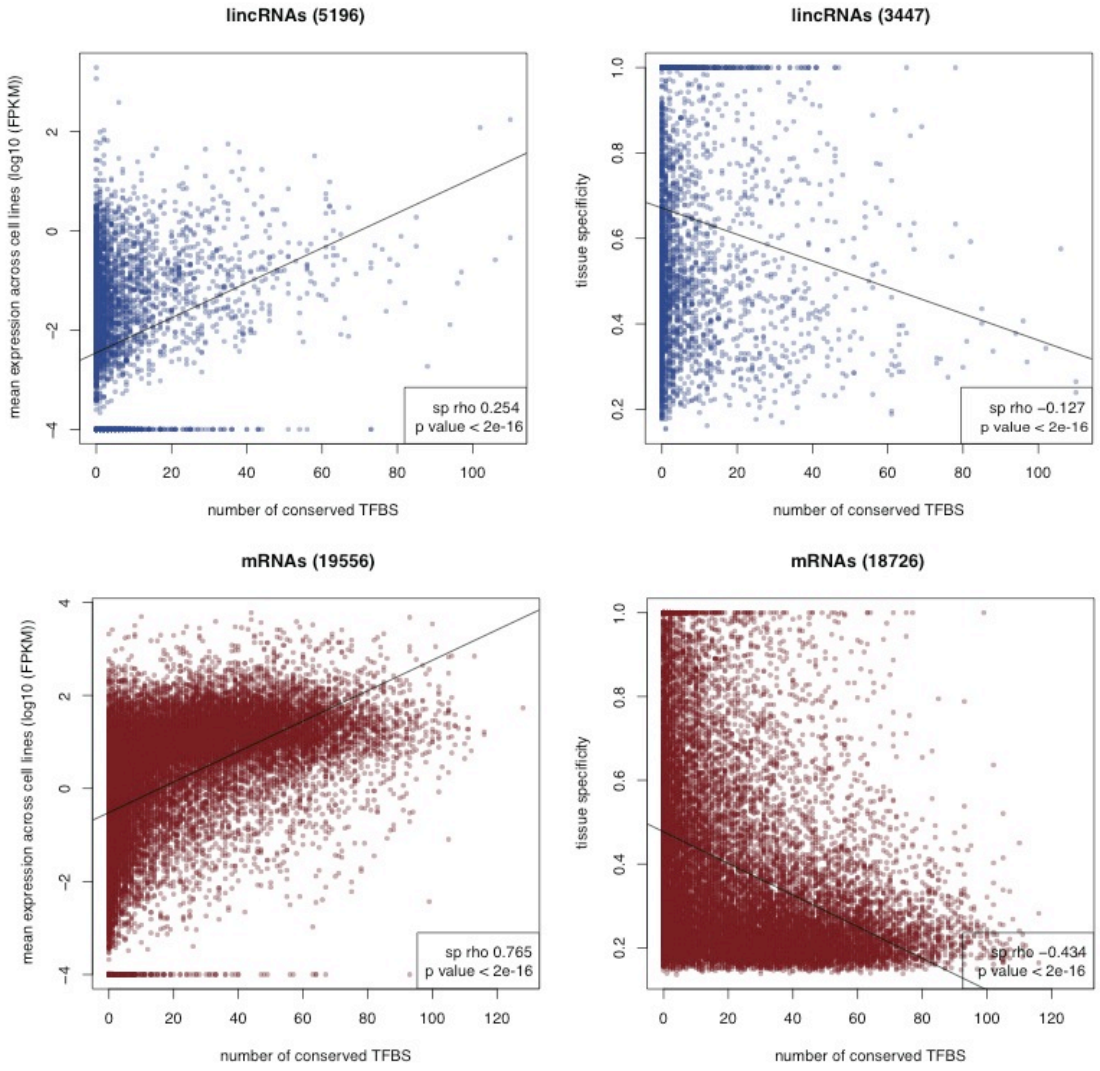
Supplemental Fig S7. Average conservation levels of transcription factor (TF) binding sites (TFBS) for TFs with known motifs that intersect lincRNA (red) and mRNA (blue) promoters. For each TF, we mapped its known motif to all ChIP-seq peaks intersecting a lincRNA or an mRNA promoter region and calculated average conservation score per nucleotide centered on the mapped motifs. We show TFs for which the number of intersections was higher than one hundred. Extension of the vertical lines represents standard error. GATA2 and GATA3 show larger conservation in lincRNAs than in mRNAs.



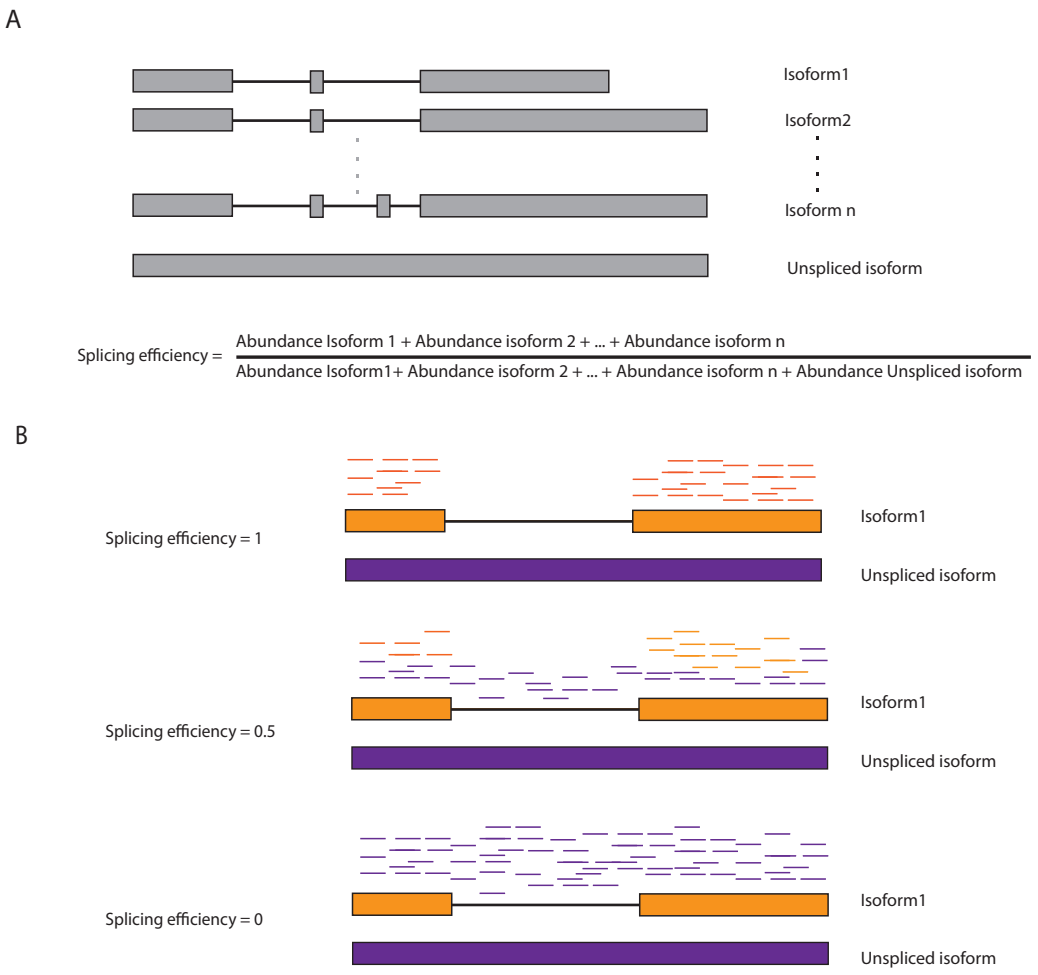
Supplemental Fig S8. Number of mRNAs with certain number of TFBS in their promoter and their median expression values Bars represent the number of mRNAs that have a specific number of TFBS conserved in their promoter. Dots represent the median value of expression for all mRNAs with that number of TFBS in their promoter. The mRNAs in **A.** had similar average expression levels than the lincRNAs average expression calculated across 7 cell lines). In **B,** all mRNAs are represented.



Supplemental Fig S9. Number of lincRNAs (A) expression matched mRNAs (B) and all mRNAS (C) with certain number of TFBS in their promoter and their median tissue specificity values. Bars represent the number of lincRNAs or mRNAS that have a specific number of TFBS conserved in their promoter. Dots represent the median value of tissue specificity for all lincRNAs or mRNAs with that number of TFBS in their promoter. In A. Presence of certain functional lincRNAs within each group is highlighted.

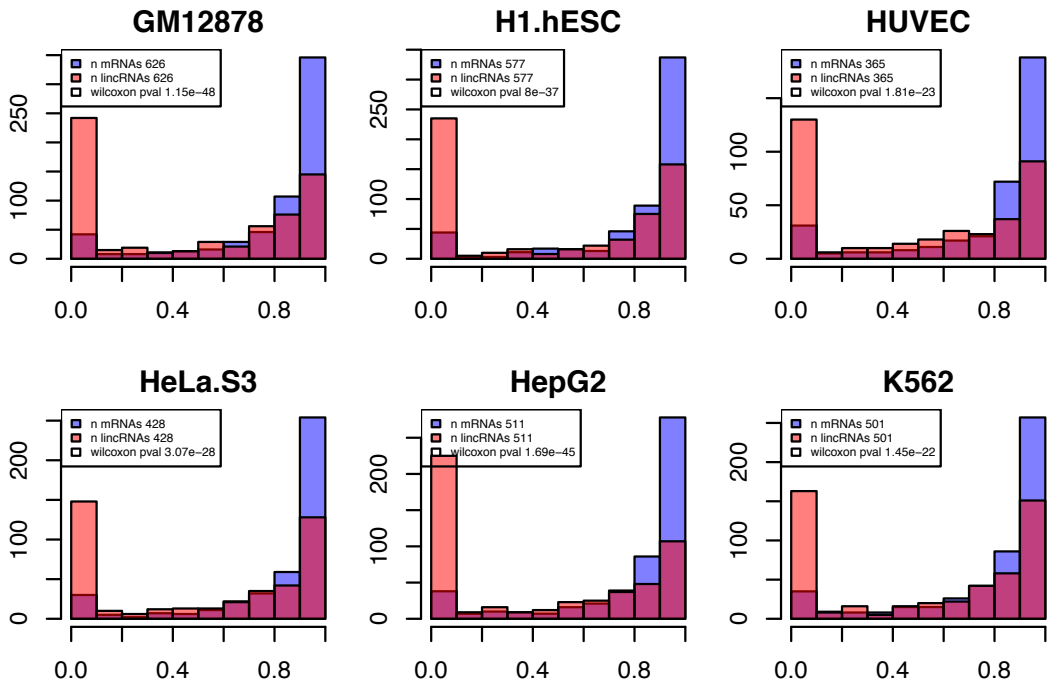


Supplemental Fig S10. Correlation between the number of different TFBS that are conserved and mean expression levels (left) or tissue specificity (right) for lincRNAs (top) or mRNAs (bottom). Numbers of genes analyzed in the tissue specificity analysis are lower than for expression because we only calculated tissue specificity for those genes for which we had expression values across all tissues within the HBM dataset plus four other tissues.

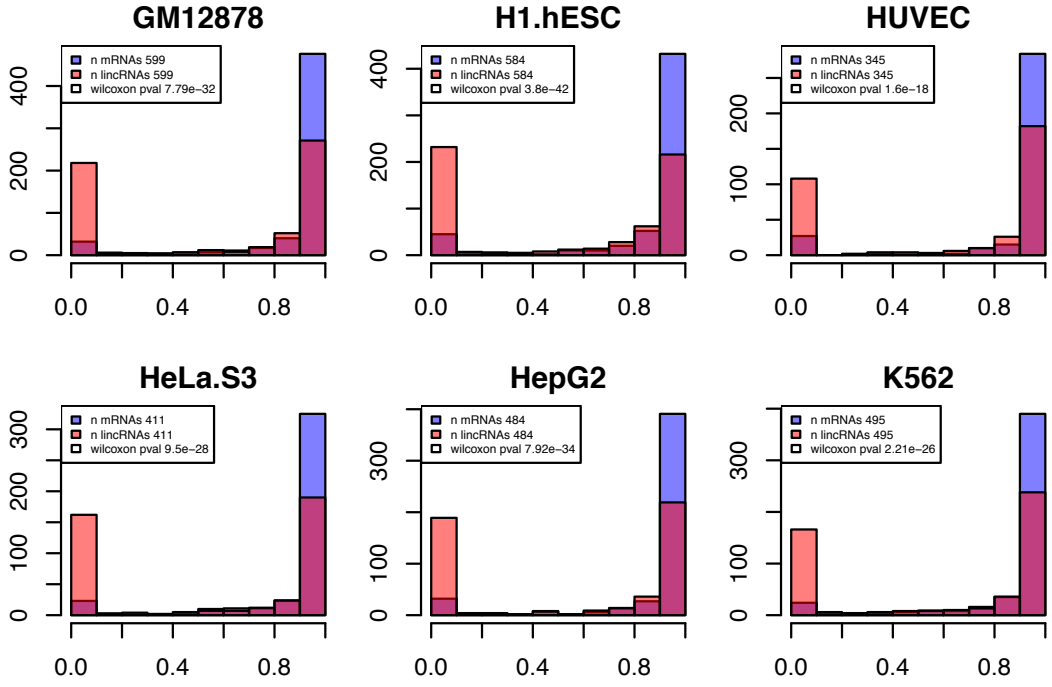


Supplemental Fig S11. Splicing efficiency calculation. A. Scheme of how splicing efficiency is calculated and **B.** examples of high and low splicing efficiency. The color of the reads represents to which isoform those reads will be assigned to and then used to calculate abundance. For a given gene, if all reads were assigned to spliced isoforms, that gene's splicing efficiency would be one whereas if all reads were assigned to the unspliced isoform (or intronic regions), its splicing efficiency would be zero (Supplemental Fig S11B). We then used Wilcoxon's effect size to obtain a quantitative measure of the strength of such differences between lincRNAs and mRNAs.

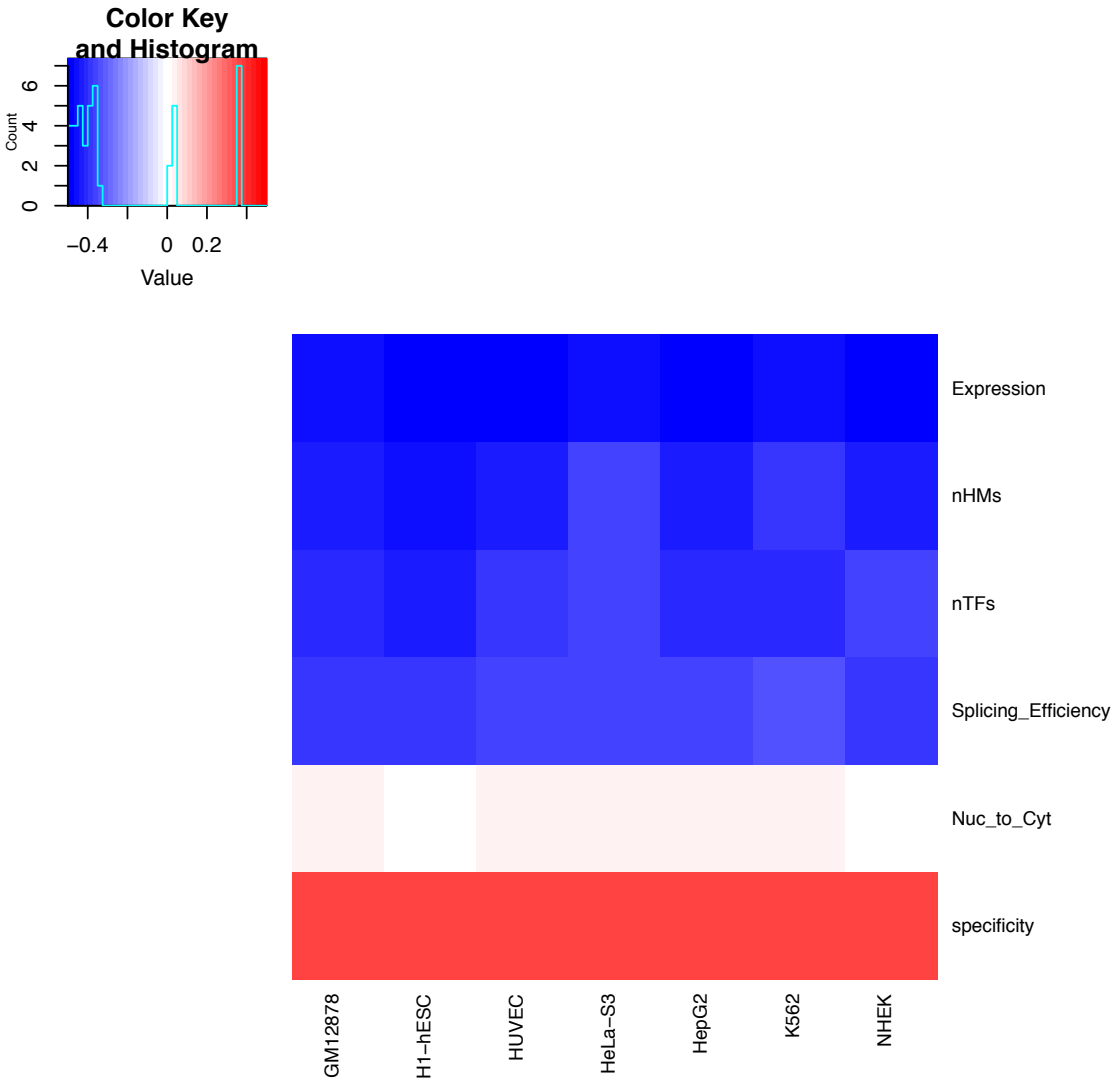
A



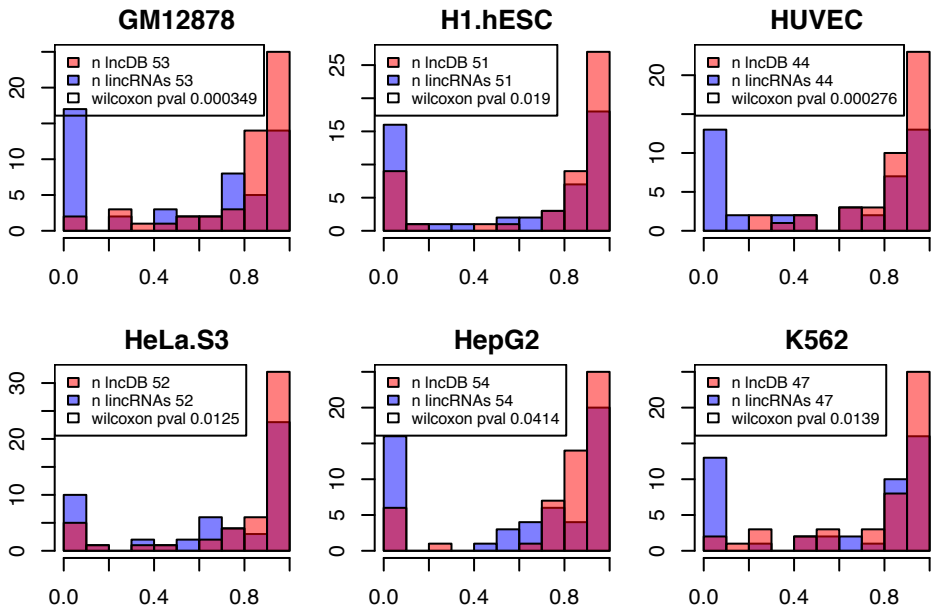
B



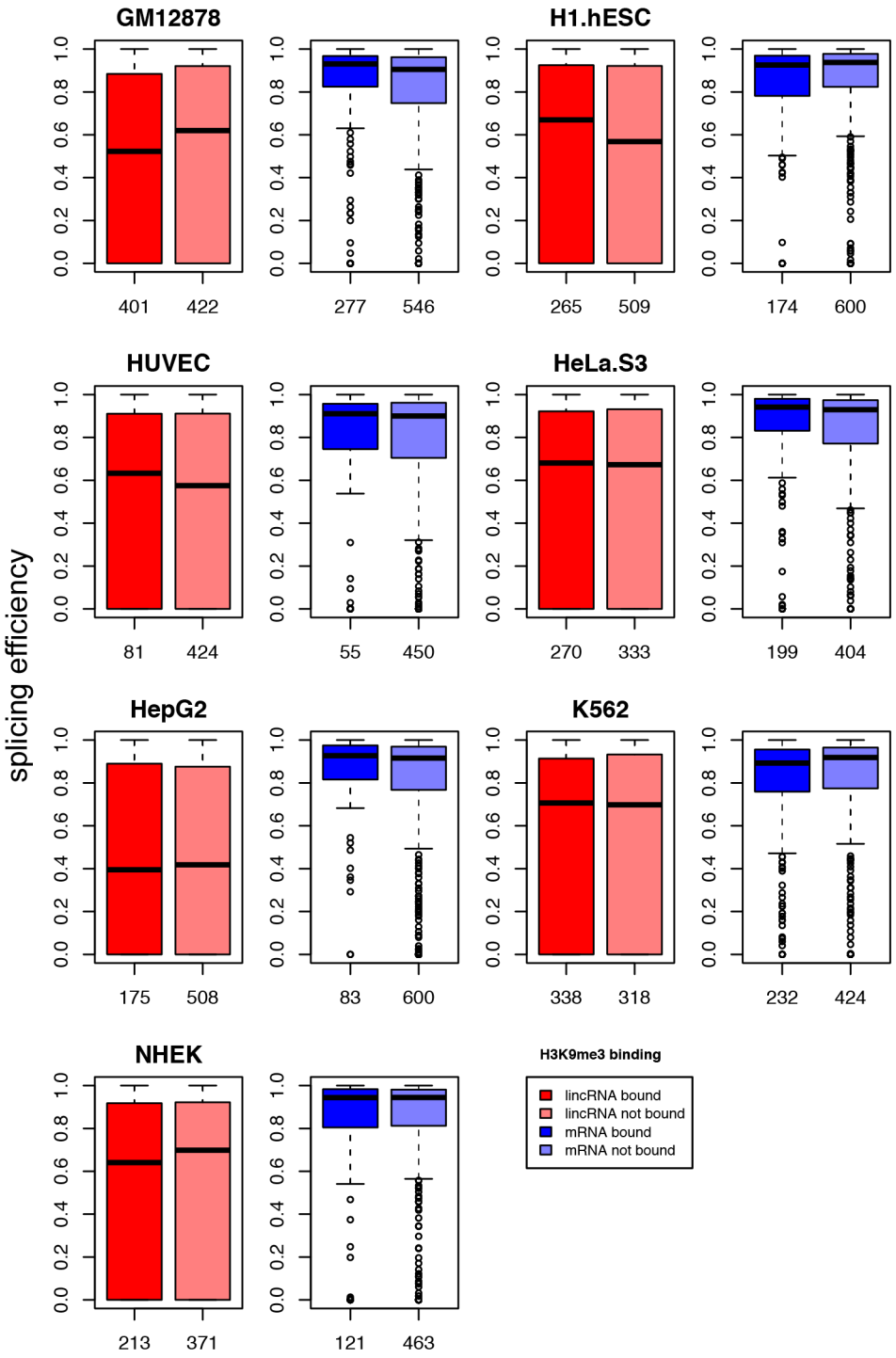
Supplemental Fig S12. Splicing efficiency histogram in lincRNAs and expression matched mRNAs across ENCODE cell lines in **A.** nuclear fraction and **B.** cytosolic cellular fraction. Only genes expressed at higher than 0.1 FPKM at the analyzed cell line were selected. All cells are significant after multiple test correction (FDR<0.05).



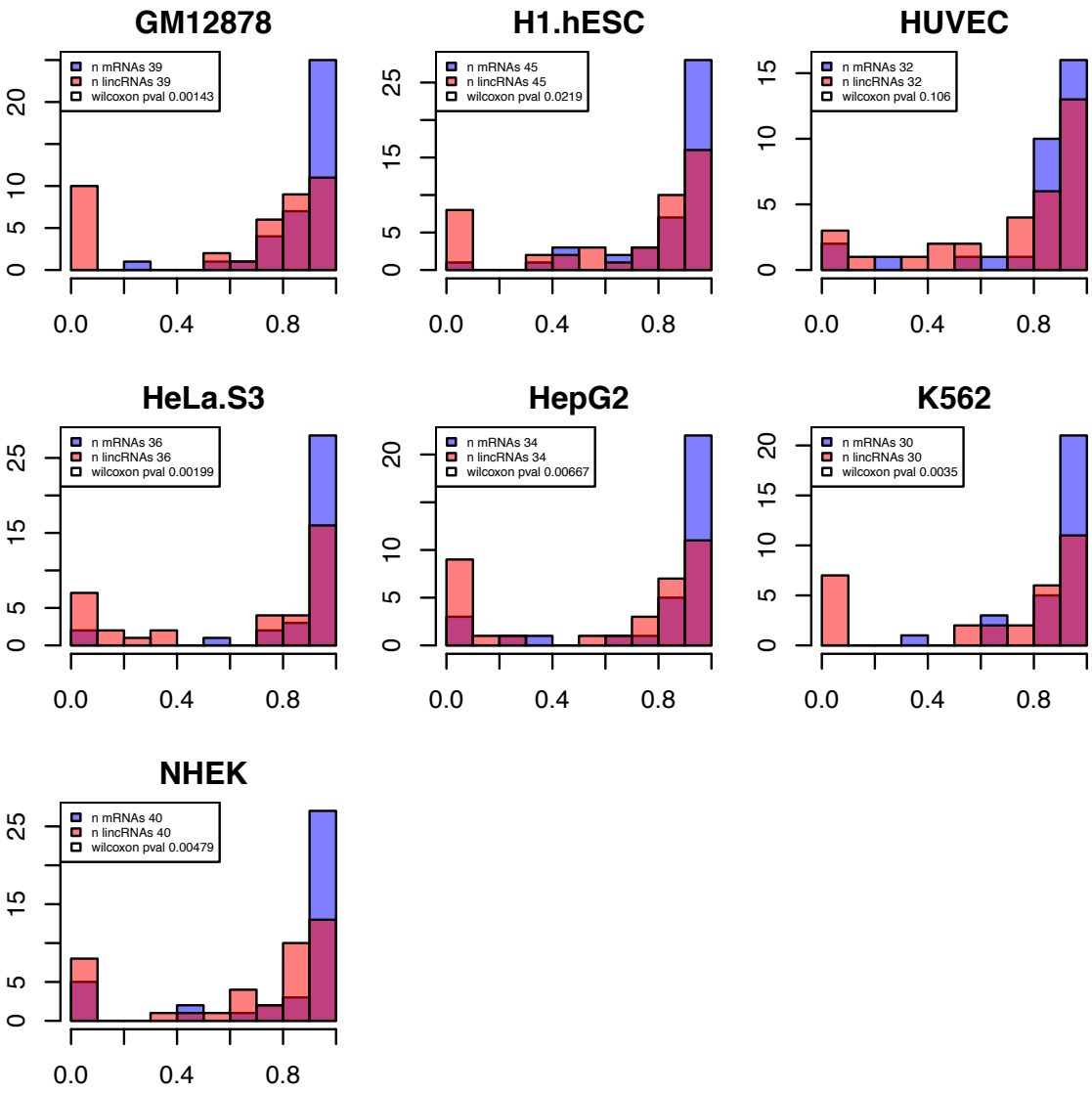
Supplemental Fig S13. Effect size differences between lincRNAs and mRNAs across several gene/promoter properties. For each cell line, we analyzed all promoters of lincRNAs and mRNAs. Blue corresponds to larger values in mRNAs and red to larger values in lincRNAs. Effect sizes are in general larger when comparing lincRNAs and mRNAs if we did not correct for expression levels.



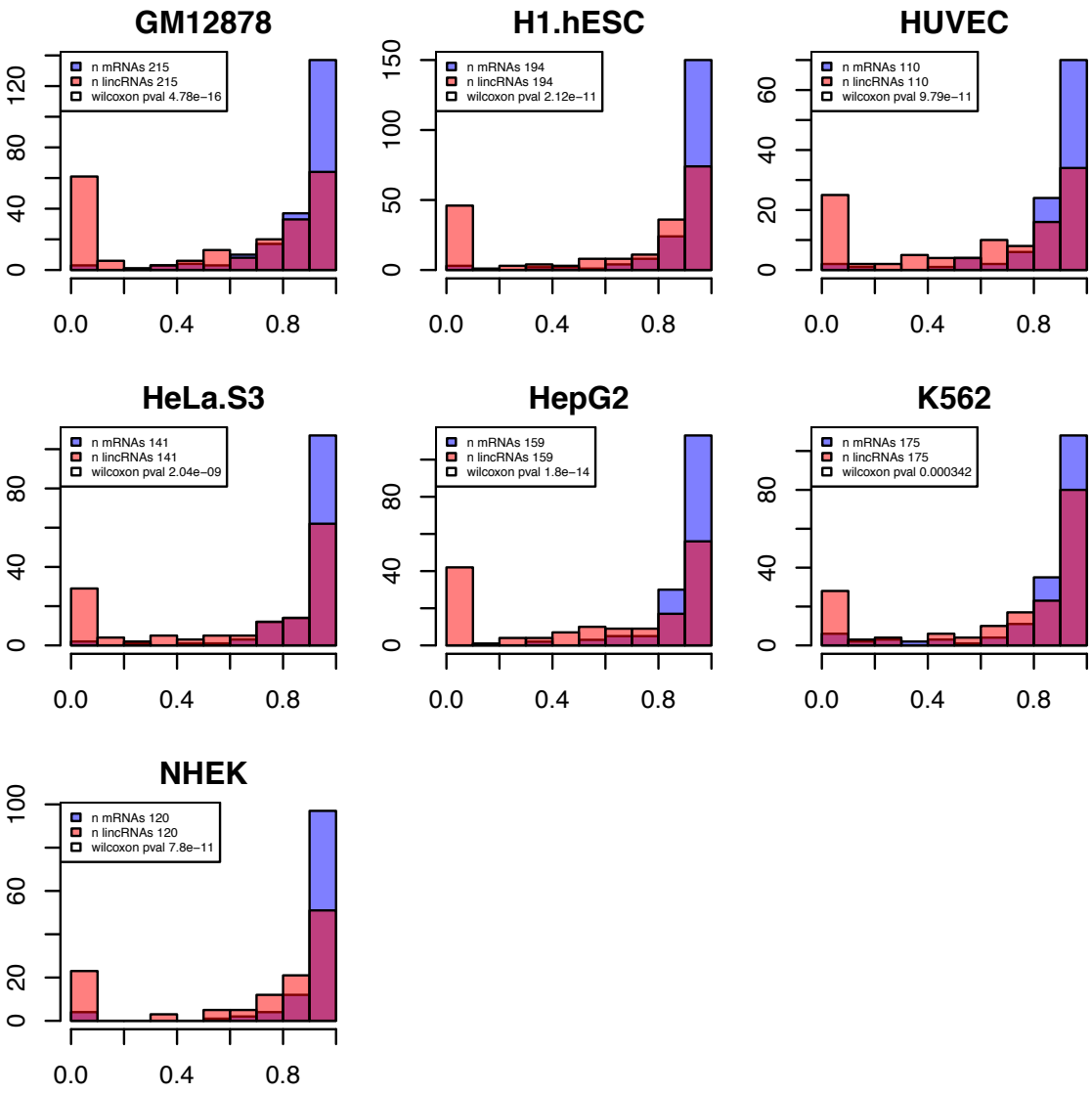
Supplemental Fig S14. Splicing efficiency histogram in lncRNAs annotated in lncRNA DB and expression matched lincRNAs. Only genes expressed at higher than 0.1 FPKM at the analyzed cell line were selected. All cells are significant after multiple test correction (FDR<0.05).



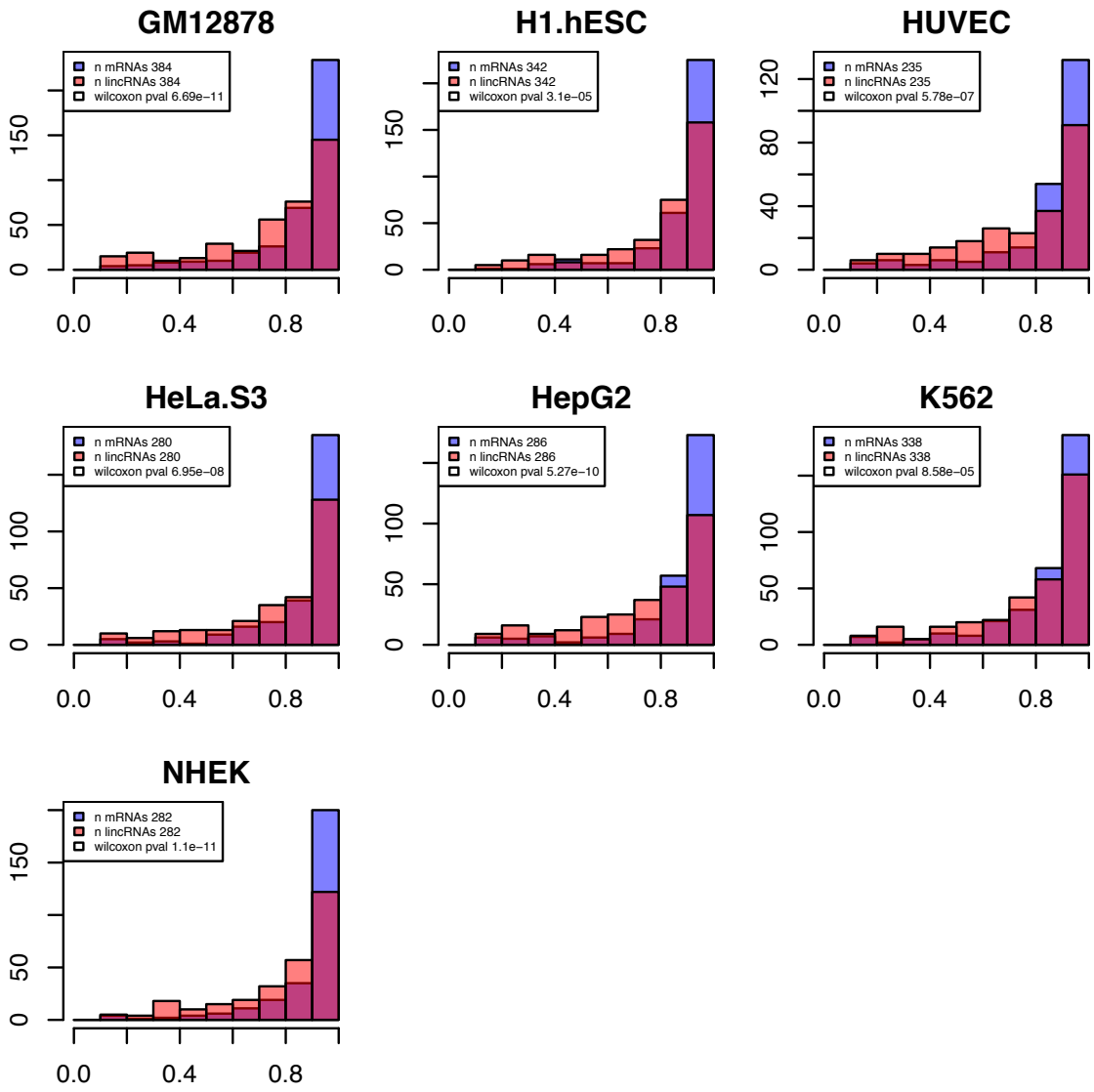
Supplemental Fig S15. Splicing efficiency differences between expression matched lincRNAs and mRNAs expressed at higher than 0.1 FPKM comparing those with or without H3K9me3 in their promoter. None of the comparisons was significant (Wilcoxon p-value > 0.05).



Supplemental Fig S16. Splicing efficiency histogram in lncRNAs annotated as "known" in GENCODE and expression matched mRNAs. Only genes expressed at higher than 0.1 FPKM at the analyzed cell line were selected. All cells are significant except for one after multiple test correction (FDR<0.05).

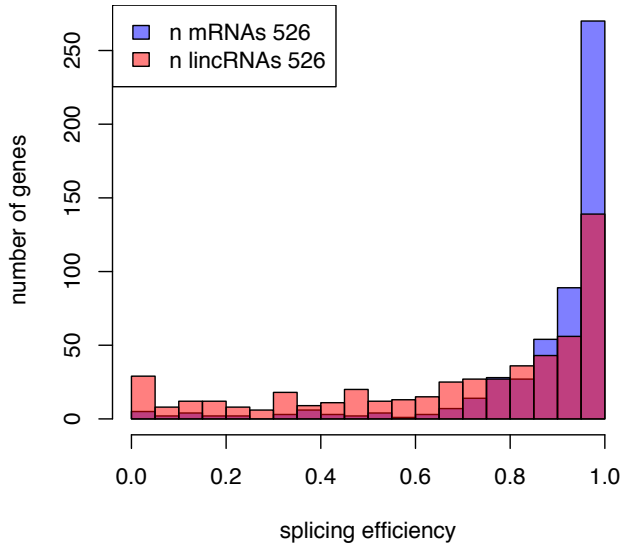


Supplemental Fig S17. Splicing efficiency histogram in lincRNAs expressed at higher than 1FPKM in the tested cell line and expression matched mRNAs. All cells are significant after multiple test correction (FDR<0.05). Only genes expressed at higher than 0.1 FPKM at the analyzed cell line were selected.

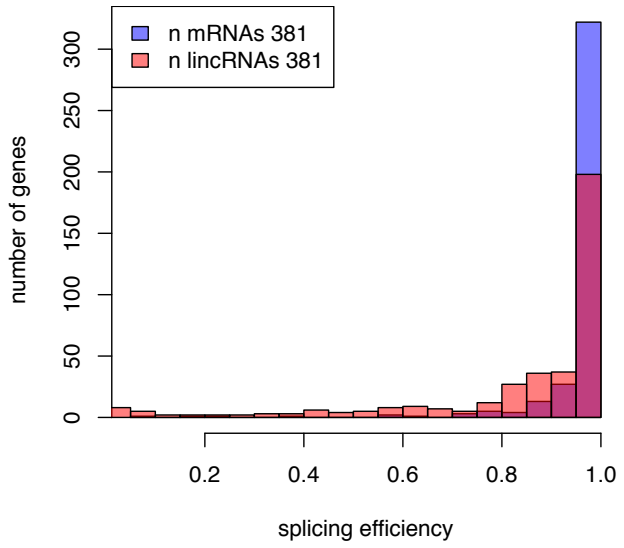


Supplemental Fig S18. Splicing efficiency histogram in lincRNAs excluding those lincRNAs with very low splicing efficiency (<0.1). All cells are significant after multiple test correction (FDR<0.05). Only genes expressed at higher than 0.1 FPKM at the analyzed cell line were selected.

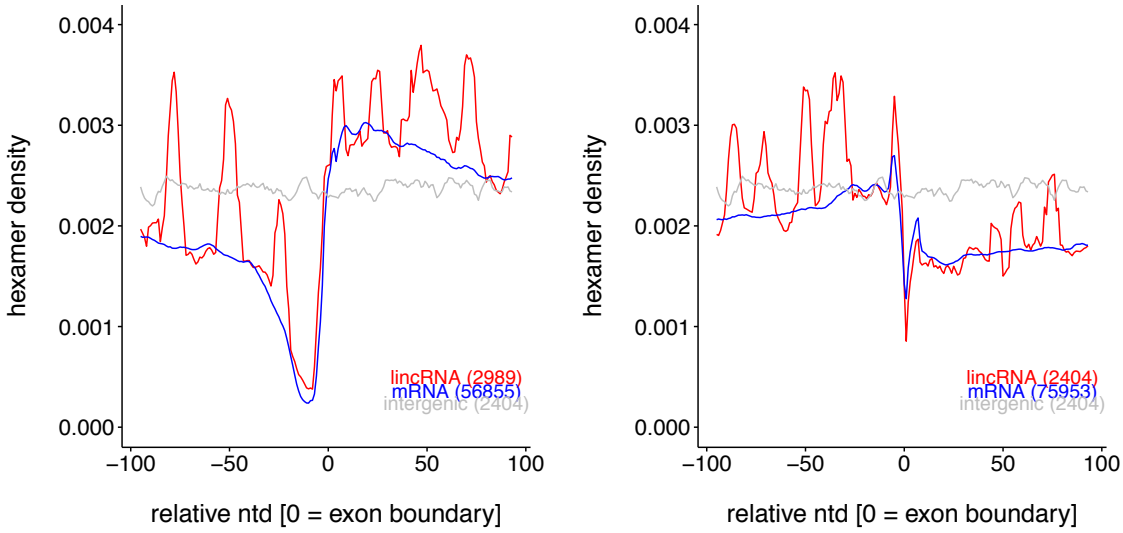
A



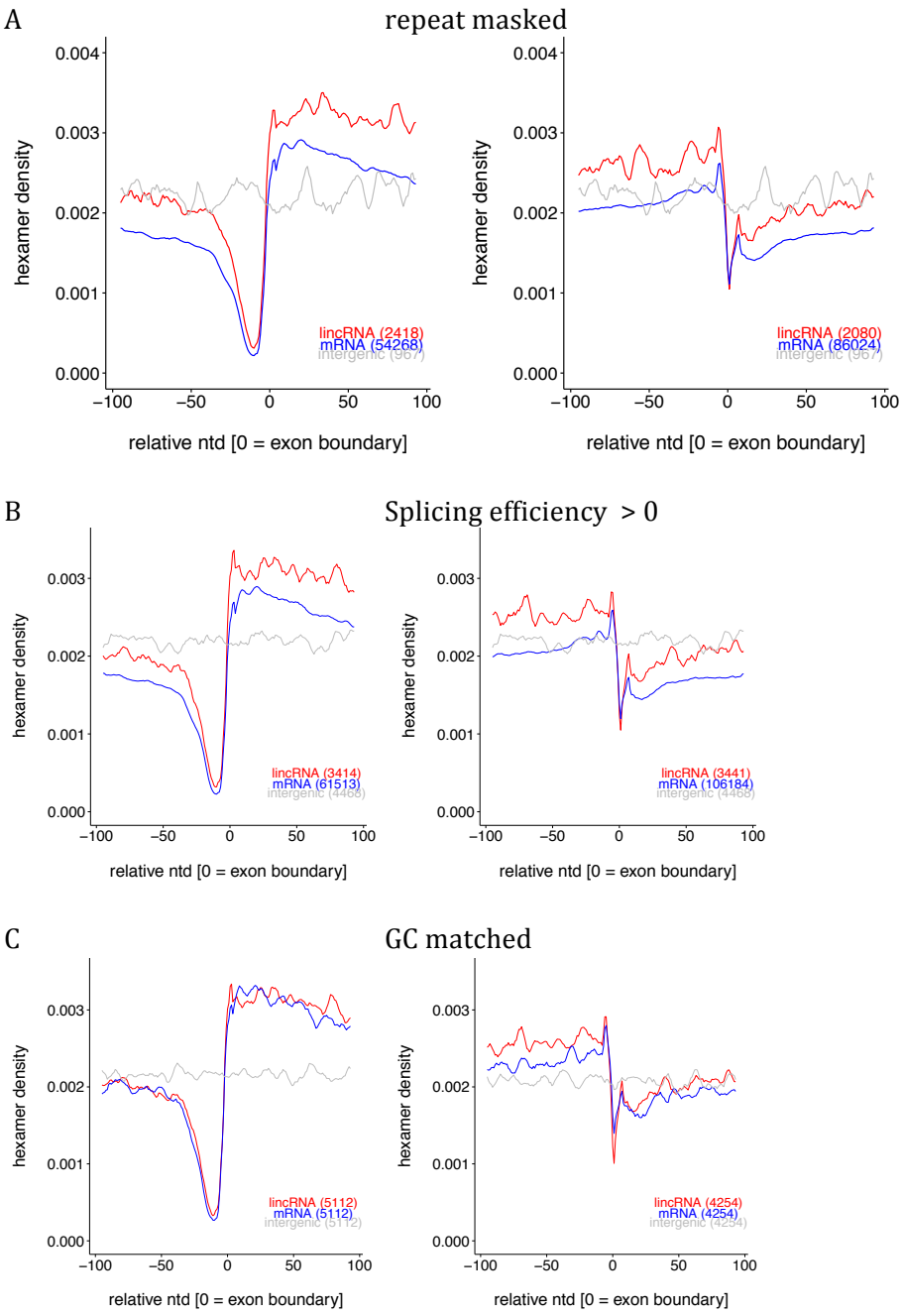
B



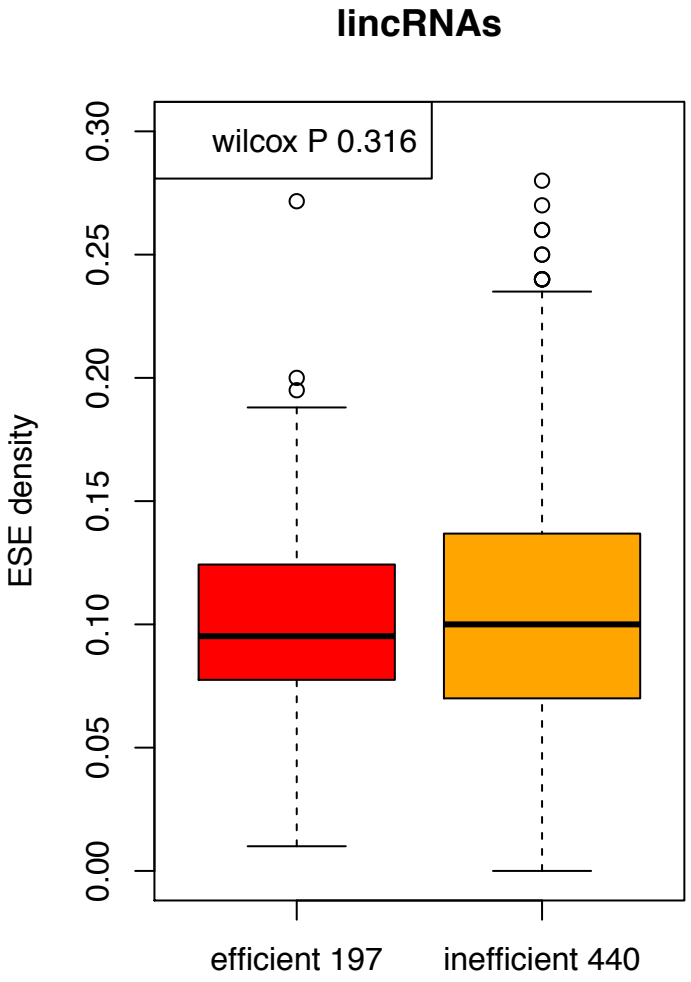
Supplemental Fig S19. Splicing efficiency in mouse ES cells in **A.** the nuclear fraction and **B.** cytosolic fraction. Only genes expressed at higher than 0.1 FPKM at the analyzed fraction were selected. Wilcoxon P < 2e-16 in both cases.



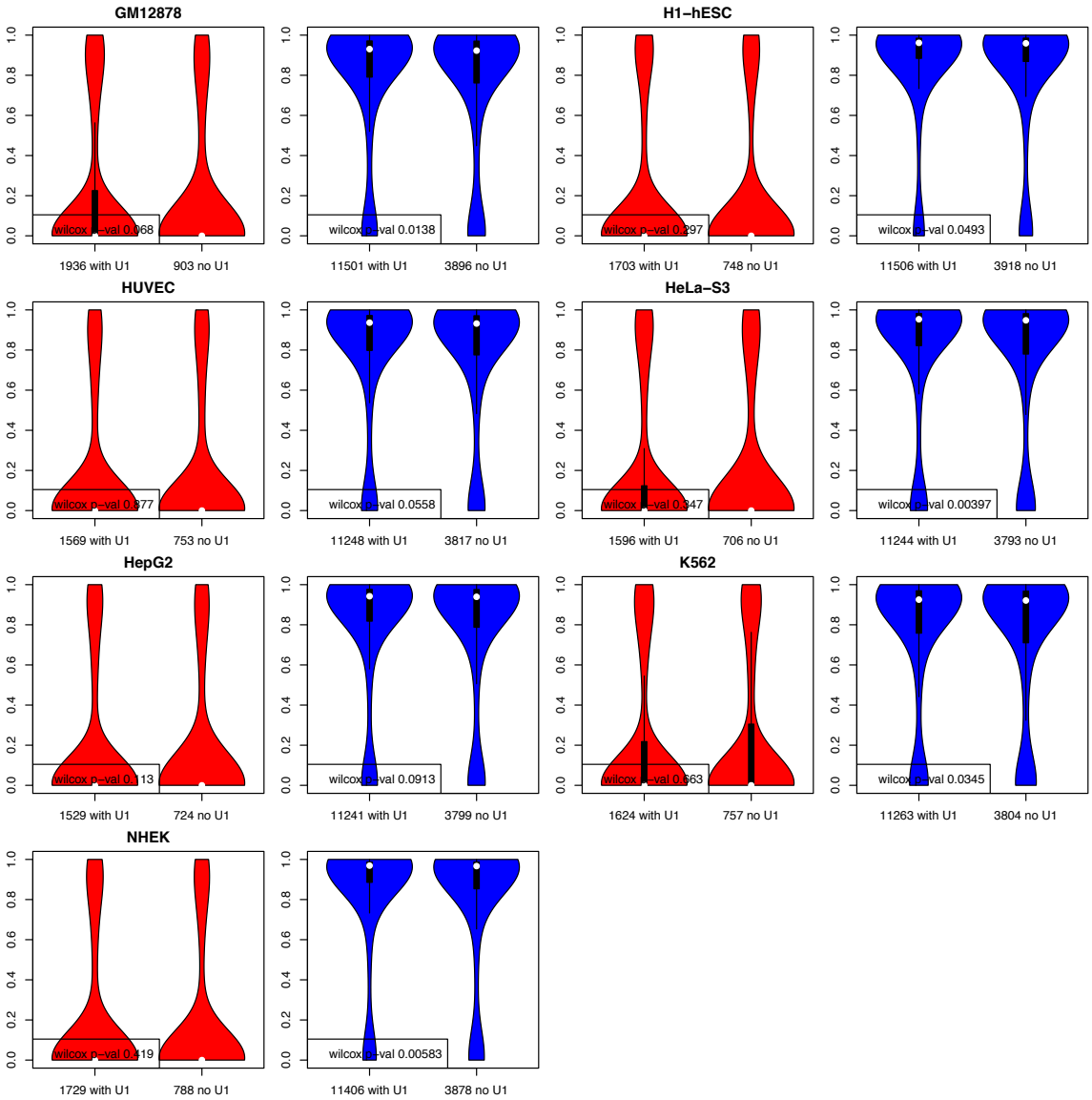
Supplemental Fig S20. Mean exonic splicing enhancer (ESE) density in mice intron exon junctions in all lincRNA and mRNA annotated exons larger than 200bp compared to random intergenic regions of the same size.



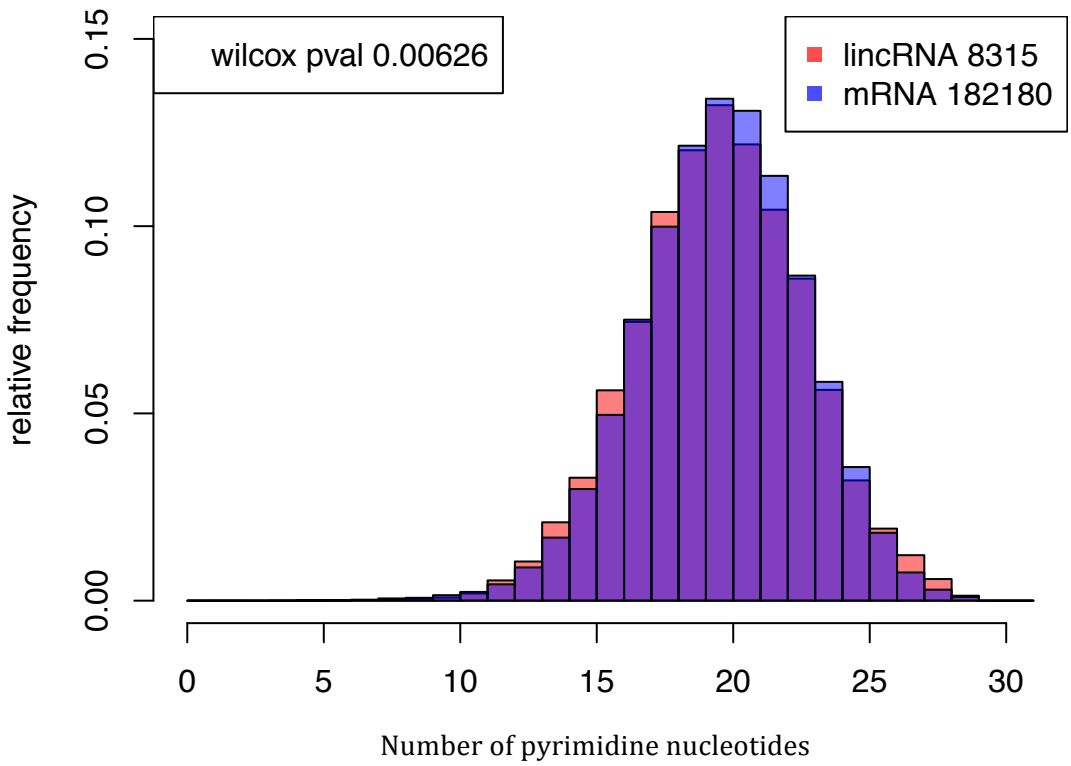
Supplemental Fig S21. Mean exonic splicing enhancer (ESE) density in human 3′ (left) and 5′ (right) splice sites of lincRNA and mRNA compared to random intergenic regions of the same length. Number of splice sites analyzed for each category is indicated in parenthesis. (A) Analysis after masking all repetitive elements (B) selecting genes with splicing efficiency larger than zero in at least one cell line (C) across splice sites in lincRNAs or mRNAs or random intergenic regions with similar GC content.



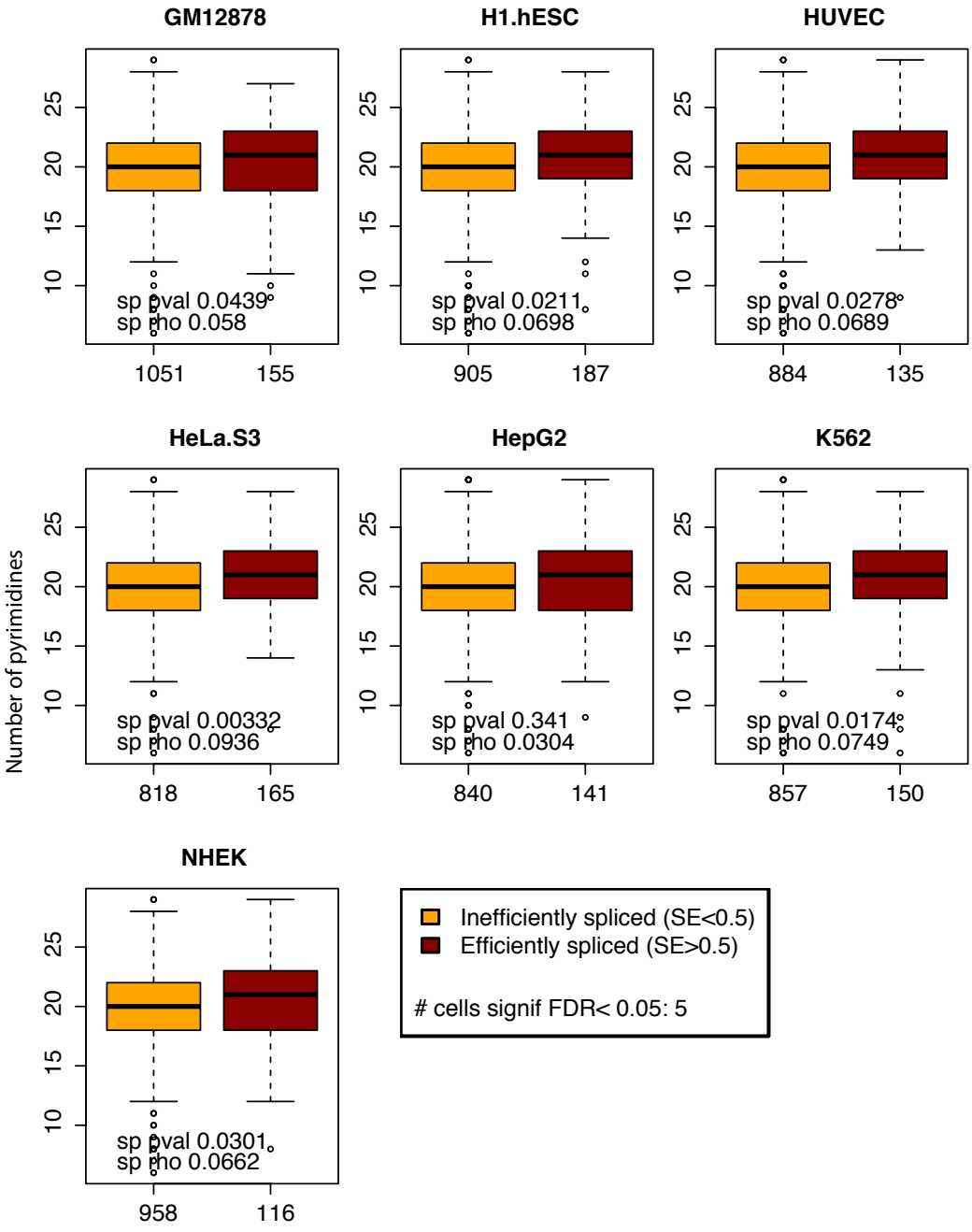
Supplemental Fig S22. ESE density distribution in lincRNAs that are efficiently and inefficiently spliced. For this analysis we selected lincRNAs that had splicing efficiency larger than zero in at least one cell line.



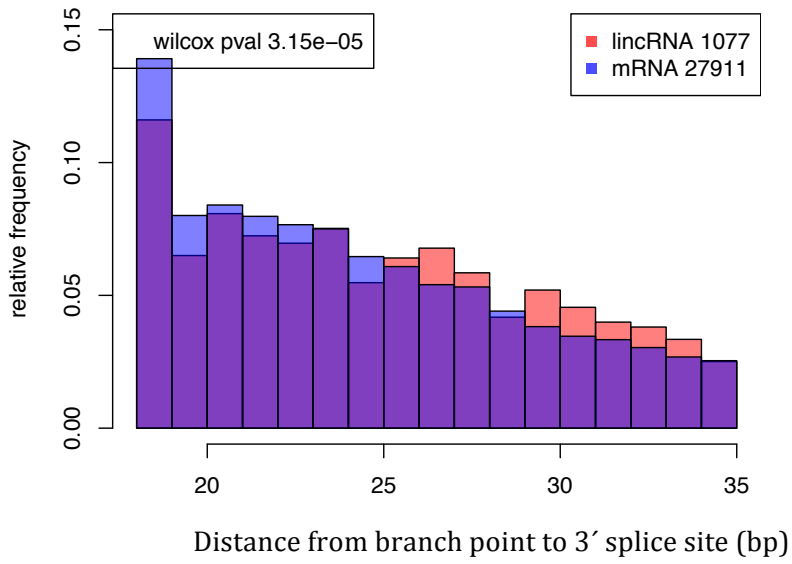
Supplemental Fig S23. Distribution of splicing efficiency (y-axis) across cell lines in lincRNAs (red) and mRNAs (blue), both with and without canonical U1 motifs at 5´ splice sites within the first gene locus Kb. After multiple test correction, only 3 out of the 7 cell lines remained significant in lincRNAs (FDR<0.05).



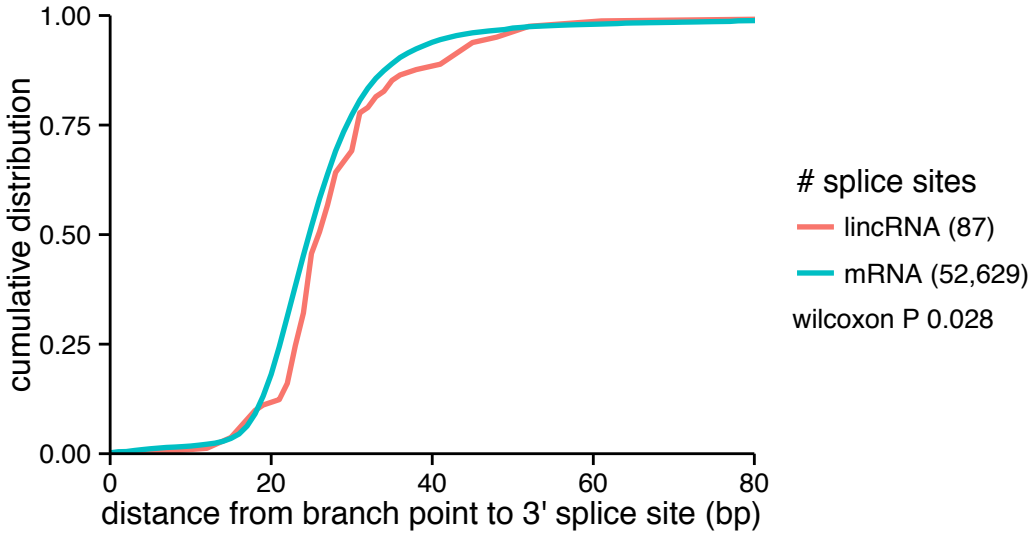
Supplemental Fig S24. Number of pyrimindine nucleotides in lincRNAs and mRNAs at the 3′ splice site upstream region when excluding lincRNAs with splicing efficiency equal to zero.



Supplemental Fig S25. Correlation between splicing efficiency and number of pyrimidines upstream of the 3´splice site region. For visual purposes we show boxplots of the distribution of the number of pyrimidine within the first 30 bp upstream of the 3´ splice site (y axis) in efficiently spliced genes versus inefficiently spliced genes (x axis). Given that splicing efficiency is a per gene measurement, we only used genes with a unique isoform and two exons to be able to have a single PPT measure per gene. Spearman rho correlation and p-values are indicated at the bottom of each plot. The number of genes within each group is indicated below the plots.

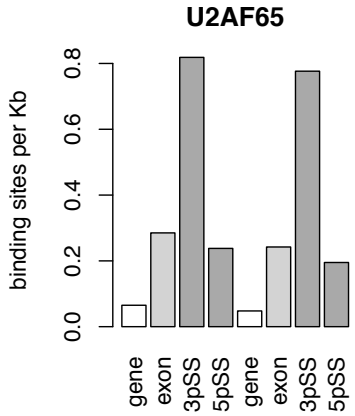


Supplemental Fig S26. Relative frequency of the Distance between 3´ splice site and the closest canonical branch point motif in lincRNAs and mRNAs (in bp). In this analysis we excluded those lincRNAs that had zero splicing efficiency across all cell lines for which splicing efficiency could be calculated.

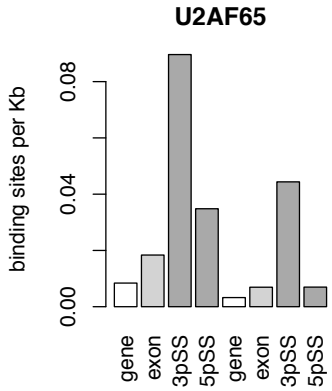


Supplemental Fig S27. Distance (bp) from experimentally determined branch points to the closest acceptor site in lincRNAs and mRNAs. The number of 3´ splice sites analyzed is indicated in parenthesis.

A

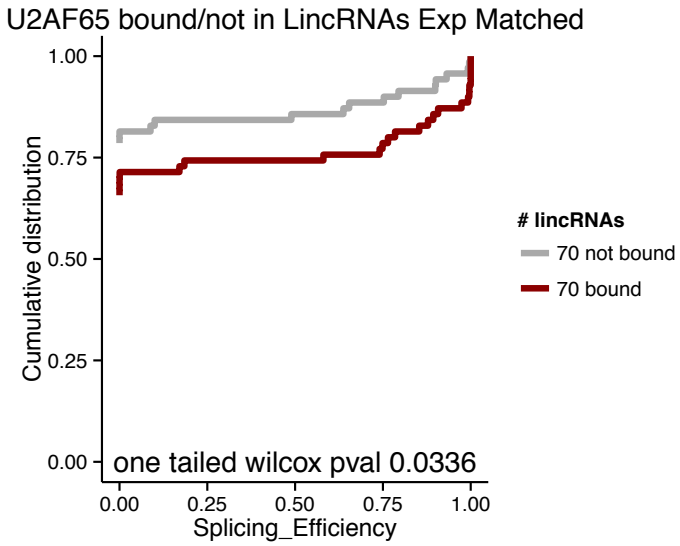


B

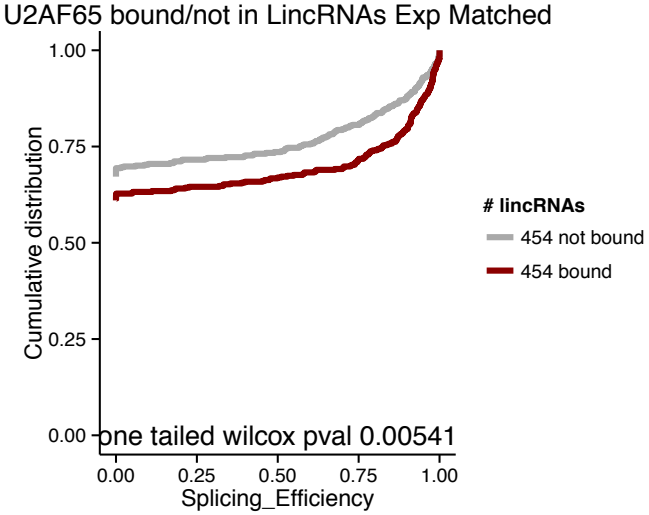


Supplemental Fig S28. Peak density for U2AF65 from two independent CLiP-seq experiment's in lincRNAS (left) and expression –matched mRNAs (right). A from Zarnack et al. 2013 . B from Shao et al. 2014. U2af65 has larger density in 3 ' splice sites. LincRNAs and mRNAs have similar binding sites per Kb. In A. peak density is larger than in B. because the total number of peaks detected in the first dataset was also larger.

A

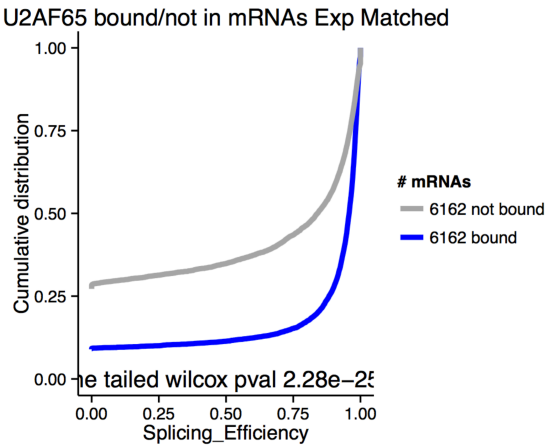


B

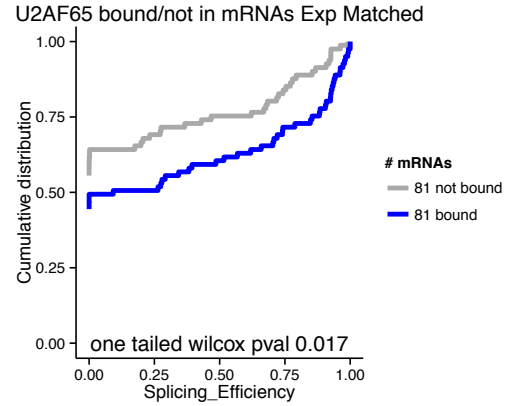


Supplemental Fig S29. Splicing efficiency of lincRNAs with U2af65 bound or not bound with similar expression levels using Zarnack et al. 2013 dataset. LincRNAs between bound/unbound groups are expression matched using abundance values from the same cell line where the CLIP experiment was carried out. **A.** Only lincRNAs with a single isoform of two exons were analyzed. **B.** LincRNAs with splicing efficiency of zero were excluded from the analysis.

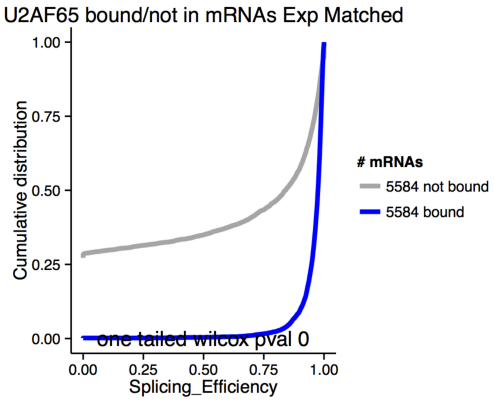
A.



B.

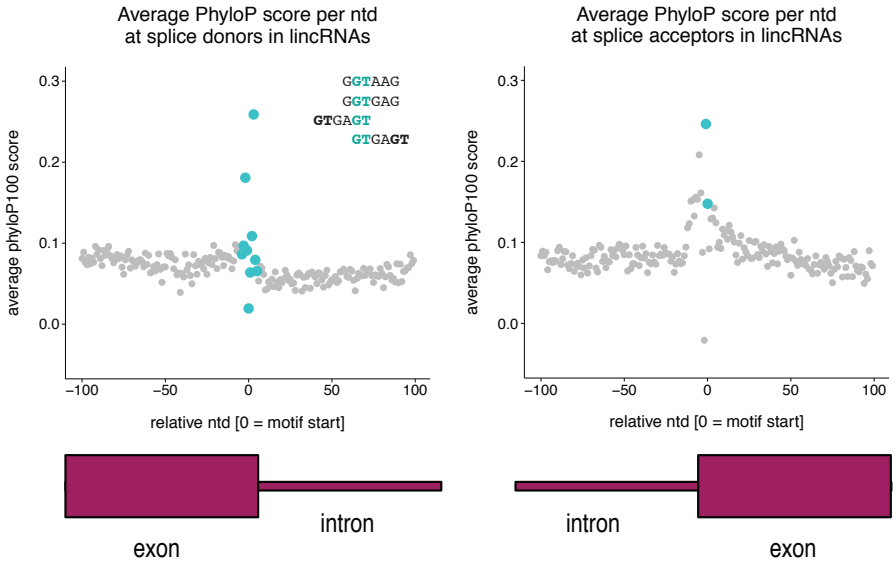


C.



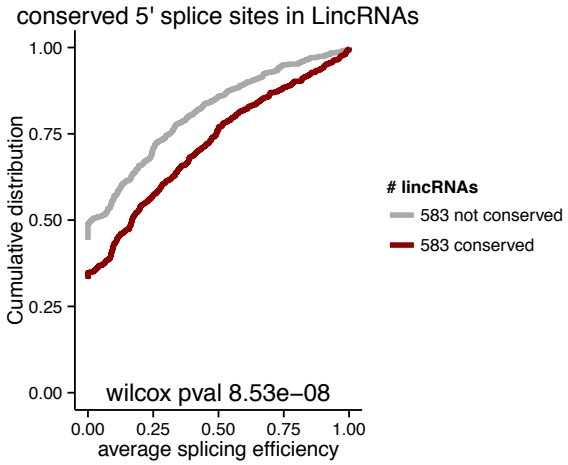
Supplemental Fig S30. Splicing efficiency of mRNAs with U2af65 bound or not bound with similar expression levels using Zarnack et al. (2013) dataset. mRNAs between bound/unbound groups are expression matched using abundance values from the same cell line where the CLIP experiment was carried out. **A.** All mRNAs. **B**. only mRNAs with a single

isoform of two exons were analyzed. **C.** mRNAs with splicing efficiency of zero were excluded from the analysis.

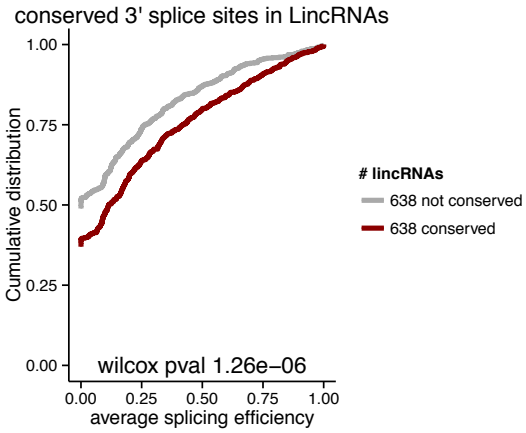


Supplemental Fig S31. Average conservation scores at canonical 5 prime splice sites (left) and 3 prime splice sites (right) in lincRNAs. Because U1 binds to 5′ splice sites, we calculated conservation at 5′ splice sites considering the canonical GT donor dinucleotides plus and minus four nucleotides to complete the U1 binding motif. The three canonical U1 motifs are represented in the left plot with their four possible alignments at the canonical splice sites. On average, both 5′ and 3′ splice junctions were significantly more conserved than neighboring regions. Empirical P<0.011 and P<0.005 respectively.

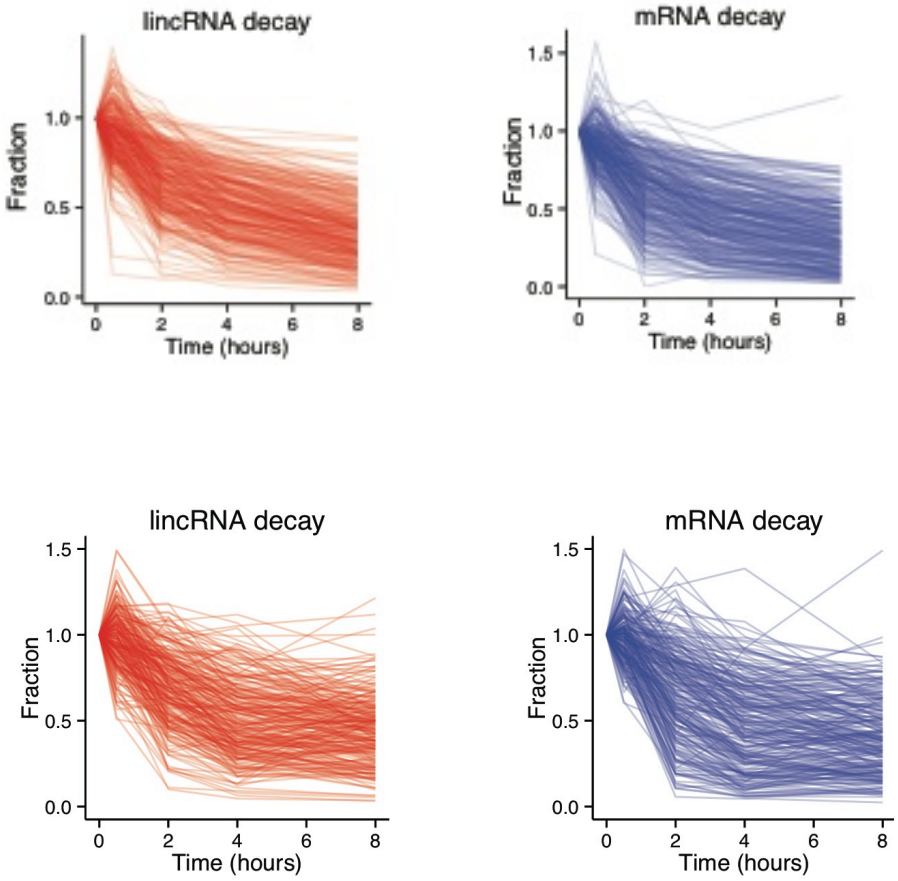
A



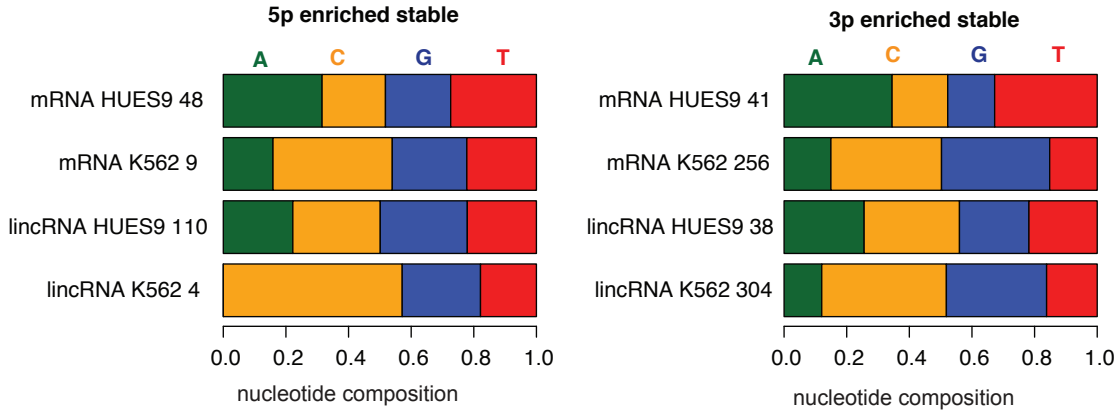
B



Supplemental Fig S32. Splicing efficiency for lincRNAs with 5′ splice site conserved (A) and 3′ splice site conserved (B) compared to expression-matched lincRNAs. Both, lincRNAs with 5′ and 3′ splice sites conserved are more efficiently spliced than equally expressed lincRNAs with no splice site conserved.

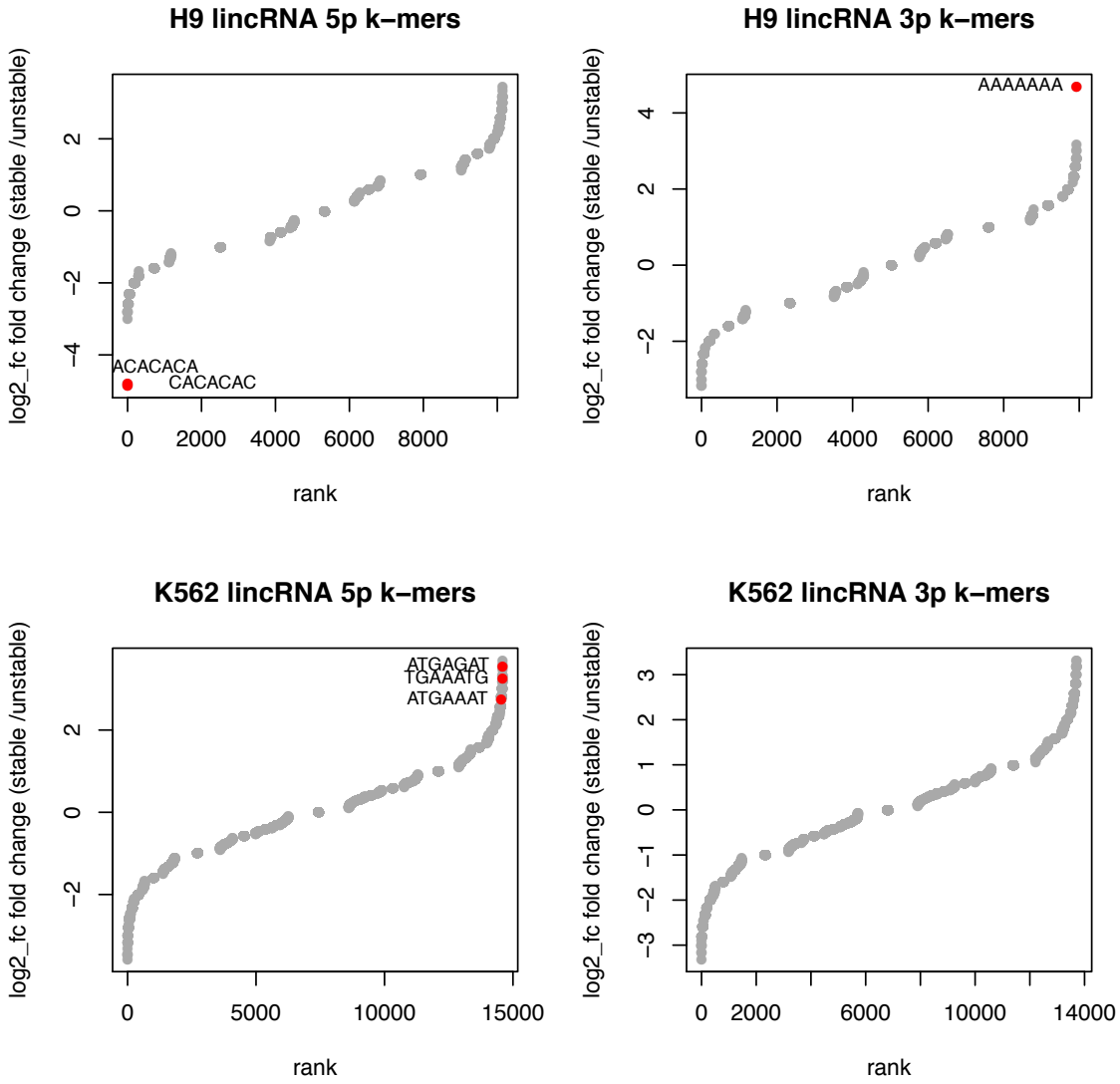


Supplemental Fig S33. RNA decay of lincRNAs (left) and mRNAs (right) after actynomycin treatment in K562 (top) and HUES9 cells (bottom).

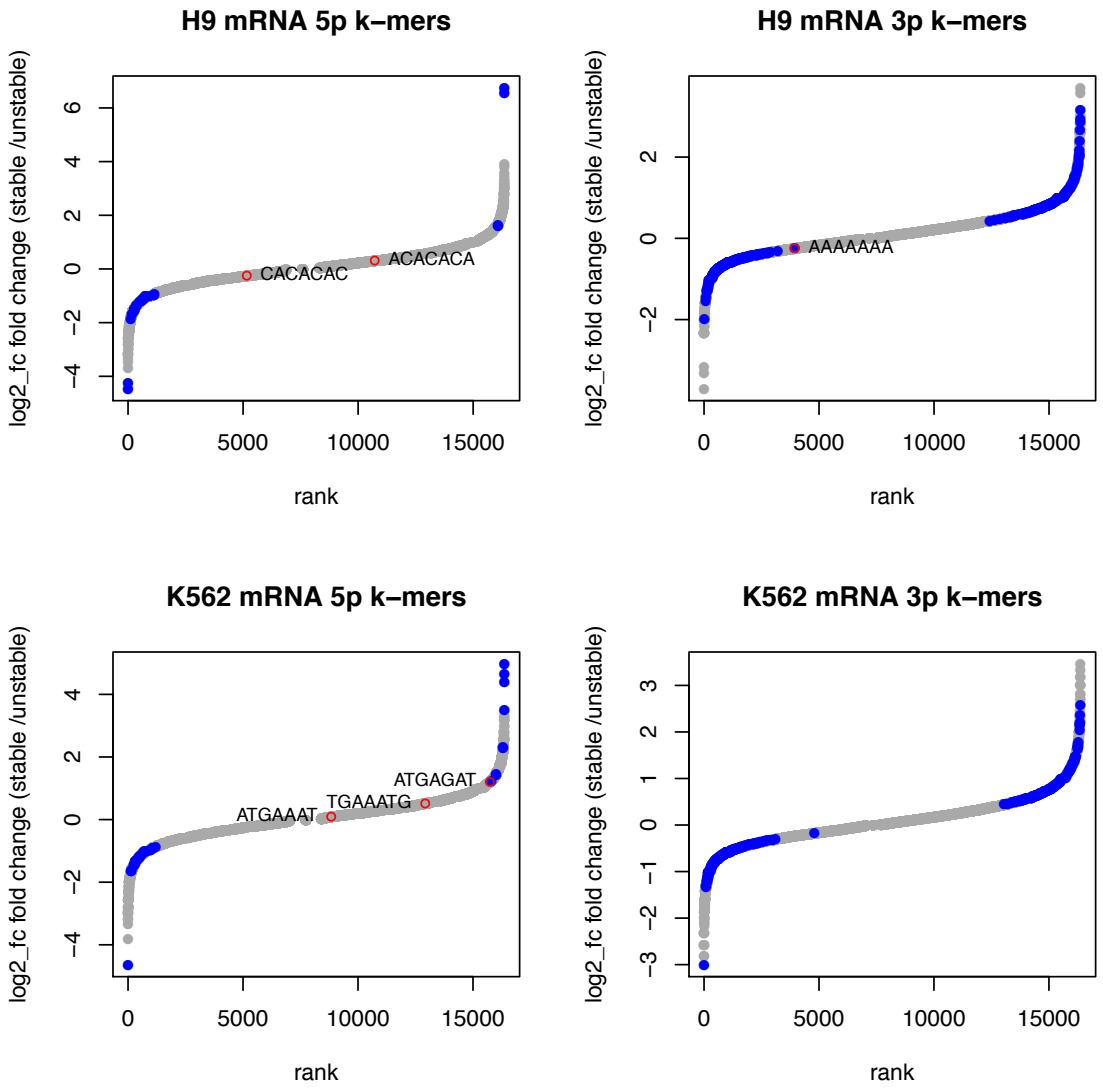


Supplemental Fig S34. Nucleotide composition of 7-mers that are enriched in stable mRNAs and lincRNAs in both the 5′ (left) or 3′ ends (right). For mRNAs, we selected 7-mers to be significant at 0.05<FDR. For lincRNAs, due to their low sample size, we selected 7-

mers with p-values < 0.05 to confirm that the trend was also present. Numbers beside cell type labels correspond to the number of *k*-mers analyzed.

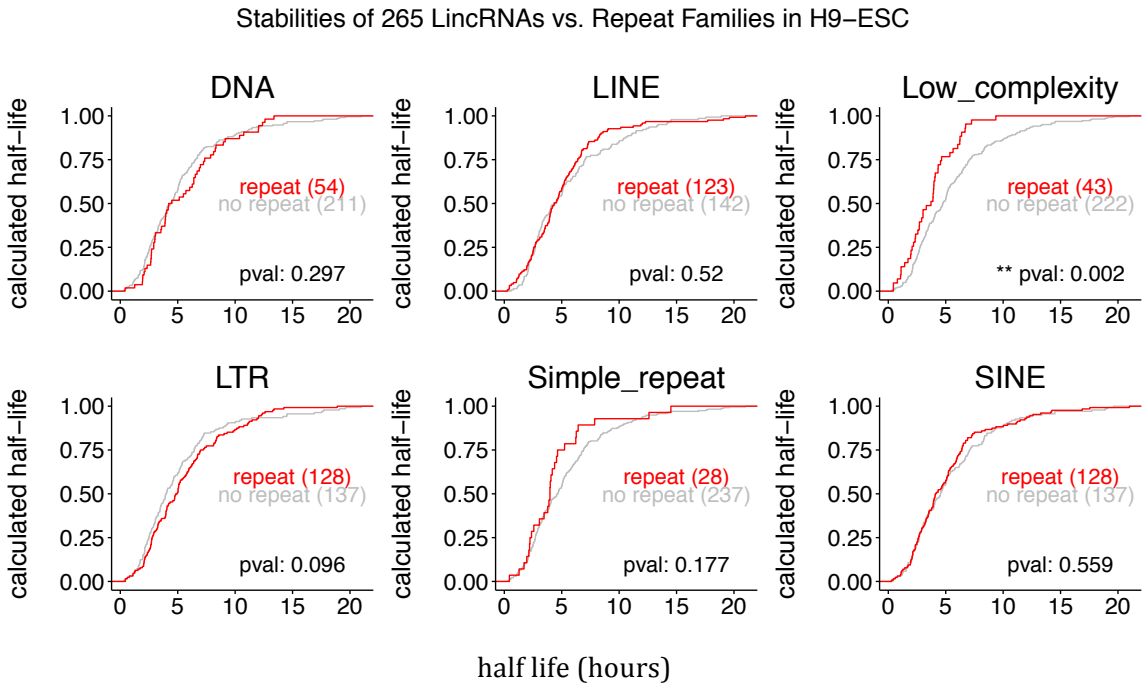


Supplemental Fig S35. *K*-mer analysis in lincRNA and mRNA based on stability. A. Rank of all 7-mers by enrichment in the first 5′ (left) or last 3′ (right) 500 bp in stable transcripts relative to unstable transcripts in lincRNAs. Colored in red are those 7-mers found to be significantly differentially represented in stable vs unstable lincRNAs transcripts. For this analysis we looked at genes expressed at > 1FPKM at time zero and selected the longest isoform.

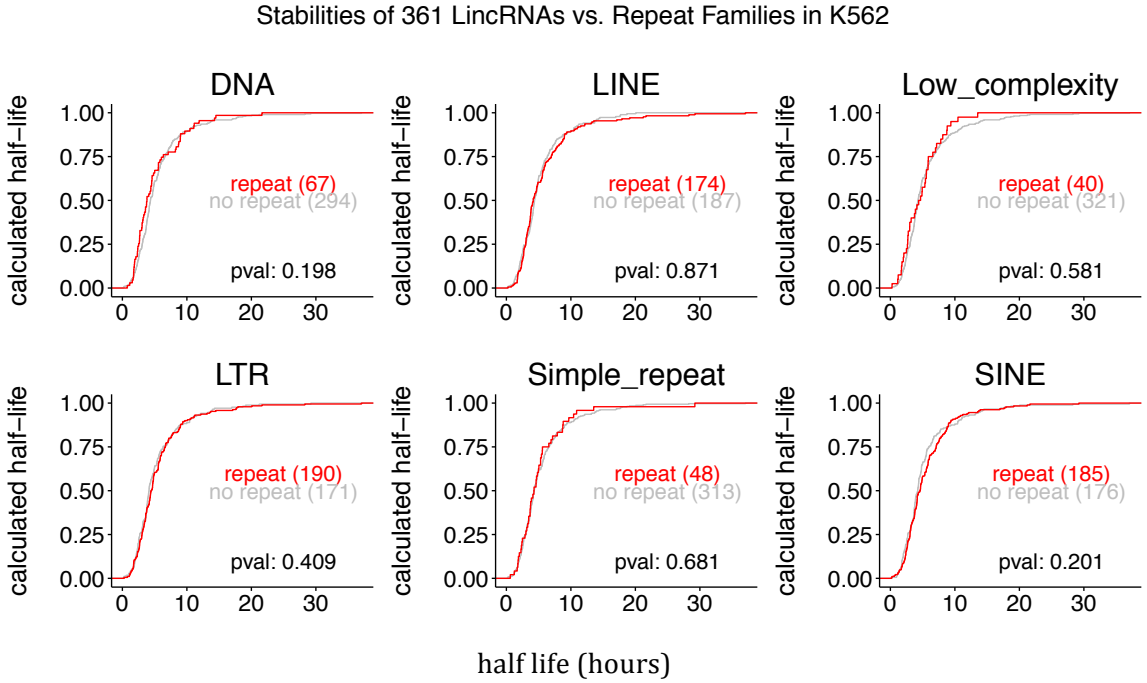


Supplemental Fig S36. *K*-mer analysis in mRNA and mRNA based on stability. A. Rank of all 7-mers by enrichment in the first 5´ (left) or last 3´ (right) 500 bp in stable transcripts relative to unstable transcripts in lincRNAs. Colored in blue are those 7-mers found to be significantly differentially represented in stable vs unstable mRNAs transcripts. In red, we highlight those *k*-mers that are significantly differentially represented in stable vs unstable lincRNAs transcripts in that transcript region and cell line. For this analysis we looked at genes expressed at > 1FPKM at time zero and selected the longest isoform.

A

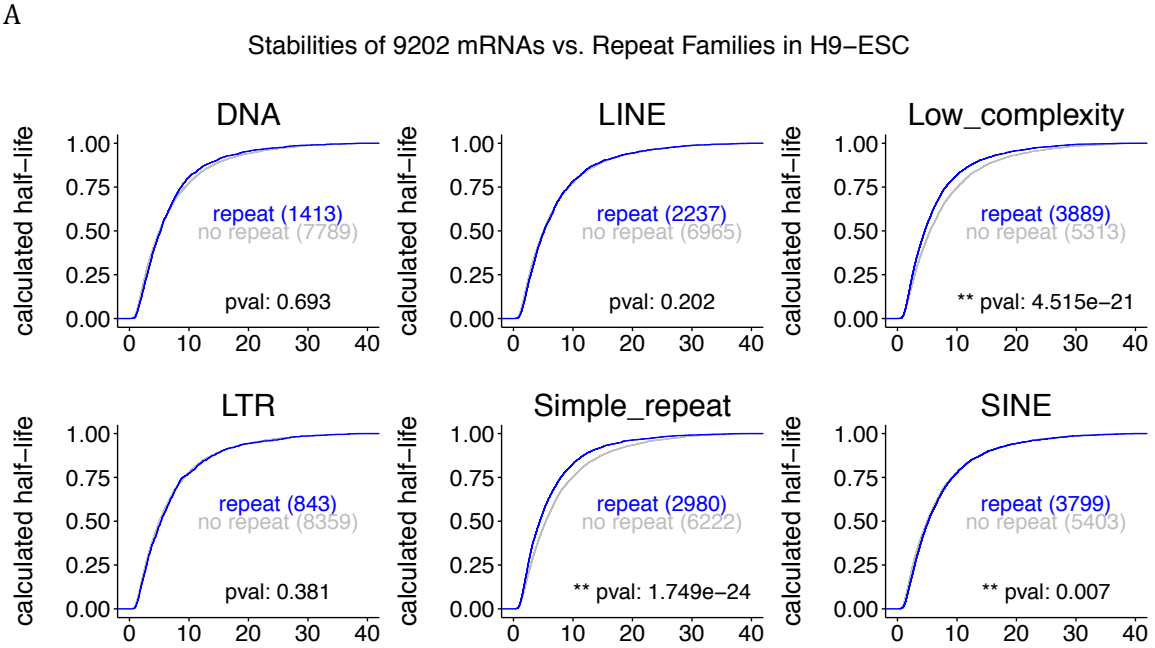


B

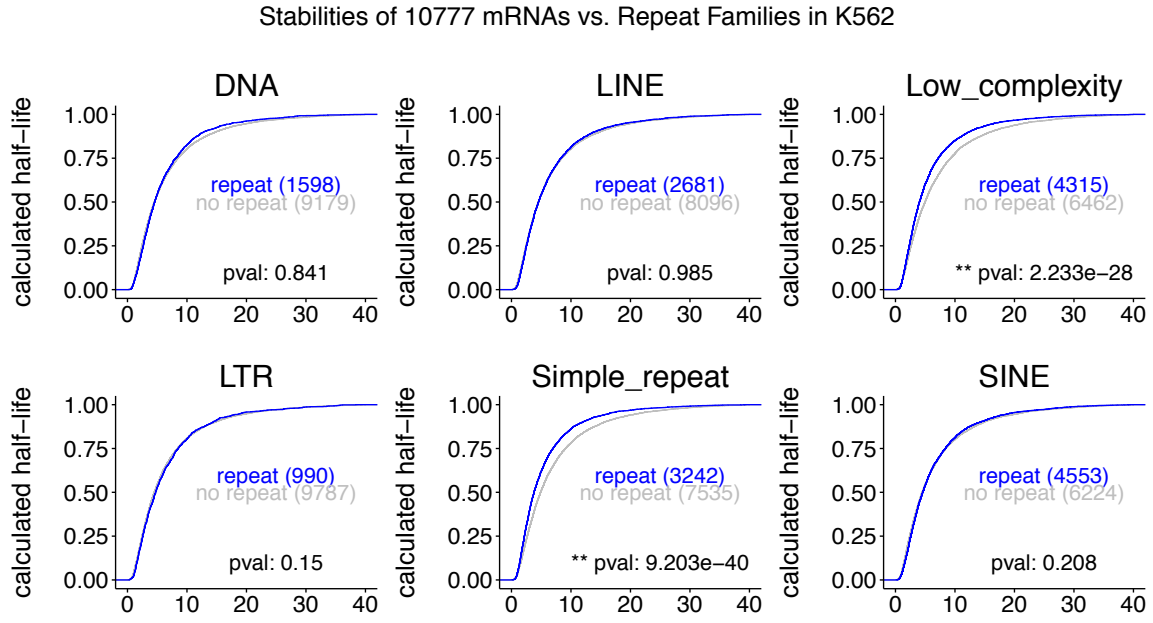


Supplemental Fig S37. Cumulative distribution (y-axis) of the half-lives in hours (x-axis) of lincRNA transcripts with or without certain repeat families overlapping any of their exons in HUES9 (A) and K562 (B).

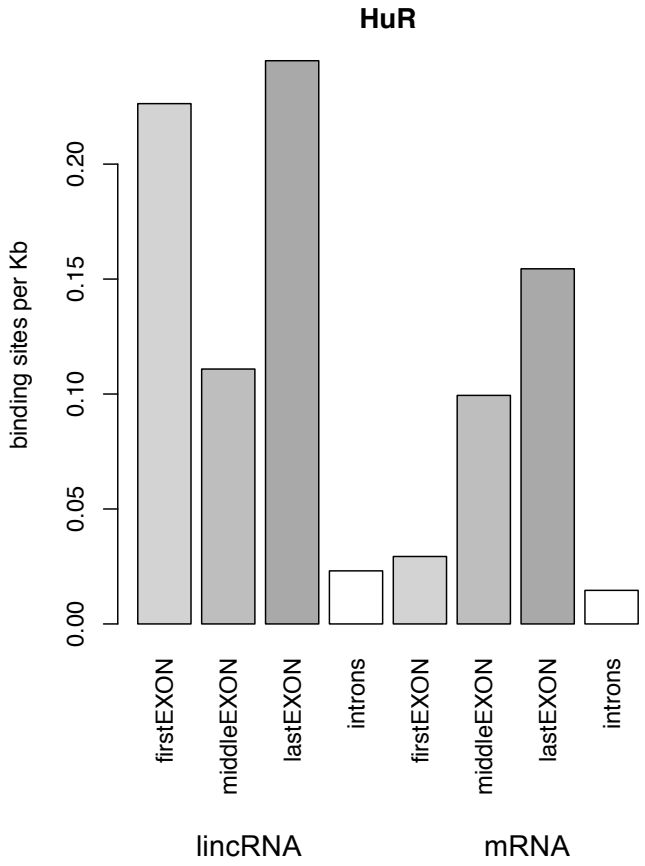
A



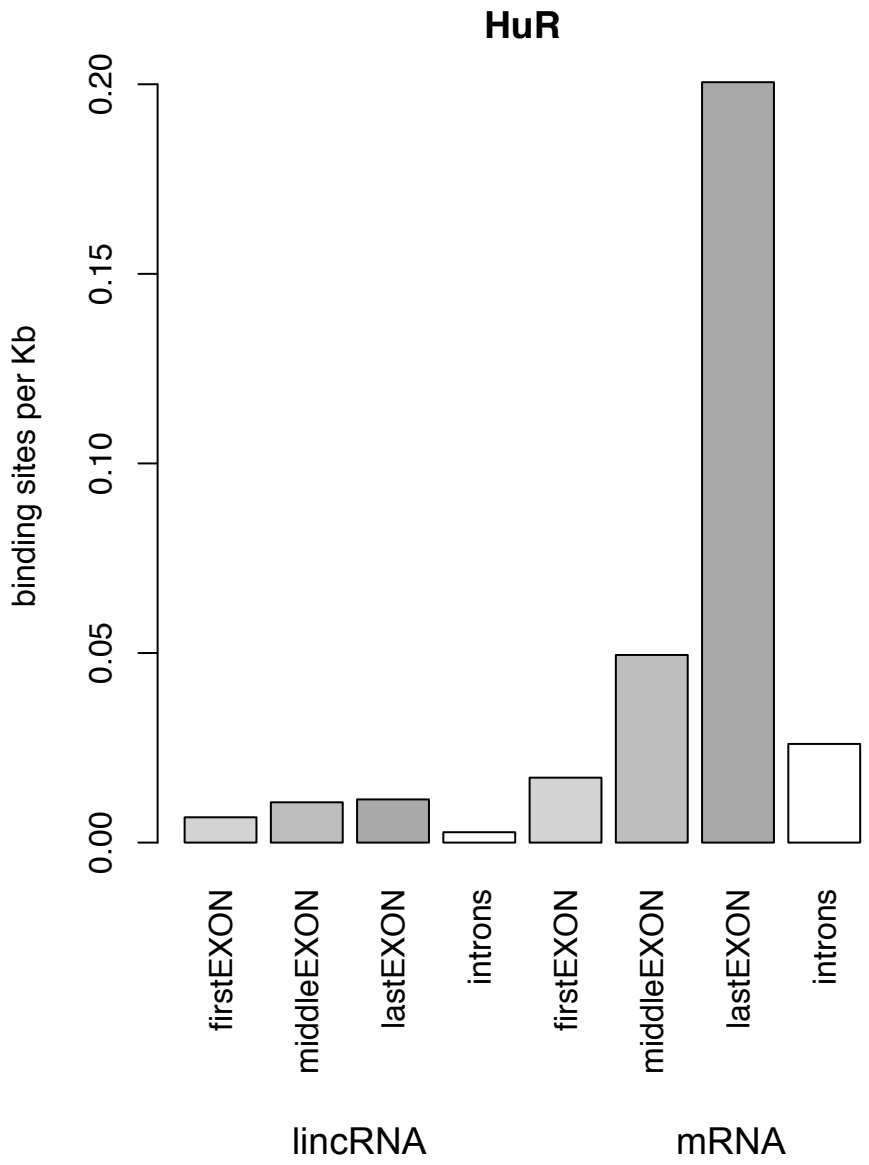
B



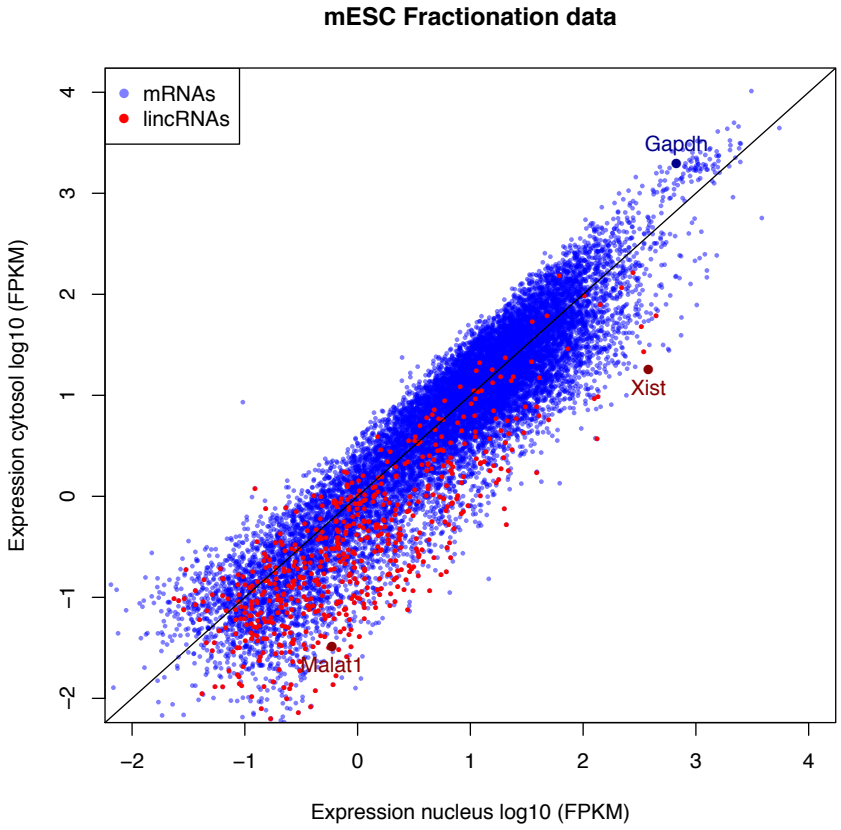
Supplemental Fig S38. Cumulative distribution (y-axis) of the half-lives in hours (x-axis) of mRNA transcripts with or without certain repeat families overlapping any of their exons in HUES9 (A) and K562 (B).



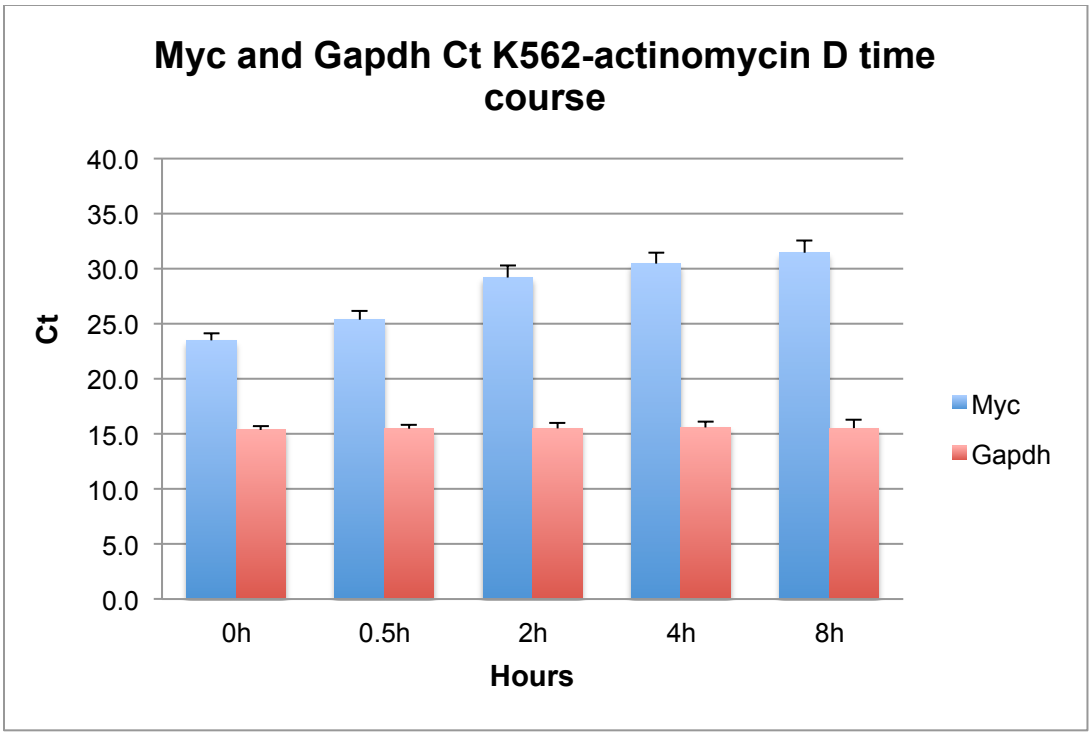
Supplemental Fig S39. Density of HuR binding sites in different regions of lincRNAs (left) and expression matched mRNAs (right). Expression was matched using expression levels in HEK293 which is the same cell line for which the CLIP-seq was carried out.

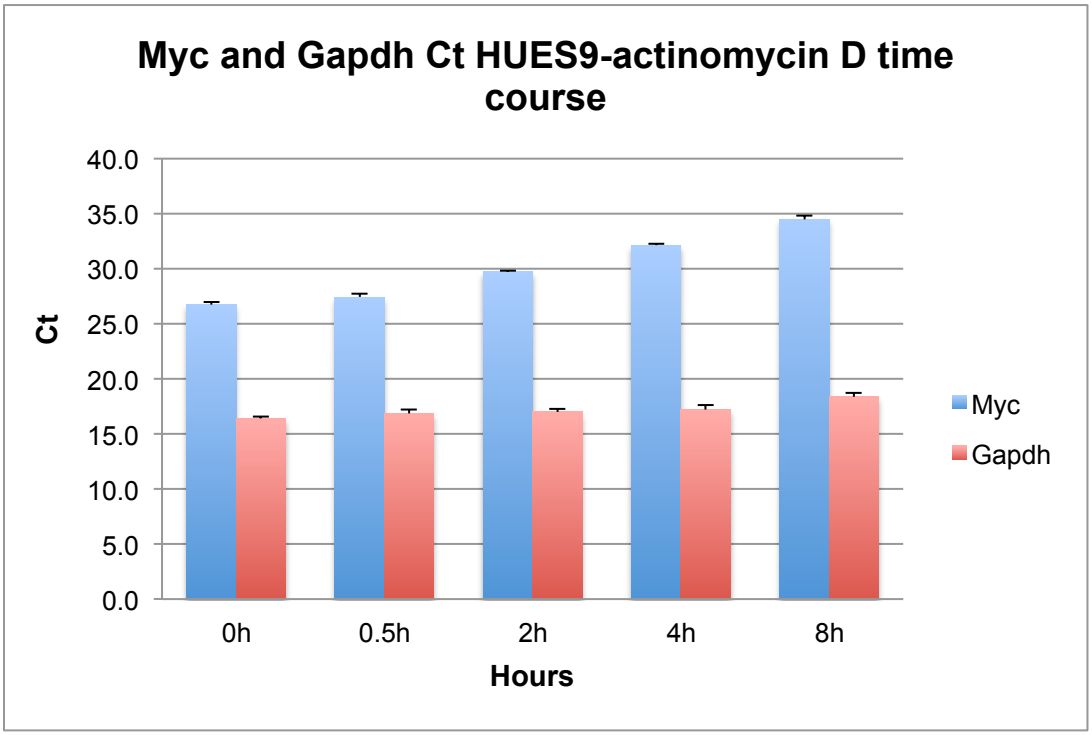


Supplemental Fig S40. Density of HuR binding sites in different regions of efficiently spliced lincRNAs (left) and mRNAs (right). Efficiently spliced lincRNAs were selected as those that had average splicing efficiency higher than 0.5 across seven ENCODE cell lines.



Supplemental Fig S41. Gene quantifications in cytosol and nuclear fractions. We selected mRNAs and lincNRAs with mean (expression nuc + expression cyt) > 0.1 FPKM. XIST and MALAT1 are lincRNAs known to be localized in the nucleus and not be exported in the cytoplasm. GAPDH is a housekeeping gene highly expressed.





Supplemental Fig S42. MYC and GAPDH qPCR levels in stability assay. Expression levels across stability time course for Myc and GAPDH in K562 (top) and HUES9 (bottom) cells.

Supplementary tables

Supplemental Table S1. LincRNA gene properties across ENCODE cell lines. Values of expression abundance measured in FPKMs (Expression), splicing efficiency (Splicing_Efficiency), ratio of nuclear/cytosol abundance (Nuc_to_Cyt), number of histone marks (nHM) detected in the promoter and number of transcription factors (nTFs) detected in the promoter across seven encode cell lines. Tissue specificity was calculated using expression values for 20 diverse tissue types as in (Cabili et al. 2011).

Supplemental Table S2. Histone mark enrichment differences between lincRNA vs mRNA promoters across ENCODE cell lines. To calculate Histone Mark enrichment we calculated the number of lincRNAs and mRNAs that had the specific histone mark peak in their promoter (n.linc.bind and n.mrna.bind). LincRNAs and mRNAs are expression matched and have expression higher than 0.1 FPKM in the tested cell line. We then used Phi effect size to calculate enrichment (Effect.Size). Values larger than zero indicate enrichment in mRNAs and values below zero enrichment in lincRNAs. P-values for Fisher's exact test are also indicated (fisher.pval). "Lab.Name" refers to the lab that performed the ENCODE ChIP-seq, and "Cell.Line" to the cell line used to perform the ChIP-seq.

Supplemental Table S3. Transcription factor enrichment differences between lincRNA vs mRNA promoters across ENCODE cell lines. To calculate transcription factor (TF) enrichment we calculated the number of lincRNAs and mRNAs that had the specific histone mark peak in their promoter (n.linc.bind and n.mrna.bind). LincRNAs and mRNAs are expression matched and have expression higher than 0.1 FPKM in the tested cell line. We then used Phi effect size to calculate enrichment (Effect.Size). Values larger than zero indicate enrichment in mRNAs and values below zero enrichment in lincRNAs. P-values for Fisher's exact test are also indicated. "Lab.Name" refers to the lab that performed the ENCODE ChIP-seq, and "Cell.Line" to the cell line used to perform the ChIP-seq.

Supplemental Table S4. Transcription factor conservation tests across lincRNA and mRNA promoters for promoters defined as -/+5Kb. We calculated conservation tests per TF by centering all the ChIP-seq peaks overlapping a lincRNA or mRNA promoter at the peak maxima, calculating average conservation per nucleotide across these regions and comparing it to the regions at each side as background. The table shows the raw p-values for the Wilcoxon test and the adjusted p-values using FDR.

Supplemental Table S5. List of TF that have their binding sites significantly more conserved than background in lincRNA promoters for promoters defined as -2Kb/+1Kb. We calculated conservation tests per TF by centering all the ChIP-seq peaks overlapping a lincRNA or mRNA promoter at the peak maxima, calculating average conservation per nucleotide across these regions and comparing it to the regions at each side as background. The table shows the raw p-values for the Wilcoxon test and the adjusted p-values using FDR.

Supplemental Table S6. Presence/absence of TF and conserved TF binding sites per lincRNA promoter (-+5Kb). We compared conservation values at those nucleotides within each ChIP-seq peak present in lincRNA promoters and compared them to their surrounding regions (see methods). We assigned significance based on Wilcoxon test and FDR<0.05. The

table gives the number of transcription factor types that are bound per lincRNA promoter across the ENCODE cell lines (nTFBS_per_gene), and the number of these that are conserved (conserved_TFBS_per_gene).

Supplemental Table S7. Presence/absence of TF and conserved TF binding sites per lincRNA promoters (-2kb/+1Kb). We compared conservation values at those nucleotides within each chipSeq peak present in lincRNA promoters and compared them to their surrounding regions (see methods). We assigned significance based on Wilcoxon test and FDR<0.05. The table gives the number of transcription factor types that are bound per lincRNA promoter across the ENCODE cell lines (nTFBS_per_gene), and the number of these that are conserved (conserved_TFBS_per_gene).

Supplemental Table S8. Number of lincRNA, mRNA promoters of the same size that have at least one TFBS or at least one conserved TFBS compared to random intergenic regions. For this analysis we used 19,575 mRNA and 5,196 lincRNA and 5,196 intergenic regions.

promoter definition	Presence of	lincRNA	mRNA	Intergenic
long (-5Kb/+5Kb)	at least 1 TFBS	4507 (86.7%)	18734 (95.7%)	3226 (62.1%)
long (-5Kb/+5Kb)	at least 1 conserved TFBS	3220 (62%)	17169 (87.7%)	1833 (35.3%)
short (-2Kb/+1Kb)	at least 1 TFBS	3437 (66.1%)	17643 (90.1%)	1860 (35.8%)
short (-2Kb/+1Kb)	at least 1 conserved TFBS	1956 (37.6%)	15345 (78.4%)	803(15.4%)

Supplemental Table S9. Number of conserved 3´ or 5´ splice sites and U1 binding sites. We calculated average conservation values for those nucleotides in 3´ or 5´ splice sites or at U1 binding sites. We then calculated average conservation of adjacent nucleotides (of the same length) in the flanking regions (see methods). We considered a 3´, 5´ or U1 sites to be conserved if it had a higher average conservation score than 95% of all flanking combinations.

Supplemental Table S10. Half-life estimates for lincRNAs and mRNAs in H1-ESC and K562. Information of the half-life estimates for each of the three replicates in both K562 and H1-ESC cell lines and the correlation with the exponential decay model. "avExpT0.k562" and "avExpT0.hues9" gives the average expression value across replicates at time zero in FPKM.

Supplemental Table S11. List of candidate lincRNAs predicted to be functional RNA molecules. List of lincRNAs for which we detect evidence of conserved DNA TFBS in their promoter region (nTFcons) and conserved RNA motifs, either U1 binding sites (nU1.1Kb) or splice sites (n3pSS or n5pSS) and that are efficiently spliced (average splicing efficiency (avSE)> 0.5 across the analyzed ENCODE cell lines). The latest 11th columns with cell line names contain expression values in FPKM across each cell line.

Supplemental Table S12. Information on all ChIP-seq data analyzed including histone marks and transcription factors. We include all URLs from which broadpeak and narrowpeak files were downloaded.

Bibliography

Almada AE, Wu X, Kriz AJ, Burge CB, Sharp PA. 2013. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**: 360–3. http://dx.doi.org/10.1038/nature12349 (Accessed February 1, 2016).

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**: 1915–27.

Goff LA, Groff AF, Sauvageau M, Trayes-Gibson Z, Sanchez-Gomez DB, Morse M, Martin RD, Elcavage LE, Liapis SC, Gonzalez-Celeiro M, et al. 2015. Spatiotemporal expression and transcriptional perturbations by long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci U S A* **112**: 6855–62. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4460505&tool=pmcentrez&rendertype=abstract (Accessed April 5, 2016).

Hacisuleyman E, Goff LA, Trapnell C, Williams A, Henao-Mejia J, Sun L, McClanahan P, Hendrickson DG, Sauvageau M, Kelley DR, et al. 2014. Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nat Struct Mol Biol* **21**: 198–206. http://dx.doi.org/10.1038/nsmb.2764 (Accessed August 20, 2015).

Ho DE, Imai K, King G, Stuart EA. 2011. MatchIt : Nonparametric Preprocessing for Parametric Causal Inference. *J Stat Softw* **42**: 1–28. https://www.jstatsoft.org/index.php/jss/article/view/v042i08/v42i08.pdf (Accessed March 21, 2016).

Kelley DR, Hendrickson DG, Tenen D, Rinn JL. 2014. Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions. *Genome Biol* **15**: 537. http://genomebiology.com/2014/15/12/537 (Accessed February 1, 2016).

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–9. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2723002&tool=pmcentrez&rendertype=abstract (Accessed July 9, 2014).

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**: 764–70. http://bioinformatics.oxfordjournals.org/content/27/6/764.abstract (Accessed July 10, 2014).

Sauvageau M, Goff LA, Lodato S, Bonev B, Groff AF, Gerhardinger C, Sanchez-Gomez DB, Hacisuleyman E, Li E, Spence M, et al. 2013. Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife* **2**: e01749. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3874104&tool=pmcentrez&rendertype=abstract (Accessed December 7, 2014).

Zarnack K, König J, Tajnik M, Martincorena I, Eustermann S, Stévant I, Reyes A, Anders S, Luscombe NM, Ule J. 2013. Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell* **152**: 453–66. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3629564&tool=pmcentr

ez&rendertype=abstract (Accessed February 2, 2016).