Short RNAs are transcribed from repressed Polycomb target genes and interact with Polycomb Repressive Complex-2

Kanhere et al.

Supplemental Data

Index

Supplemental Figures and legends

Supplemental Tables

Figure S1

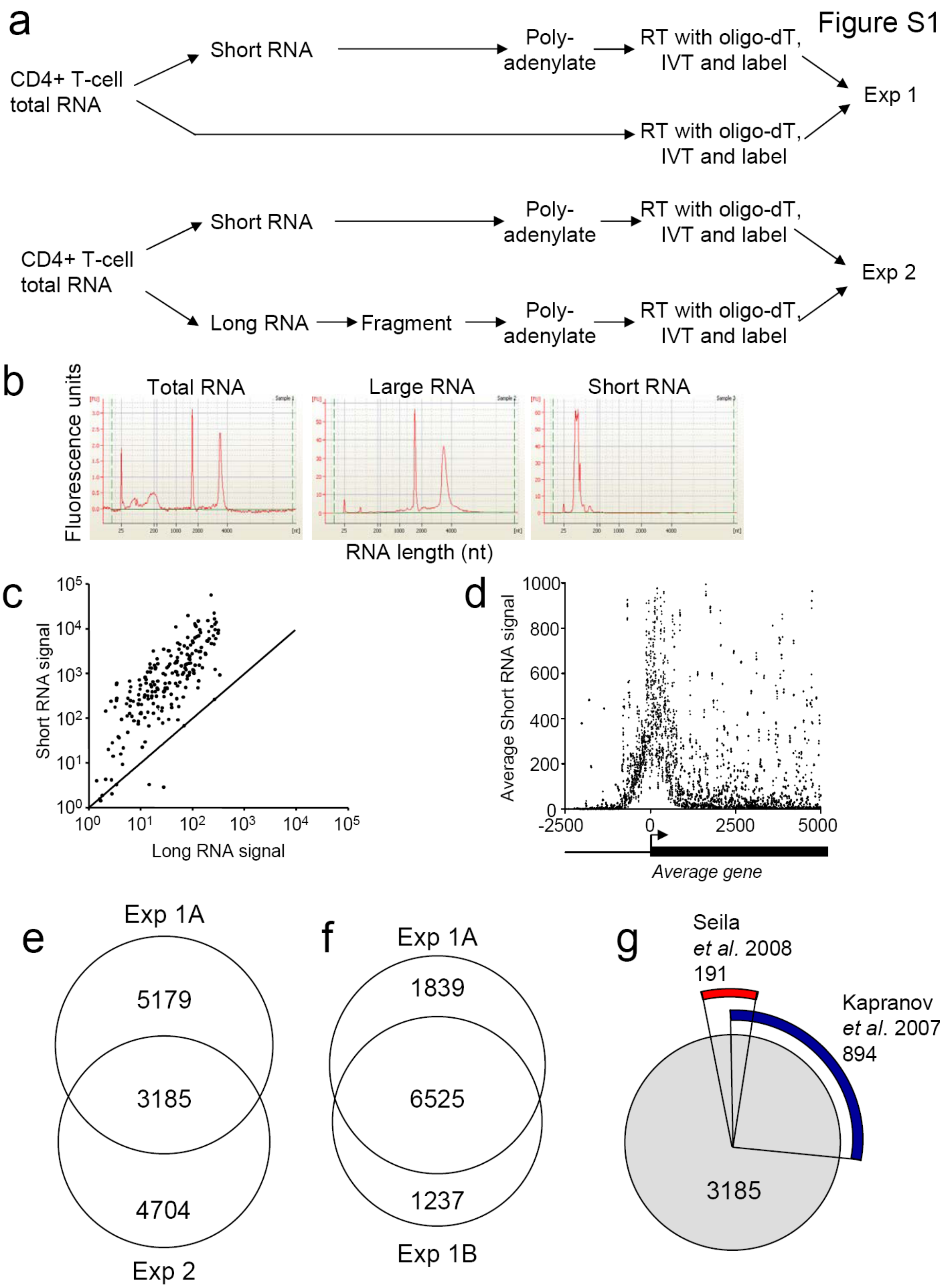

a
CD4+ T-cell total RNA
Short RNA
Poly-adenylate
RT with oligo-dT, IVT and label
Exp 1
RT with oligo-dT, IVT and label
CD4+ T-cell total RNA
Short RNA
Poly-adenylate
RT with oligo-dT, IVT and label
Exp 2
Long RNA
Fragment
Poly-adenylate
RT with oligo-dT, IVT and label
b
Fluorescence units
Total RNA
Large RNA
Short RNA
RNA length (nt)
c
Short RNA signal
10^0
10^1
10^2
10^3
10^4
10^5
Long RNA signal
d
Average Short RNA signal
0
200
400
600
800
1000
-2500
0
2500
5000
Average gene
e
Exp 1A
5179
3185
4704
Exp 2
f
Exp 1A
1839
6525
1237
Exp 1B
g
Seila *et al.* 2008
191
Kapranov *et al.* 2007
894
3185

Figure S2

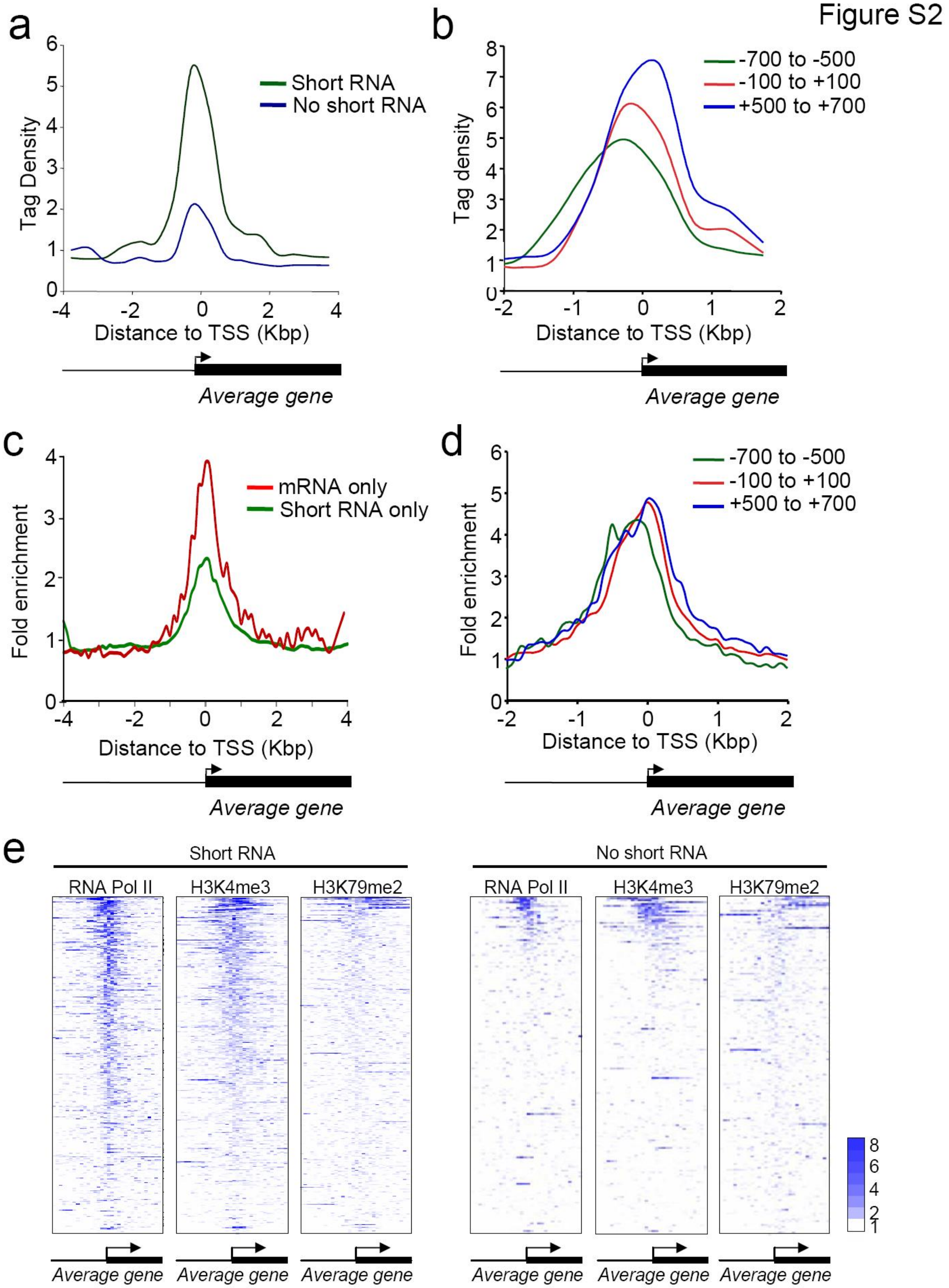

a
Tag Density
6
5
4
3
2
1
0
-4
-2
0
2
4
Distance to TSS (Kbp)
Short RNA
No short RNA
Average gene
b
Tag density
8
7
6
5
4
3
2
1
0
-2
-1
0
1
2
Distance to TSS (Kbp)
-700 to -500
-100 to +100
+500 to +700
Average gene
c
Fold enrichment
4
3
2
1
0
-4
-2
0
2
4
Distance to TSS (Kbp)
mRNA only
Short RNA only
Average gene
d
Fold enrichment
6
5
4
3
2
1
0
-2
-1
0
1
2
Distance to TSS (Kbp)
-700 to -500
-100 to +100
+500 to +700
Average gene
e
Short RNA
No short RNA
RNA Pol II
H3K4me3
H3K79me2
RNA Pol II
H3K4me3
H3K79me2
8
6
4
2
1
Average gene
Average gene
Average gene
Average gene
Average gene
Average gene

Figure S3

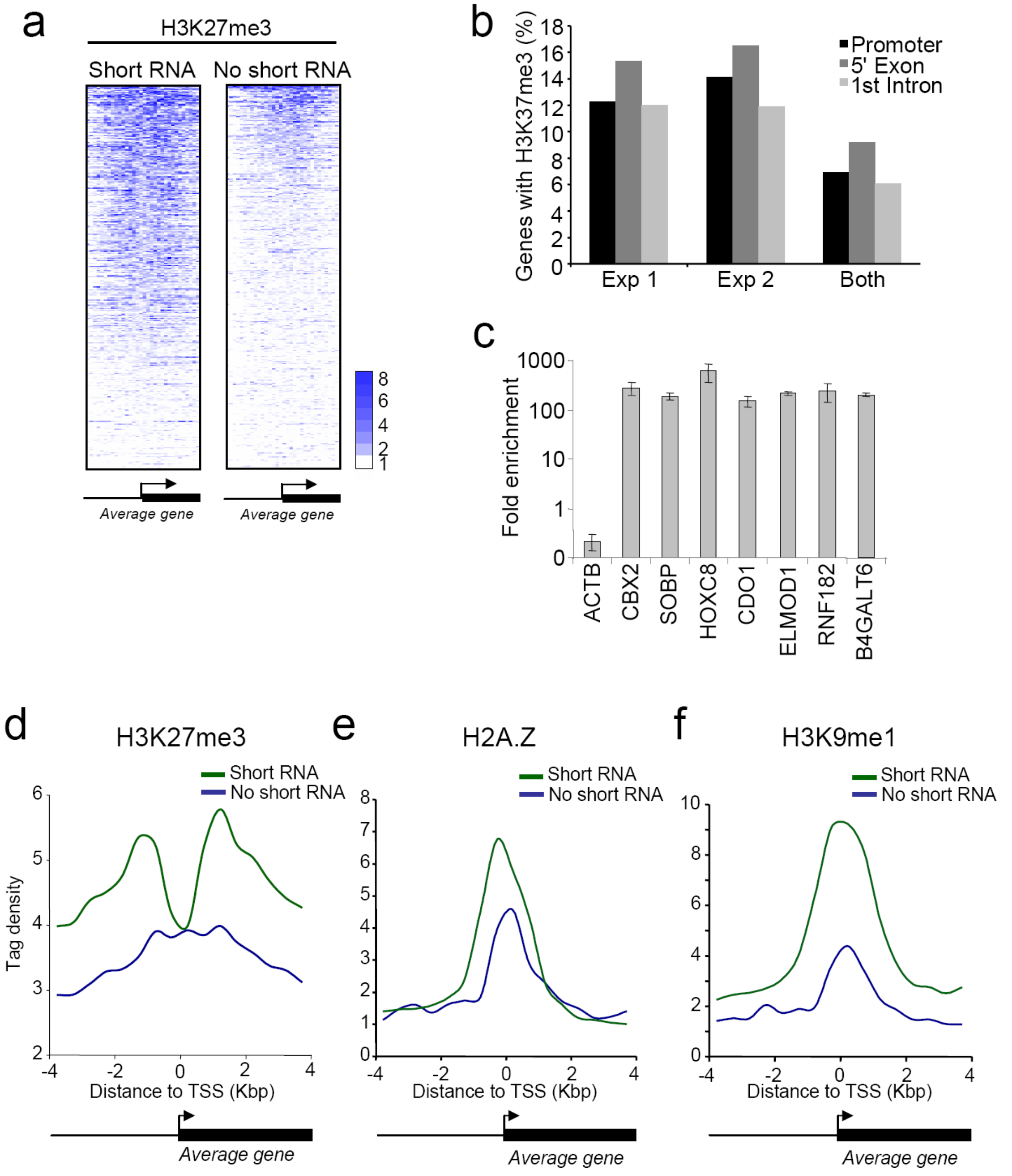

a
H3K27me3
Short RNA
No short RNA
8
6
4
2
1
Average gene
Average gene
b
Genes with H3K37me3 (%)
0
2
4
6
8
10
12
14
16
18
Promoter
5' Exon
1st Intron
Exp 1
Exp 2
Both
c
Fold enrichment
1000
100
10
1
0
ACTB
CBX2
SOBP
HOXC8
CDO1
ELMOD1
RNF182
B4GALT6
d
H3K27me3
Short RNA
No short RNA
Tag density
6
5
4
3
2
-4
-2
0
2
4
Distance to TSS (Kbp)
Average gene
e
H2A.Z
Short RNA
No short RNA
8
7
6
5
4
3
2
1
0
-4
-2
0
2
4
Distance to TSS (Kbp)
Average gene
f
H3K9me1
Short RNA
No short RNA
10
8
6
4
2
0
-4
-2
0
2
4
Distance to TSS (Kbp)
Average gene

Figure S3 cont

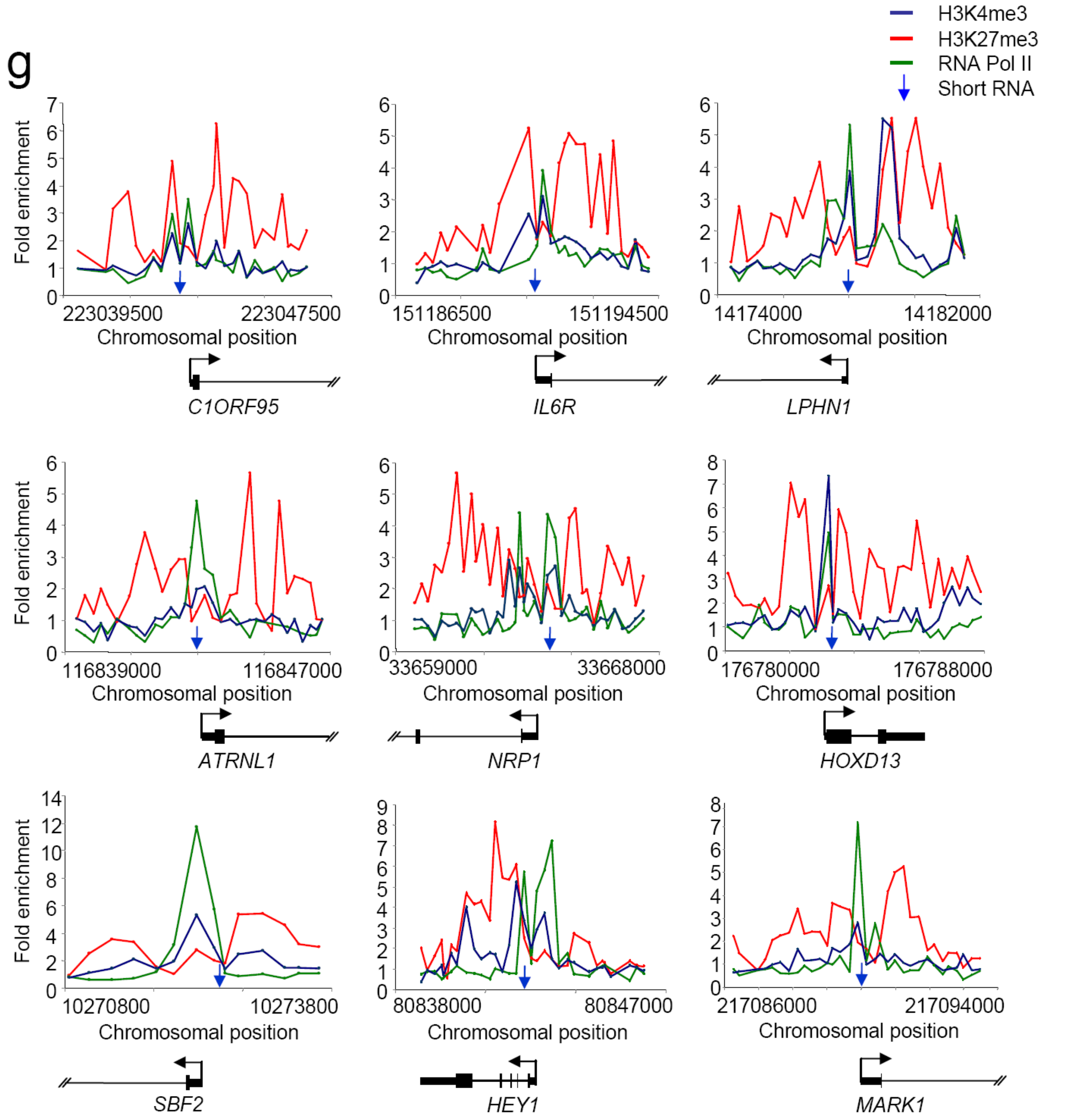


Figure S3 cont.

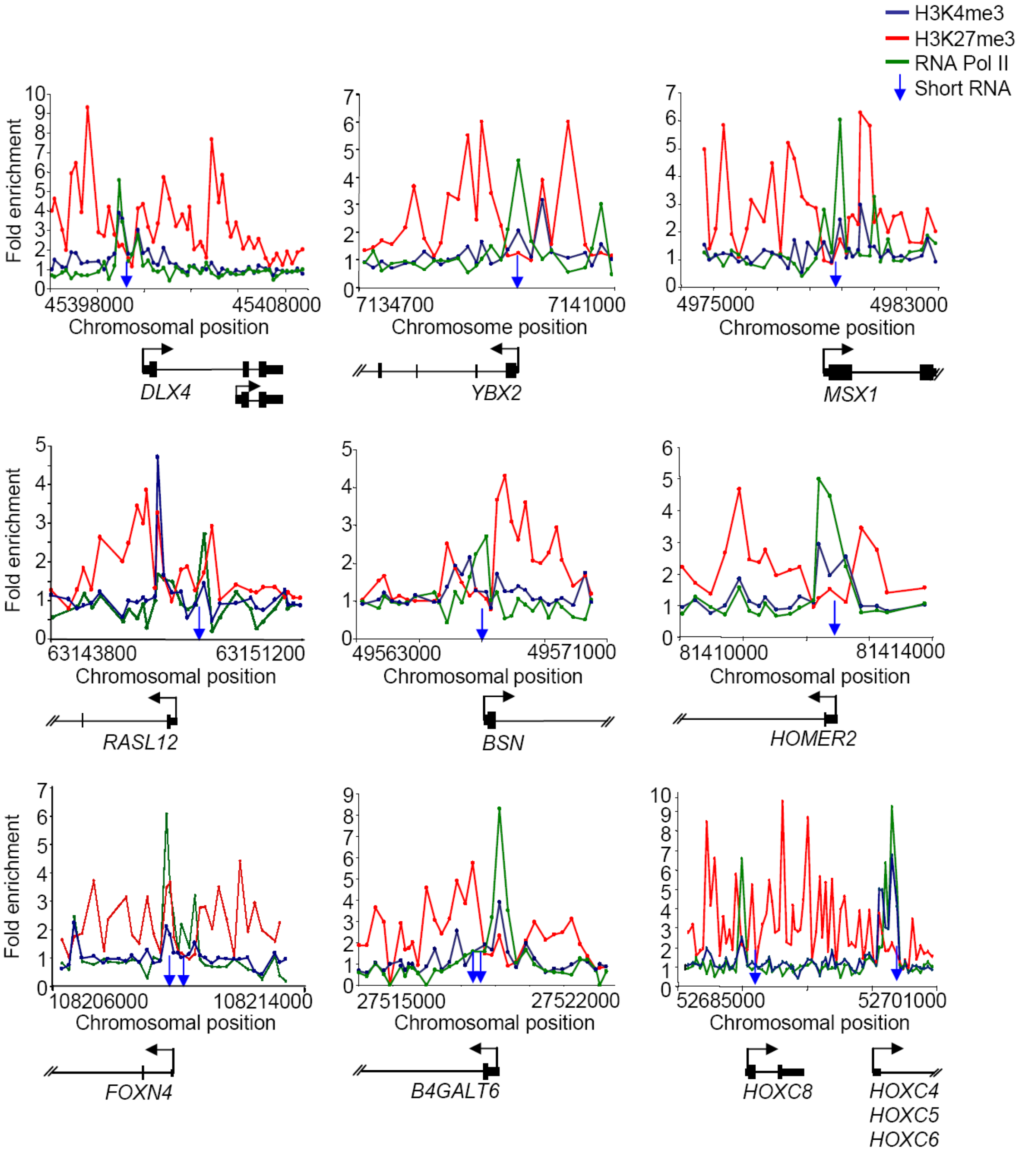


Figure S4

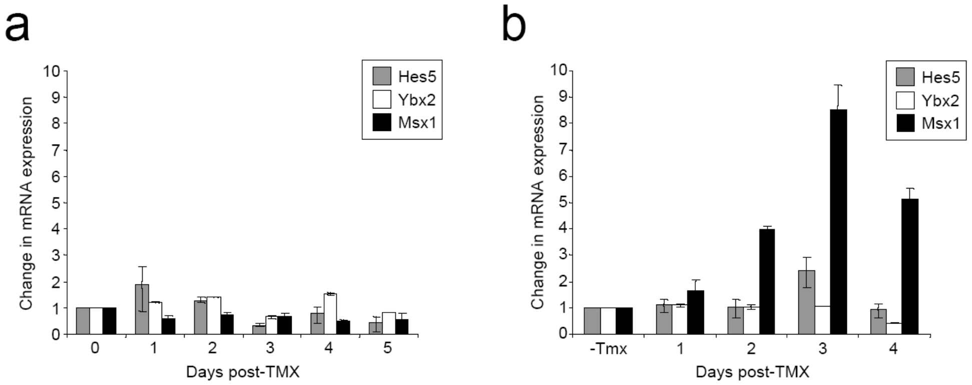

a
Change in mRNA expression
10
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
Days post-TMX
Hes5
Ybx2
Msx1
b
Change in mRNA expression
10
9
8
7
6
5
4
3
2
1
0
-Tmx
1
2
3
4
Days post-TMX
Hes5
Ybx2
Msx1

Figure S5

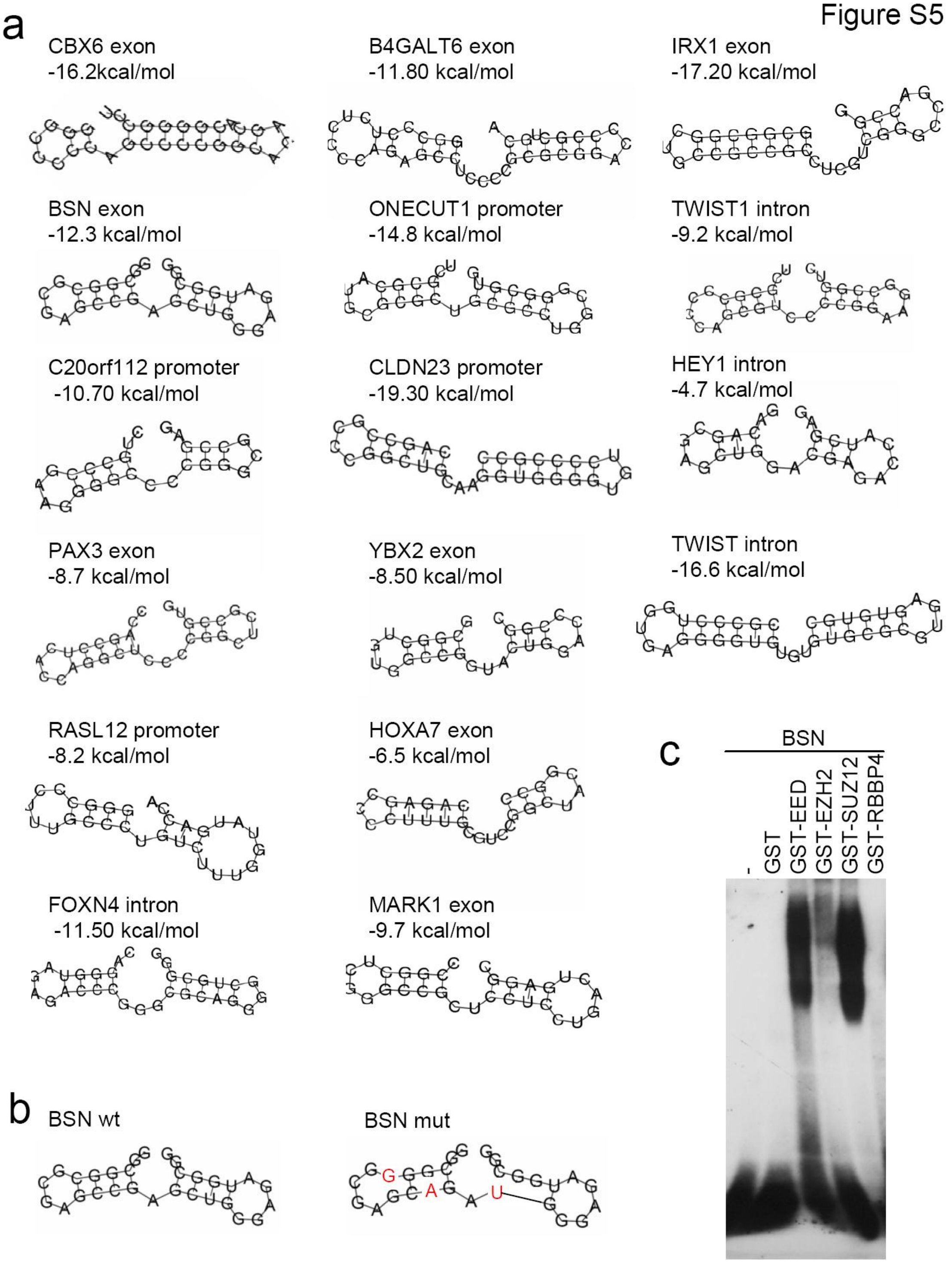

a
CBX6 exon
-16.2kcal/mol
B4GALT6 exon
-11.80 kcal/mol
IRX1 exon
-17.20 kcal/mol
BSN exon
-12.3 kcal/mol
ONECUT1 promoter
-14.8 kcal/mol
TWIST1 intron
-9.2 kcal/mol
C20orf112 promoter
-10.70 kcal/mol
CLDN23 promoter
-19.30 kcal/mol
HEY1 intron
-4.7 kcal/mol
PAX3 exon
-8.7 kcal/mol
YBX2 exon
-8.50 kcal/mol
TWIST intron
-16.6 kcal/mol
RASL12 promoter
-8.2 kcal/mol
HOXA7 exon
-6.5 kcal/mol
FOXN4 intron
-11.50 kcal/mol
MARK1 exon
-9.7 kcal/mol
b
BSN wt
BSN mut
G
A
U
c
BSN
-
GST
GST-EED
GST-EZH2
GST-SUZ12
GST-RBBP4

Figure S6

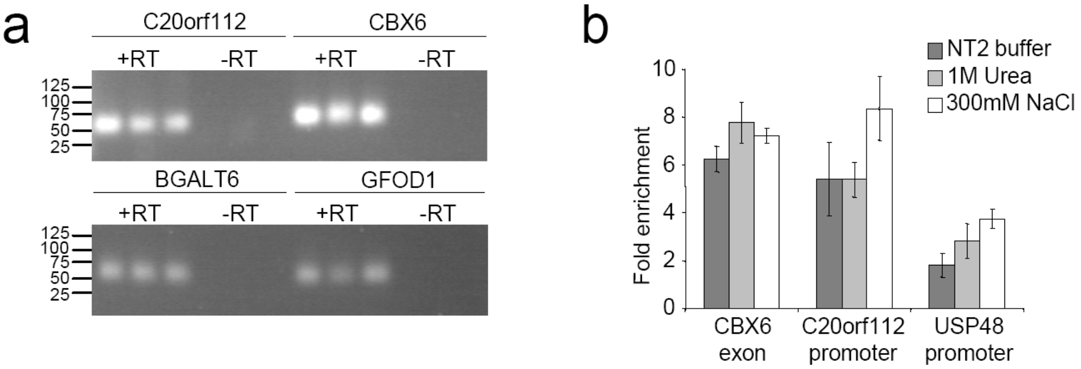
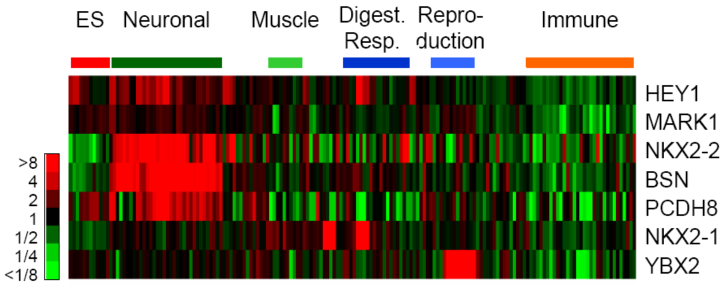


Figure S7



Supplemental Figure Legends

Figure S1. Replicate experiments identifying short RNA transcripts.
Related to Figure 1.

A. Summary of the experimental protocol for replicate experiments to detect short RNA transcripts. In both experiments, total RNA purified from CD4+ T cells was fractionated into long and short (<200 nt) populations and the short RNA poly-adenylated, reverse transcribed with oligo-dT and amplified and labeled by in vitro transcription. For the first strategy, Cy5-labeled short RNA was hybridised to custom microarrays (see Supplemental methods for details) together with mRNA labelled with Cy3. For the second strategy, Cy5-labeled short RNA was hybridised to arrays together with fragmented, poly-adenylated long RNA labeled with Cy3.

B. Typical Agilent Bioanlyzer electropherograms for total RNA and long and short RNA after fractionation.

C. Background-subtracted array signals for control snRNA and snoRNA probes in the long RNA channel (x-axis) and the short RNA channel (y-axis) for the array from our second strategy.

D. Distribution of short RNAs relative to the transcription start site of the average gene, using data from our second array experiment, plotted as a moving average with a window of 10 bp.

E. Venn diagram showing the overlap between the short RNAs identified using the two different methods outlined in A.

F. Venn diagram showing the overlap between the short RNAs identified in biological replicates of the first experimental strategy (mRNA reference).

G. Pie chart showing the short RNAs identified in this study that were previously identified in two other studies (sequences identified must be within 50bp of each other along the genome). 28% of short RNAs identified in this study were also identified by Kapranov and colleagues and 6% were also identified by Seila and colleagues.

Figure S2. Markers of transcriptional initiation at short RNA loci.
Related to Figure 2.

A. Composite enrichment profile of RNA pol II at genes for which no mRNA can be detected. The plot shows average tag density (ChIP-Seq data from Barski et al., 2007) and genes are divided into those that are associated with short RNA (green) and those not associated with short RNA (blue). The start and direction of transcription of the average gene is indicated by an arrow.

B. As for A. except genes are divided into those that detect a short RNA with probes positioned -100 to +100 (red), -700 to -500 (green) or +500 to +700 (blue) relative to the mRNA TSS.

C. Composite enrichment profile of RNA pol II from ChIP-Chip data at genes that express mRNA but not short RNA (red) and short RNA but not mRNA (green).

D. Composite enrichment profile of RNA pol II from ChIP-ChIP data at genes that express short RNAs and mRNA. Details as for B.

E. Heat maps showing enrichment of RNA polymerase II, H3K4me3 and H3K79me2 at genes that express short RNAs in the absence of detectable mRNA expression compared with those that do not express short RNA or mRNA. Each row represents one gene and each column represents the data from one oligonucleotide probe. Oligos are ordered by their position relative to the transcription start site, as shown by the diagram below. Fold enrichment is indicated by color, according to the scale on the right.

Figure S3. Enrichment of H3K27me3 at genes associated with short RNAs. Related to Figure 3.

A. Heat maps showing enrichment H3K27me3 at genes that do not express detectable mRNA. Genes are divided into those at which short RNAs can be detected and those where they cannot. Each row represents one gene and genes are ordered by the enrichment of H3K27me3. Each column represents the data from one oligonucleotide probe. Oligos are ordered by their position relative to the transcription start site, as shown by the diagram below. A scale for the enrichment ratios is shown on the left.

B. Percentage of H3K27-methylated genes at which short RNAs can be detected in the promoter, 5' exons or introns in replicate experiments.

C. Enrichment of short RNA loci by histone H3 methylated ChIP compared to total histone H3 ChIP, measured by quantitative PCR. Enrichment of ACTB is shown for comparison. Error bars indicate standard deviation (n=3).

D. Composite enrichment profile of H3K27me3 at genes for which no mRNA can be detected. The plot shows average tag density (ChIP-Seq data from Barski et al., 2007) and genes are divided into those that are associated with short RNA (green) and those not associated with short RNA (blue).

E. As D., except for H2A.Z.

F. As D., except for H3K9me1.

G. Examples of RNA polymerase II (green), H3K4me3 (blue) and H3K27me3 (red) ChIP signals at genes associated with short RNAs. The plots show unprocessed enrichment ratios for all probes within a genomic region (RNA Pol II ChIP vs. whole genomic DNA or methylated H3 ChIP vs total H3 ChIP). Chromosomal positions are from NCBI build 35 of the human genome. Genes are shown to scale below and aligned with the plots by chromosomal position (exons are represented by vertical bars, the start and direction of transcription by an arrow). The position of the short RNAs is indicated by the vertical arrow.

Figure S4. mRNA expression in Ezh2 and Ring1 conditional knockout cell lines. Related to Figure 4.

A. Quantitative PCR showing the expression (mean and SD, n=3) of Hes5, Msx1 and Ybx2 mRNA at timepoints after the addition of tamoxifen to the mES cell line Ezh2-1.3, which induces deletion of Ezh2. Expression is relative to day 0 and normalized to Actin.

B. As A., except for the line ES-ERT2, in which tamoxifen induces deletion of Ring1b. Expression is relative to cells harvested at the same timepoint but not treated with tamoxifen and normalized to Actin.

Figure S5. Predicted stem-loop structures encoded by short RNAs.
Related to Figure 5.
A. Predicted structure of stem-loop motifs within short RNA sequences. Genes proximal to each short RNA are named, along with the free energy of the structure.
B. Wild-type and mutated BSN stem-loop sequences used in EMSA experiments. Red indicates an altered nucleotide.
C. Longer exposure of the second panel of Figure 5E.

Figure S6. Enrichment of short RNAs by SUZ12 RNA IP.
Related to Figure 6.
A. Amplification products from SUZ12 IP cDNA no reverse transcriptase control reactions using primers specific for small RNAs (n=3 for each PCR).
B. Fold enrichment (mean and SD, n=3) of different RNA species in SUZ12 immunoprecipitate washed under 3 different conditions compared with input RNA and normalized to GAPDH.

Figure S7. Expression of mRNA from genes associated with short RNAs across different cell and tissue types.
Related to Figure 7.
Heat maps showing the tissue-specific expression of genes associated with H3K27me3 and short RNAs in CD4+ T cells for which northern blotting was performed in PBMC and neurons (Fig 7A). The figure was generated with expression data taken from 3 studies (Abeyta et al., 2004; Sato et al., 2003; Su et al., 2004). Each row represents one gene and each column one tissue or cell sample. Tissues are grouped according to cell type or organ system. Gene expression is shown relative to the median expression across all tissue types according to the scale on the left. *FOXN4* and *HES5* were not represented in the expression dataset.

Sequence class	*Number of sequences*	*Average number of probes per sequence*	*Total number of probes*
5’ end of RefSeq transcripts	18983	2.99	56833
3’ end of RefSeq transcripts	18563	2.91	53995
First introns	17505	2.39	41780
Proximal promoter regions (upstream of transcription start site)	17346	2.96	51345
miRNA stem loops	315	1	315
Genomic regions upstream and downstream of miRNA stem loops	320	8.42	5319
Non-coding RNAs	417	2.12	885
HIV genome	1	120	120
Agilent 1A v2 probes	20173	1	20173
Agilent hybridisation controls	NA	NA	2296
Total	NA	NA	233061

Table S1. Array design

Table S2. Array probes detecting short RNAs in replicate experiments. Due to its size, this table is provided as a separate file.

Gene	*Position of short RNA*	*Species*	*Result*	*Sequence*
5S rRNA	NA	Human (& mouse)	Yes	TTAGCTTCCGAGATCAGACGAGATCGGGCGCGTTCAGGGTGGTATGGCC
ASCL1	Intron	Human	No	GGAAAAGAACAGGAGAGGTTAATTTGAACGTGTAGGCTAGTGGTAGAGG
BSN	Exon	Human	Yes	GCGGTGCTCACACTCTCGGCGCCGCCGCTGCCGCCGCCATCTCCCAGCT
C20orf112	Promoter	Human	Yes	TGTCTGCGGTCAGGCCAGGAGAGGGAGACTTGGCCCAAATAAAGTGACT
FOXD1	Promoter	Human	No	TCACATGGTGTGCACGTCAGAGCGCTGCCGAGGGAAGGAAAGCAAGCCT
FOXN4	Intron	Human	Yes	CAATGCCCGGCATTGCCCGGGAGGAGGGAGCAAAGCCGACCCTGCAAGG
HES5	Promoter	Human	Yes	GATGCCGGGAGCCCCGCGCCTCAATATGCTGCCTTTTCCCAGGCCGCCA
Hes5	Promoter	Mouse	Yes	GATGCTGAGAGCCCCGCGCCTCAATATGCTGCCTTTTCCCAGGCCGCGG
HEY1	Intron	Human	Yes	CGCGGCAGGCCTGCGCTCGCCTCCCGCTCTGGCTCGGCTCCGCTCCGCC
HOXA7	Exon	Human	Yes	AGGCGAAGGCGCCGGCGCCCGCCCCGTAGCCGCTTCTCTGTGAGTTGGG
HOXC6	Intron	Human	Yes	GCCATTAGCACCAATTATTAGAGAGATCCCGAGTGCCCAGGACCCTCCC
MAPT	Exon	Human	Yes	GGTGGCAGCGGCGCTGCTGTTGGTGCCGGAGCTGGTGGGTGGCGGTGAC
MARK1	Exon	Human	Yes	GGCGCGAATGTCTCGGCTCGGTCCGCGCGGTCACAGCCACCGCCGCCGC
Msx1	Exon	Mouse	Yes	GCTTCCTGTGATCGGCCATGAGGGCCTCCACGCTGAAGGGCAGGAGTGA
NBEA	Intron	Human	No	CGGCGGGGGATCCGCAGGCTTATTCTCGGCGGTGGCGGTACCGCTAACC
NKX2-1	Exon	Human	Yes	ATGAGCGAGCGAGTCTGGGGACGAACCCTGGGGCCGCACTGTTGGTCTA
NKX2-2	Promoter	Human	Yes	GGAGGAGGGAAAAAATCCTCTTTAACATTCACCGGTTCCTACCTCCCCG
PAX3	Exon	Human	Yes	CGGCGAGCCGGGGAGCCTGGTGAGGCTGGAGCGCGGCCTGCCTGAGTCT
Pcdh8	Exon	Mouse	Yes	ATCCTCTTCGAACGTGCTGTATCGGACTGTCTTGCTCTGGGCCACTGAG
RASL12	Promoter	Human	Yes	TGGCACTGCCTCTTGCCTACAGCAATTCTGGGATTGGTGGTGAGCCTGG
SOBP	Exon	Human	No	AGTGGCTGTCACGGGTAAGGATAGTGCGGAGAGTCTCGGGGATGCCGCC
TBX3	Intron	Human	No	AGGGATGTCTGTGCTCCACTGAAAAATTCTGTCTGTTCGGGAGGGAACC
TMOD2	Intron	Human	No	TGCGGCCCAGCGGGCTGCAGAGGCTGCGGCACCGCAGTGCGGGGCGCGG
YBX2	Exon	Human (& mouse)	Yes	CTCATCCCGCCGGGTCCAGTACCGGCCACAGCCGCCACCGCCCCGGCCC

Table S3. Probes used for Northern blotting.
All sequences are 5’ to 3’. Result indicates if a short RNA species was detected in DNase-treated RNA.

RNA	***Sequence***
Xist-RepA	UUGCCCAUCGGGGCCACGGAUACCUGCU
Xist-RepA mut	UUGCGCAUCGAGGCCACAUACCUGCU
BSN	GGCGGCGCGAGCCGAGCUGGGAGAUGGCGG
BSN mut	GGCGGGGCGAGCAGAUGGGAGAUGGCGG
C20orf112	CUGCCCGAAGGGGCCCCGGGCGCCGAG
HEY1	GACAGCGAGCUGGACGAGACCAUCGAG
MARK1	CCGGCUCGGGCCGCUCCUCCUGACUGAGGC
PAX3	CCAGCCUCACCAGGCUCCCCGGCUCGCCGUG

Table S4. Probes used for EMSA.

RNA	***Forward***	***Reverse***	***Probe***
Xist RNA	GGGGCTGCGGATACCTGG	CGATGGGCCAGAGTGTTGG	SYBR Green
Actin mRNA	CAGCTCACCATGGATGATGATATC	AAGCCGGCCTTGCACAT	CCGCGCTCGTCGTCGACAAC
GAPDH mRNA	GGCTGAGAACGGGAAGCTT	AGGGATCTCGCTCCTGGAA	TCATCAATGGAAATCCCATCACCA
C20orf112 Short RNA	GTTCCTGAGGGTGCCTGAGT	AGGCCAGGAGAGGGAGACT	SYBR Green
GFOD1 Short RNA	CTAGCCGCATTGATGAGGTG	GCAGGTTAATGCACACCAAG	SYBR Green
CBX6 Short RNA	GAGAGGGAGCGTGAGCTGTA	GAGGAAAGTTTTGGGTTTGG	SYBR Green
B4GALT6 Short RNA	GACTCTCTGGCCTCCCATTC	CGAGCGGAAAAGAGGAAAT	SYBR Green
POU4F1 Short RNA	GTCTCCGAAACGCTGCAT	GGGACCTGCACACACCA	SYBR Green
KIAA0776 Short RNA	ACCCTCCCTCTCCTTTGTG	CCCAGATACAGGCGTGTCC	SYBR Green
USP48 Short RNA	AAGTGCCAGTGAGGACGTG	CGGGCTGAGGTAGAGGAAG	SYBR Green
TSPYL4 Short RNA	ACCGAGACCAGTGCCAAG	TGTTCGCCATCACCTGTGT	SYBR Green
CLSTN3 Short RNA	CCGTCCTGGGGTCTCTATTC	AGGTAGCCACCAGCGTGTAG	SYBR Green
SLC17A3 Short RNA	GTGTCCTGCCCAACATTCAG	GTGGTGGCTCTTGGTCTTCT	SYBR Green
U1 snRNA	CTCCGGATGTGCTGACCC	CAAATTATGCAGTCGAGTTTCCC	SYBR Green
U2 snRNA	ATGGATTTTTGGAGCAGGGA	GTCGATGCGTGGAGTGGAC	SYBR Green
U3 snRNA	CGTGTAGAGCACCGAAAACCA	GGCTTCACGCTCAGGAGAAA	SYBR Green
7SK RNA	AGAACGTAGGGTAGTCAAGC	AGAAAGGCAGACTGCCACAT	SYBR Green
5S rRNA	GATCTCGGAAGCTAAGCAGG	AAGCCTACAGCACCCGGTAT	SYBR Green
Luciferase	CACCATGGAAGATGCCAAAA	CCCGTCTTCGAGTGGGTAGA	CATTAAGAAGGGCCCAGCGCCA

Table S5. Primers used for quantitative PCR of immunoprecipitated RNA
All sequences are 5’ to 3’. HPRT1: Applied Biosystems Pre-developed assay 433768F

DNA	*Forward*	*Reverse*	*Probe*
ACTB	CAGCTCACCATGGATGATGATATC	AAGCCGGCCTTGCACAT	CCGCGCTCGTCGTCGACAAC
CDO1	CACGTCCATTCCTCCTCAG	GCCAAGTTCGACCAGTACA	SYBR Green
B4GALT6	GACAGGACGAAGAGAGGGAGA	CTGTGCTCAGGCGGATGAT	SYBR Green
CBX2	AACTTAAAGGGTGATTCTTCATGG	CTCCCTGTGATCGTCGTCAG	SYBR Green
SOBP	GGCAGGGGAAGAGGCTTT	CTAACTTTGGAGGGACCCTGA	SYBR Green
ELMOD1	CCTCTAAGCGGCGAGTTC	AACTTGATCTGAAGCAAGTGAGC	SYBR Green
HOXC8	GGGATGGCCCATGATTTATT	AAATAATTGATCTGCTGGAATGTG	SYBR Green
RNF182	GTGTGGCAGCAGCTGAGAT	TCTGATGTAAACACTGGATACACTTG	SYBR Green
LTR-luciferase A	AGGCCAATGAAGGAGAGAACAA	CTTTCTCCGCGTCCTCCAT	AGCTTGTTACACCCTATGAGCCTGCATGG
LTR-luciferase B	CACCATGGAAGATGCCAAAA	CCCGTCTTCGAGTGGGTAGA	CATTAAGAAGGGCCCAGCGCCA
LTR-luciferase C	CCACGCTGGGCTACTTGATC	GCAAGAATAGCTCCTCCTCGAA	SYBR Green
ACCN2	TCTTTGGAGATTTGGCAGTAAGG	TTGACCACTCAGATCCCATCCT	SYBR Green
HAPLN2	TCCCAACCCCAGCATCTTC	CCTGTGTCCAGCCCTGTGA	SYBR Green
HMX2	CCGGCCAGGTTTATGGAGTA	GGATTGCCTCAGGTAGGGATT	SYBR Green

Table S6. Primers used for quantitative PCR of H3K27me3 and H3 ChIP DNA. All sequences are 5’ to 3’.

DNA	*Forward*	*Reverse*	*Probe*
Ybx2	GGAACCAGCCAGCTCATACC	GGAGGTCGAAGGGAAACAAAA	SYBR Green
Msx1	CCAGCCCTATAGAAAGCAAGGA	CCCCTCAGAGCAATGCTTTG	SYBR Green
Hes5	TTTGTATGGGTGGGTGCATGT	AAGCCTTCAGAACAGCCTGTGT	SYBR Green
Pcdh8	ATTGGGATTTTATCTTTCACCAGAA	CCACAGACTCAAGATCTACAAGTTGTT	SYBR Green
Actin	TTGTCCCCCCAACTTGATGT	CCCTGGCTGCCTCAACAC	SYBR Green

Table S7. Primers used for quantitative RT-PCR of mouse mRNA. All sequences are 5’ to 3’

Supplemental Experimental Procedures

CD4+ T cell purification

Peripheral blood mononuclear cells were isolated from buffy coat using lymphoprep and CD4+ cells purified by magnetic selection (CD4+ isolation kit II, Miltenyi Biotech). Cells were routinely >95% positive for CD4 and <5% positive for CD69, an activation marker.

RNA purification

Cells were dissolved in Trizol (Invitrogen) and total RNA purified according to the manufacturer’s instructions. Total RNA was checked for degradation using an Agilent Bioanalyzer. Short RNA (<200bp) was purified from the total RNA using Ambion’s mirVana miRNA purification system. RNA size fractionation was confirmed using a Bioanalyzer (Figure S1)

Labelling RNA for microarray analysis

In our first experiment, 200ng of short RNA was poly-adenylated using polyA polymerase (Ambion). The addition of a poly-A tail was verified by quantifying the amount of cDNA synthesised from 5S RNA reverse transcribed using oligo-dT compared with random primers. The poly-adenylated short RNA was then reverse transcribed, in vitro transcribed and labelled with Cy5 using Agilent’s Low RNA Input Linear Amplification Kit. mRNA was also labeled with Cy3 using the same protocol. 4 μg of Cy5-labelled short RNA was mixed with 1 μg of Cy-3 labelled mRNA and hybridized to a custom 244K element microarray according the Agilent protocols, except for the addition of 40 μg of yeast tRNA. The arrays contained probes for the 5’ and 3’ ends of human RefSeq transcripts (see Array design, below).

In our second experiment, 500 ng of short RNA was poly-adenylated and then labelled with the Agilent protocol. To avoid labelling biases inherent in reverse transcribing long mRNA molecules from the 3’ end with the regular labelling protocol, we purified long RNA using the mirVana microRNA purification kit, fragmented this into approximately the same size range as the short RNA fraction by heating at 94^{o}C for 10 minutes in 20 mM magnesium acetate and 100 mM potassium acestate, and poly-adenylated the RNA. The RNA was then labeled with Cy3 with the Agilent protocol as before.

Array design

We designed a microarray containing probes to the 5’ and 3’ ends of human RefSeq transcripts, to proximal promoter regions just upstream of the transcription start site for each RefSeq transcript and to the first intron within each RefSeq transcript. Each transcript is represented by around 3 5’-end probes, 3 3’-end probes, 2-3 first intron probes and 3 promoter probes. The probes to RefSeq mRNAs were designed using ArrayOligoSelector (AOS) and are spaced approximately every 230 bp along the RNA sequence. Redundant probes designed from splice variants were removed. The array also includes the set of probes present on Agilent's Whole Human Genome expression array. These reported similar expression data to the 3'-end probes designed in-house. The promoter and intron probes were taken from our previously designed genomic oligo database used for our ChIP-Chip arrays and are spaced approximately every 190 bp (Lee et al., 2006). The array also contains probes to genomic regions around miRNA stem loops and to known non-coding RNAs designed using AOS. A summary of the array design is given in Supplemental Table S1.

Identification of probes detecting short RNAs

Image analysis and spot intensity measurements were carried out using Agilent Feature Extraction software. The software processes signal intensities by applying background subtraction and lowess normalization and uses normalized intensities to calculated log ratios. Log ratios and log ratio error are used to calculate z-statistics and p-values. The log ratio distribution of probes corresponding to known small RNAs was used to devise an algorithm for the prediction of novel small RNAs detected by promoter, 5' exon or first intron probes ($\log_2$ ratio >1.5, p-value <0.01, normalized signal >50). Only those probes predicted to bind short RNA using both arrays were considered for further analyses and these are listed in Supplemental Table S2. Genes were assigned as not being associated with a short RNA if all the promoter, 5' exon or first intron probes showed $\log_2$ ratio < 1.5 and Cy5 signals <20.

mRNA expression analysis

We considered a gene to produce mRNA if at least two probes to the 3' end of the mRNA reported a $\log_2$ ratio < -1.5 and p-value <0.01 in our first experiment. We considered a gene not to produce mRNA if the $\log_2$ ratio for all of the three 3' probes and the control Agilent Whole Human Genome expression array probe was > -1.5 and the Cy3 intensity was <20.

We also used absent/present calls from Affymetrix gene expression array analysis of CD4+ T cells (Su et al., 2004). We considered a gene to produce mRNA if the probe was called Present in both replicates and not to produce mRNA if called Absent in both replicates. To guard against inclusion of poor probes, we only considered probes that were able to detect mRNA expression in other tissue types.

ChIP-Chip

CD4+ T cells were purified as described above and crosslinked with 1% formaldehyde (by volume) for 20 minutes before the reaction was quenched by addition of glycine. ChIP was performed as described (Lee et al., 2006). Briefly, 10^8 cells were lysed for each ChIP experiment and sonicated on ice at 24W for 5 minutes total (pulses of 30s separated for gaps of 1 minute) with a Misonix Sonicator 3000.

The resulting whole cell extract was incubated overnight at 4°C with 100 µl of Dynal Protein G magnetic beads that had been preincubated with 10 µg of the appropriate antibody. The antibodies used for ChIP were as follows: anti-RNA pol II (8WG16, Abcam), anti-H3 (Abcam ab1791), anti-H3K4me3 (Abcam ab8580), anti-H3K79me2 (Abcam ab3594) and anti-H3K27me3 (Abcam ab6002).

The immunoprecipitation was allowed to proceed overnight. Beads were washed 5 times with RIPA buffer and once with TE containing 50 mM NaCl. Bound complexes were eluted from the beads by heating at 65°C for 1 hour with occasional vortexing and crosslinking was reversed by incubation at 65°C for a further 6 hours. Whole cell extract DNA (reserved from the sonication step) was also treated for crosslink reversal.

Immunoprecipitated DNA and whole cell extract DNA were then purified by treatment with RNAse A, proteinase K and phenol:chloroform:isoamyl alcohol extraction. Purified DNA was blunted and ligated to linker and amplified using a two-stage PCR protocol. Amplified DNA was labeled and purified using Bioprime random primer labeling kits (Invitrogen). Immunoenriched DNA was labeled with Cy5 fluorophore, whole cell extract DNA or histone H3 ChIP DNA was labeled with Cy3 fluorophore.

Labeled DNA was mixed with Herring sperm DNA, yeast tRNA and Cot-1 human DNA and hybridized to Agilent oligonucleotide microarrays in Agilent hybridization chambers for 40 hours at 40°C.

For our ChIP-Chip experiments, we designed a 2-slide set to cover 8 kb (approximately 4 kb upstream and 4 kb downstream) around the transcription start site of 18,450 RefSeq-annotated human genes and around the stem-loop of miRNA genes in mIRBase. Probes were designed against build 35 of the human genome sequence according to previously published criteria (Lee et al., 2006). Each array contains ~244,000 60-mer oligonucleotide probes. Oligonucleotide probes were spaced ~250 bp along the genome, on average.

Arrays were then washed and scanned using an Agilent DNA microarray scanner BA and the data extracted using Agilent Feature Extractor. We used the default CGH protocol with the following modifications: the background was set to the average intensity of all negative control spots, rank consistent probes was used only to calculate the normalization factor (linear normalization) and spacial detrending was turned off. We then calculated the log of the ratio of intensity in the IP-enriched channel to intensity in the genomic DNA channel for each probe and used a whole chip error model (Lee et al., 2006) to calculate confidence values for each spot on each array.

ChIP-Seq data analysis

ChIP-Seq data (Barski et al., 2007) for different histone modifications and RNA polymerase II binding in CD4$^+$ T cells were obtained in the form of summary BED files corresponding to each histone modification and RNA polymerase II. These files detailed number of sequence tags within 200bp windows of genomic DNA. The genes in each set were aligned with respect to Transcription start site (TSS). Average tag density within -4000 bp to 4000 bp with respect to TSS was calculated.

Analysis of previously published short RNA expression data

We compared short RNA data from this study with short RNA data from two previous publications on a gene-by-gene basis (Kapranov et al., 2007; Seila et al., 2008). The genomic coordinates of short RNAs (<200nt) from HepG2 cells detected by Kapranov and colleagues were mapped to RefSeq genes. If at least one short RNA was mapped within -1500bp to +1500bp of the TSS, the gene was considered to transcribe a short RNA. The genome coordinates of short RNAs sequenced by Seila and collegues were mapped to RefSeq genes. For comparisons of genes at which short RNAs were detected (Figure 1E), if at least one sense-strand short RNA was mapped within -1500bp to +1500bp of the RefSeq TSS, the gene was considered to transcribe a short RNA. For comparison of short RNAs themselves between studies (Figure S2F), a short RNA was judged to have been detected if the sequenced fragment (Seila et al., 2008) or the array probe (Kapranov et al., 2007) was within 50bp of short RNA detected in this study.

Murine ES cell culture

ES cell-derived motor neuron precursors were generated as described elsewhere (Wichterle et al., 2002; Wichterle and Peljto, 2008). v6.5 mouse ES cells were partially dissociated and cultured on Nunc Delta dishes in ADFNK medium optimized for ES cell to HB9+ motor neurons (Advanced DMEM/F12 and Neurobasal media (1:1 ratio), 0.1mM 2-mercaptoethanol, 2mM L-glutamine, 1 x Penicillin/Streptomycin, and 10% Knockout Serum Replacement). After two days, EBs were transferred to non-treated suspension culture dishes in ADFNK media supplemented with 1uM retinoic acid and 0.5ug/ml Hedgehog agonist Hh-Ag1.3. EBs were cultured for an additional 24 hours to achieve neural precursor cells, or 48 hours to achieve motor neuron precursors.

Ezh2 ko: The Ezh2-1.3 ES cell line was derived from a cross between mice carrying the floxed SET domain of Ezh2 (Su et. al., 2003) and mice carrying tamoxifen-inducible Cre at the Rosa locus (Cre-ERT2) and carries a conditional *Ezh2* mutation on both alleles that is induced by 4-hydroxy-tamoxifen (800nM). Loss of Ezh2 results in loss of H3K27me3 and a reduction in H3K27me2/me1 at day 4 post tamoxifen addition, without consequent changes in ES cell self-renewal. Cells were maintained on 0.1% gelatin-coated surfaces using KO-DMEM medium,10% FCS, 5% knock-out serum replacement (Invitrogen), non-essential amino acids, L-glutamine, 2-mercaptoethanol, antibiotics and 1000 U/ml of leukaemia inhibitory factor (ESGRO-LIF, Chemicon/Millipore).

Ring1b ko: We cultured ES-ERT2, an ES cell line, that carries a tamoxifen-inducible, conditional knockout of the core PRC1 protein Ring1B, and is also homozygous null for the functional homologue Ring1A. Following addition of tamoxifen, ES-ERT2 cells are progressively depleted of Ring1B protein and global H2A ubiquitination while PRC2 and H3K27me3 remain unchanged (Stock et al., 2007).

Northern Blotting

Short RNAs were isolated from PBMCs, neurons and ES cells as described above, treated with DNase-turbo (Ambion) and purified by ethanol precipitation. We found the use of non-DNase treated RNA could result in the detection of short DNA fragments in addition to short RNA species. Short RNA was resolved under denaturing conditions by running 10μg of sample (5μg for ES cell experiments) per well on 15% acrylamide-7M Urea TBE Novex gels (Invitrogen) at 200 V for 1 hr. The gel was electroblotted to Nytran Supercharge membranes (Whatman) at room temperature in 0.5X TBE with a starting current of 200mA for 2 hours. RNA was fixed to the membrane by UV-crosslinking followed by baking at 80°C for 1 hour. Probes were designed based on the sequences of array probes that were predicted to bind short RNAs (see above) and were associated with H3K4me3 and H3K27me3. All probes were designed to be 49 nucleotides in length. Sequences of all northern probes are provided in Supplemental Table S3. The probes were synthesized by Integrated DNA Technologies with the 3' StarFire extension and were radioactively labeled according to manufacturer's instructions. Blots were prehybridized in UltraHyb buffer (Ambion) and hybridization was carried out at 35^{o}C for 16 hrs. The membrane was then washed multiple times in 2x SSC + 0.5% SDS. The first three washes were carried out at 35^{o}C for 5 min and the last wash was carried out at 42^{o}C for 10 mins. The blots were then exposed to a PhosphoScreen (Molecular Dynamics) for one to two days and scanned using a Storm 860 phophoimager.

Western blotting

Ezh2-1.3 ko cells were boiled in Laemmli buffer and resolved on SDS-PAGE. Proteins were transferred to a nitrocellulose membrane (Hybond-ECL, GE Healthcare) and blotted with the antibodies against EZH2 (Millipore 07-689), H3 (Abcam ab1791-100), H3K27me3 (Diagenode CS-069-100) and anti-Rabbit IgG-HRP (GE Healthcare). Detection was performed with ECL or ECL Advance (GE Healthcare).

Gene Ontology

Gene ontology analysis was performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Dennis et al., 2003). Biological functions and protein domains enriched in the set genes that transcribed short RNA in the absence of mRNA were identified relative to all genes represented on the array. For comparison, an analysis was performed for genes that were not associated with short RNA or mRNA. Only selected categories with p-value $<10^{-3}$ and that significantly differ between two the gene lists are plotted in Figure 4E.

CpG analysis

It has been previously noted that there is a high correlation between polycomb binding sites and CpG islands (Ku et al., 2008). Given that short RNA loci were associated with H3K27me3, we considered that genomic regions encoding short RNAs might exhibit CpG density. To test this we defined the set of genes which expressed short RNAs in the absence of detectable mRNA and extracted short RNA sequences and an additional

100bp of sequence on each side. For each sequence, the ratio of the observed CpG density to the expected CpG density was calculated. Expected CpG was calculated as [C density]*[G density]. This ratio is reflective of CpG islands and has been recently shown to be much more useful in dividing promotes according to their CpG content (Saxonov et al., 2005, Weber et al., 2007). As a control, we used a set of probes which had same distribution with respect to the mRNA TSS but that did not detect any short RNA.

RNA secondary structure

A stem-loop-stem structure motif in Xist has been shown to interact with PRC2 (Zhao et al., 2008). We used RNAmotif (Macke et al., 2001)to perform a secondary structure motif search to detect similar motifs within 200 bp sequences centered on our array probes that detected short RNAs. The structures were evaluated using a nearest-neighbour energy function and only those motifs with free energy <= -6.5 kcal/mol were considered for further analysis. This cut-off was arrived based on the previously published structure motif (Zhao et al., 2008). We compared the results to a control set of probes that have the same distribution around the mRNA TSS but that didn't detect short RNA. RNA free energy structures were generated using RNAfold (Hofacker and Stadler, 2006).

Recombinant protein production

Full-length SUZ12 was amplified from the IMAGE clone and cloned into pGEX-4T1 (GE). Plasmids encoding GST-EED, GST-EZH2 and GST-RBBP4 were kindly provided by Dr.Y. Zhang. GST-fusion proteins were produced in *Escherichia coli* (BL-21 strain) and purified with Glutathione Sepharose 4B beads (GE Healthcare) according to the manufacturer's instructions. The fusion proteins were eluted from the beads with reduced glutathione (Sigma) and quantified by Coomassie staining after SDS-PAGE.

EMSA

Probes used in EMSA (Table S4) were end-labeled with [γ-^{32}P]ATP using T4-polynucleotide kinase (Promega) and purified with MicroBioSpin-6 columns (Bio-Rad). Nuclear extracts were prepared from the T cell line CEM using Nuclear Extract Kit (Active Motif). EMSAs were performed as described previously (Zhao et al. 2008 Science). 10 or 20 μg of nuclear extract 4 μg of GST fusion proteins were incubated with 0.5 pmol of radioactive probe for 30 minutes and protein-nucleotide complexes were resolved by acrylamide gel elecrophoresis. In competition experiments 200 pmol or 400 pmol of cold unlabelled probe was incubated with nuclear extract for 20 minutes before radioactive probe was added.

RNA immunoprecipitation

RNA associated with PRC2 was enriched from the T cell line CEM with an antibody to SUZ12 (Abcam) or unspecific rabbit antibody (Santa Cruz) following published protocols (Keene et al., 2006; Zhao et al., 2008). Cells were washed with 1x PBS and the pellet resuspend in an equal volume of cold lysis buffer (10mM Hepes, 100mM KCl, 5mM $MgCl_2$, 0.5% NP-40 supplemented with 1mM DTT, Complete protease inhibitor and RNaseOUT (100U/ml)). Lysis was allowed to proceed on ice for 5 minutes during which time the cells were passed 4 times through a 27-guage needle, before being frozen at -80°C. The lysate was then thawed on ice and diluted 10:1 in NT2 buffer (50mM Tris pH7.5, 150mM NaCl, 1mM $MgCl_2$, 0.05% NP-40, supplemented as before). DNA was degraded with Turbo DNase (Ambion) for 30 minutes at 4°C (30 units per $2x10^8$ cells) and insoluble material removed by centrifugation. After preclearing with protein-G beads (Dynal), PRC2 was immunoprecipitated overnight at 4°C with protein-G magnetic beads pre-bound with antibody. The beads were washed 6 times in NT2 buffer (supplemented with 1M urea or 300mM NaCl for certain experiments) and the RNA extracted from the IPs and input material with Trizol-LS. The purified RNA was then treated with Turbo DNase, precipitated and reverse transcribed with SuperScript III (Invitrogen). For endogenous genes, cDNAs in the IP samples were quantified by SYBR-green real-time PCR against a dilution series of cDNA reverse transcribed from 1 μg of input RNA and fold enrichment calculated relative to cDNA reverse transcribed from an amount of input RNA comparable to mass purified in the IP. The primers used are listed in Supplemental Table S5. For luciferase, enrichment was quantified relative to input RNA using 2^dCt.

Luciferase assay

The R and U5 regions of the HIV LTR from the SF2 molecular clone were replaced with DNA encoding the murine Xist-RepA stem-loop, a mutant Xist-RepA stem-loop and the stem-loop in the C20orf112 short RNA by PCR and the wild-type and modified LTRs cloned in place of the CMV promoter in pIRESneo3 (Clontech) and sequence verified. The firefly luciferase coding sequence from pCSFLW was then cloned upstream of the LTR. Plasmid DNA from each construct was transfected into Hela cells in triplicate together with the Renilla luciferase plasmid phRL-null. 48 hours later, firefly luciferase activity was measured in relation to Renilla luciferase activity with the Dual-luciferase reporter system (Promega) using a Glomax luminomter (Promega). The experiment was repeated 7 times, 5 times with one set of clones and twice with another set. The significance of differences in the ratio of firefly to renilla luciferase activity were estimated with a one-sided paired T-test.

Luciferase construct RNA IP and ChIP

Hela cells were transfected with each construct and cells harvested after 3 days and processed for RNA IP and ChIP using the methods described above. The H3K27me3/H3 ratio was calculated (2^dCt) for each construct and divided by that measured for the wild-type LTR construct. The primers used are listed in Table S6. Their respective positions in the luciferase construct are as follows:

LTR-luciferase B primers are -280bp relative to the start of the added short RNA sequence.

LTR-luciferase B primers are +3 relative to the end of the added short RNA sequence.
LTR-luciferase C primers are +758 relative to the end of the added short RNA sequence.

Enrichment of 3 endogenous loci previously shown to be associated with H3K27me3 in Hela cells were also measured (Cuddapah et al., 2009) and the data for these 3 genes averaged.

Quantitative reverse-transcription PCR of mRNA

Around 500ng of total RNA was reverse transcribed using oligo-dT primer (Promega). mRNA expression during the ES cell differentiation time course was measured by quantitative RT-PCR using SYBRgreen. Change in the expression of Hes5, Pcdh8, Ybx2 and Msx1 were calculated relative to day 0 for the neuronal differentiation timecourse and for the Ezh2 deletion timecourse or to no tamoxifen control samples for the Ring1b deletion timecourse. Expression changes were then normalized to Actin (2^ddCt). Primers used are listed in Supplemental Table 7.

Supplemental References

Abeyta, M.J., Clark, A.T., Rodriguez, R.T., Bodnar, M.S., Pera, R.A., and Firpo, M.T. (2004). Unique gene expression signatures of independently-derived human embryonic stem cell lines. Human molecular genetics *13*, 601-608.

Cuddapah, S., Jothi, R., Schones, D.E., Roh, T.Y., Cui, K., Zhao, K. (2009). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Res. *19*:24-32.

Dennis, G., Jr., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., and Lempicki, R.A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. *4*, P3

Keene, J.D., Komisarow, J.M., and Friedersdorf, M.B. (2006). RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. Nature protocols *1*, 302-307.

Sato, N., Sanjuan, I.M., Heke, M., Uchida, M., Naef, F., and Brivanlou, A.H. (2003). Molecular signature of human embryonic stem cells and its comparison with the mouse. Developmental biology *260*, 404-413.

Saxonov, S., Berg, P., Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A. *103*,1412-1417.

Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Pääbo, S., Rebhan, M., Schübeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet. *39*, 457-466.

Wichterle, H. and Peljto, M. (2008). Differentiation of mouse embryonic stem cells to spinal motor neurons. Curr. Protoc. Stem Cell Biol. *5*,1H.1.1-1H.1.9.