Genomic positional conservation identifies topological anchor point (tap)RNAs linked to developmental loci

Paulo P. Amaral[1*], Tommaso Leonardi[2,3*], Namshik Han[1*], Emmanuelle Viré[1†], Dennis Gascoigne[1], Raúl Arias-Carrasco[4], Magdalena Büscher[1], Anda Zhang[5], Stefano Pluchino[3], Vinicius Maracaja-Coutinho[4], Helder I. Nakaya[6], Martin Hemberg[1,7], Ramin Shiekhattar[5], Anton J. Enright[2] and Tony Kouzarides[1**]

[1] *The Gurdon Institute, University of Cambridge, Tennis Court Road, Cambridge, CB2 1QN, UK.*
[2] *EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK.*
[3] *Department of Clinical Neurosciences; Wellcome Trust-Medical Research Council Stem Cell Institute, University of Cambridge, Clifford Allbutt Building-Cambridge Biosciences Campus, Hills Road, Cambridge, CB2 0PY, UK.*
[4] *Centro de Genómica y Bioinformática, Facultad de Ciencias, Universidad Mayor, Camino La Pirámide 5750, Huechuraba, Santiago, 8580000, Chile*
[5] *University of Miami Miller School of Medicine, Sylvester Comprehensive Cancer Center, Department of Human Genetics, Biomedical Research Building, Miami, FL 33136, USA.*
[6] *School of Pharmaceutical Sciences, University of São Paulo, Av. Prof. Lineu Prestes 580, São Paulo 05508, Brazil.*
[7] *Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, UK.*
[†] *Present address: MRC Prion Unit, UCL Institute of Neurology, Queen Square House, Queen Square, London WC1N 3BG.*

[*] *Equal first authors*

[**] *Correspondence: t.kouzarides@gurdon.cam.ac.uk*

April 28, 2016

Contents

1 Supplementary Figures

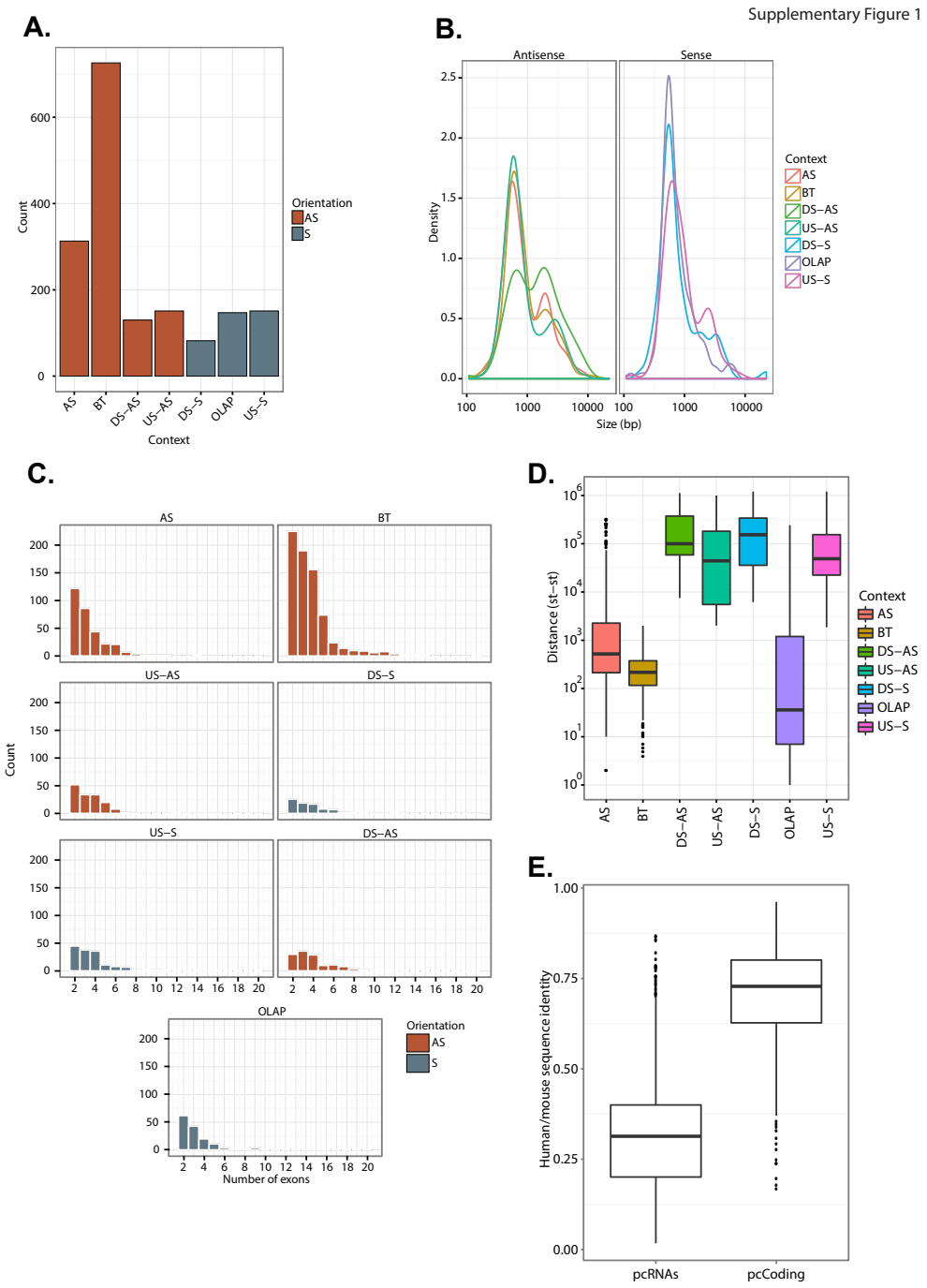


Figure 1: ***A****: Bar chart showing the number of pcRNAs in each orientation.* ***B****: Density distribution of the distance between pcRNAs and respective coding genes, color-coded by positional orientation. The left plot shows pcRNA in antisense orientations, while the right plot shows pcRNAs in sense orientations.* ***C:*** *Bar chart showing exon-number distribution for each pcRNA.* ***D****: Boxplot showing the distribution of the distances between the TSS of pcRNAs and the TSS of their corresponding coding gene.* ***E:*** *Boxplot showing the fraction of sequence identity between human and mouse pcRNAs and human and mouse pcRNA-associated protein coding genes. Sequence identity was calculated with the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970).*

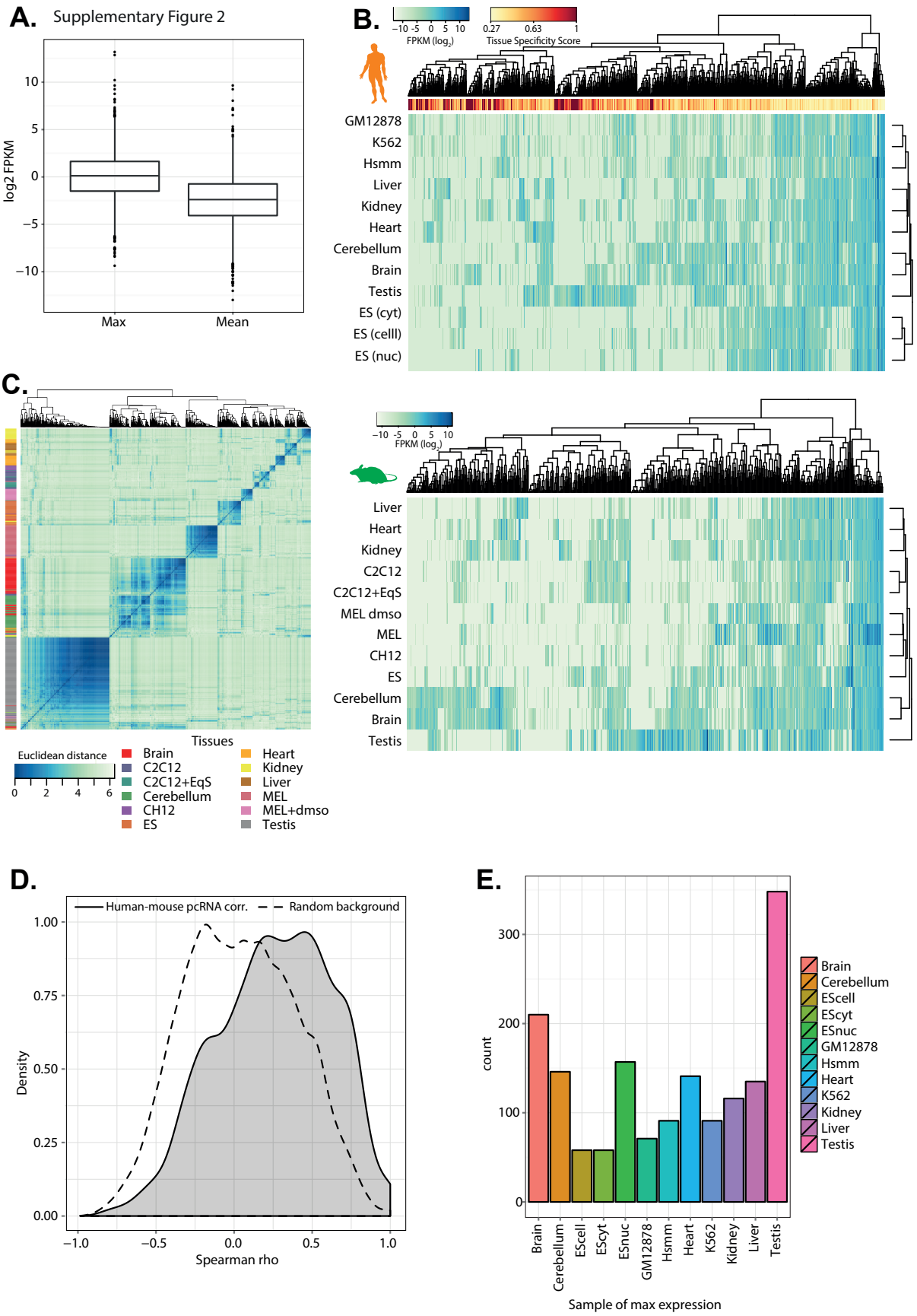

A.
Supplementary Figure 2
log2 FPKM
Max
Mean
B.
FPKM (log2)
Tissue Specificity Score
GM12878
K562
Hsmm
Liver
Kidney
Heart
Cerebellum
Brain
Testis
ES (cyt)
ES (celll)
ES (nuc)
C.
Euclidean distance
Tissues
Brain
C2C12
C2C12+EqS
Cerebellum
CH12
ES
Heart
Kidney
Liver
MEL
MEL+dmso
Testis
Liver
Heart
Kidney
C2C12
C2C12+EqS
MEL dmso
MEL
CH12
ES
Cerebellum
Brain
Testis
D.
Human-mouse pcRNA corr.
Random background
Density
Spearman rho
E.
count
Brain
Cerebellum
EScell
EScyt
ESnuc
GM12878
Hsmm
Heart
K562
Kidney
Liver
Testis
Sample of max expression

Figure 2: ***A:*** *Boxplot showing the distribution of the highest FPKM measured across all samples for each pcRNA (left) and the mean FPKM across all samples for each pcRNA (right).* ***B:*** *Heatmap showing the expression profiles of human pcRNAs (top) and mouse pcRNAs (bottom) across tissues and cell lines. The vertical sidebar reports the tissue specificity score of pcRNAs, ranging from 0.27 (white) to 1 (red).* ***C:*** *Heatmap showing the Euclidean distance between the expression profiles of mouse pcRNAs. The vertical sidebar reports the tissues in which each pcRNA has maximal expression.* ***D:*** *Density distribution of the Spearman's correlation coefficients between human and mouse pcRNA pairs. Mean Spearman's rho between human and mouse 0.26, permutation test p-value $<10^{-6}$. The dotted line shows the background distribution of all pairwise Spearman's correlations between human and mouse pcRNA.* ***E:*** *Bar chart showing the number of pcRNAs (y-axis) detected to have the highest expression in each given tissue (x-axis).*

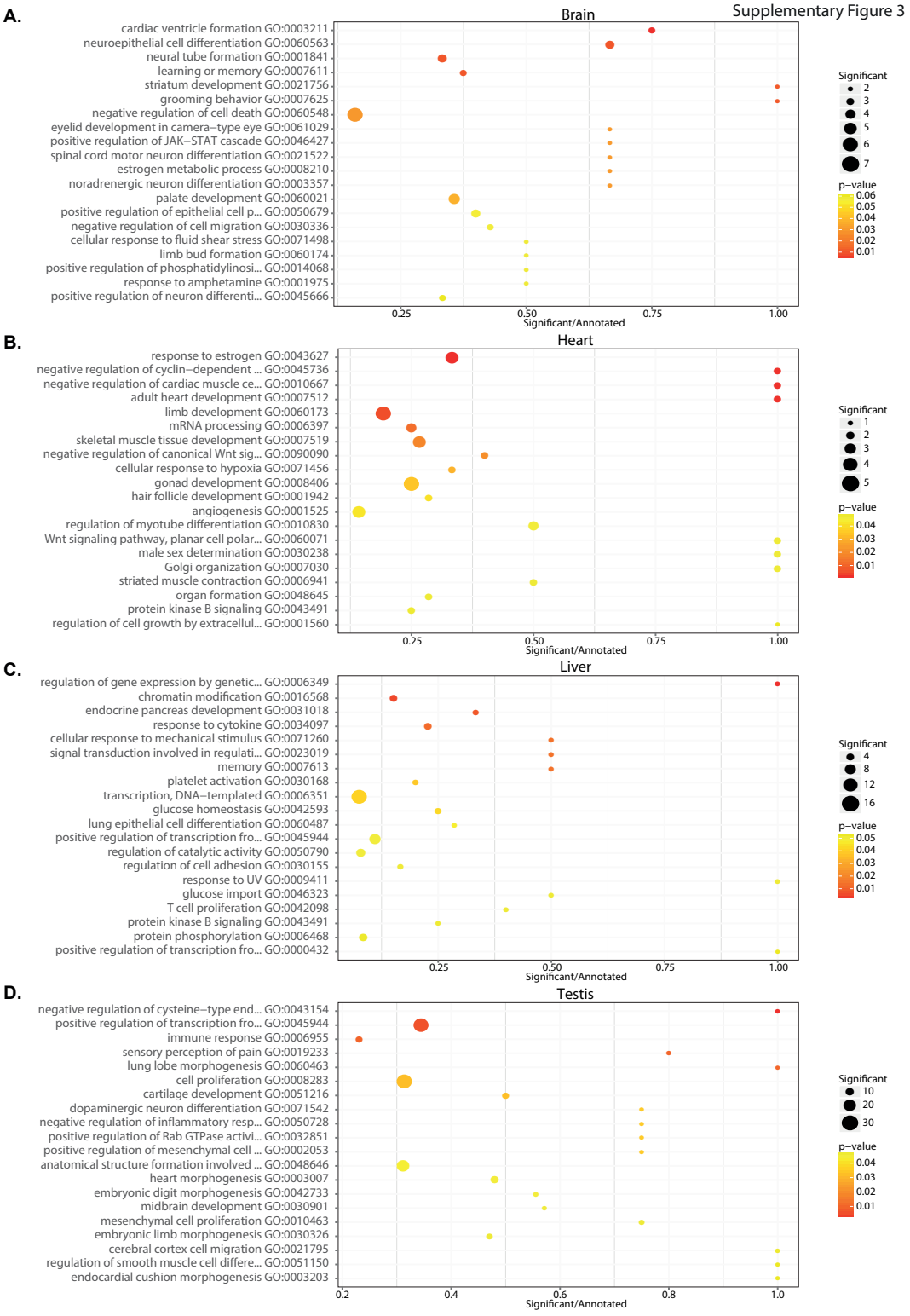


Figure 3: ***A-D:*** *GO enrichment analysis of coding genes associated to pcRNAs with expression specific for Brain (A), Heart (B), Liver (C), Testis (D). The x-axis shows the enrichment score, calculated as the number of pcRNA-associated genes in a given GO category divided by the total number of genes in the category. The size of the points indicates the absolute number of pcRNA-associated genes in the given GO category. The color-coding indicates the adjusted p-value.*

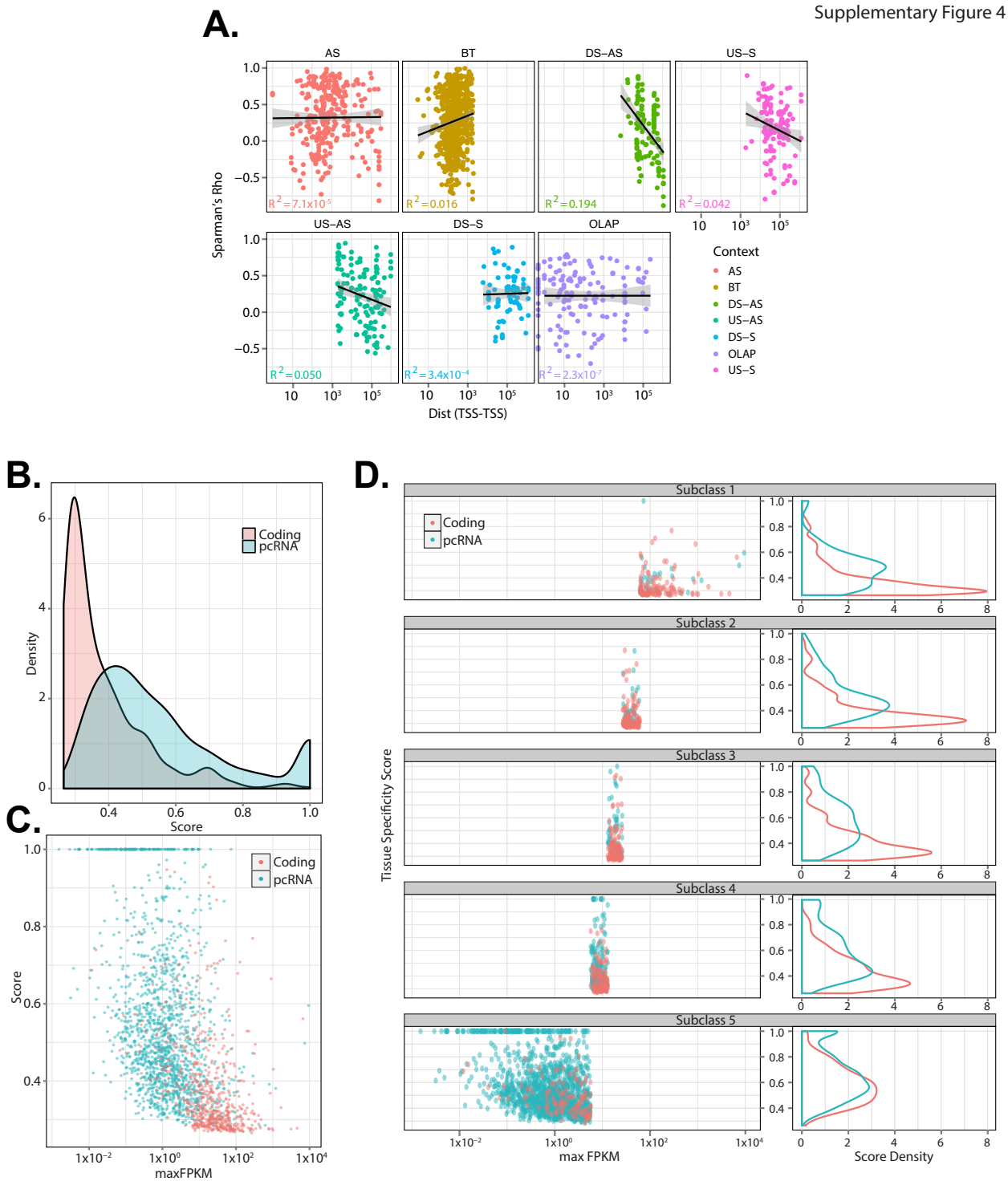

A.
AS
BT
DS–AS
US–S
US–AS
DS–S
OLAP
Sparman's Rho
Dist (TSS-TSS)
Context
AS
BT
DS–AS
US–AS
DS–S
OLAP
US–S
B.
Coding
pcRNA
Density
Score
C.
Score
maxFPKM
D.
Subclass 1
Subclass 2
Subclass 3
Subclass 4
Subclass 5
Tissue Specificity Score
max FPKM
Score Density

Figure 4: ***A:*** *Plot showing the Spearman correlation coefficient between the expression of pcRNAs and their corresponding coding genes as a function of their distance (TSS to TSS), indicating independence of TSS to TSS distance (R^2=0.008, p-value $3.23x10^{-4}$). The black lines represent the linear fit.* ***B****: Density distribution of the Tissue Specificity Score (see Supplementary Methods) for pcRNAs (blue) and pcRNA-associated coding genes (red) showing significant higher specificity for pcRNAs (mean pcRNA tissue specificity score 0.55, mean associated coding gene tissue specificity score 0.37, p-value* 4.25×10^{-220}*, Wilcoxon test).* ***C:*** *Scatterplot showing the highest FPKM observed across tissues (x-axis) for pcRNAs (blue) and pcRNA-associated coding genes (red) plotted against their tissue specificity score (y-axis).* ***D:*** *Scatterplot and density distribution of Tissue Specificity Scores for pcRNAs (blue) and pcRNA-associated coding genes (red) divided into 5 expression sub-groups. Each of the five sub-plots only displays pcRNAs and coding genes with similar expression levels (see Supplementary Methods) and shows the highest FPKM observed across tissues (x-axis) plotted against their tissue specificity score (y-axis). The right part of the plot shows the distribution of tissue specificity scores for each sub-group, showing that pcRNAs have higher tissue specificity score then pcRNA-associated coding genes independently of their expression level.*

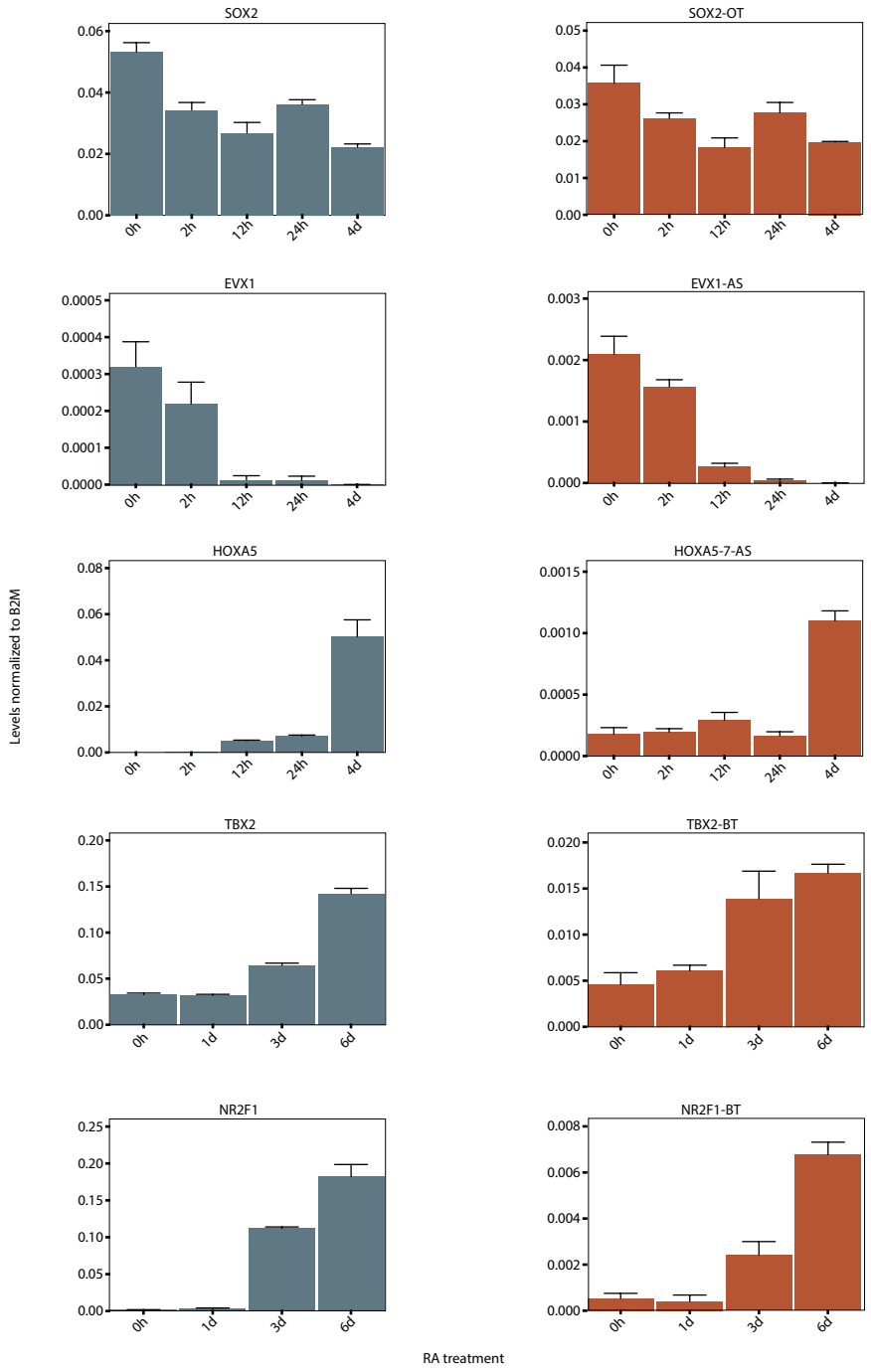
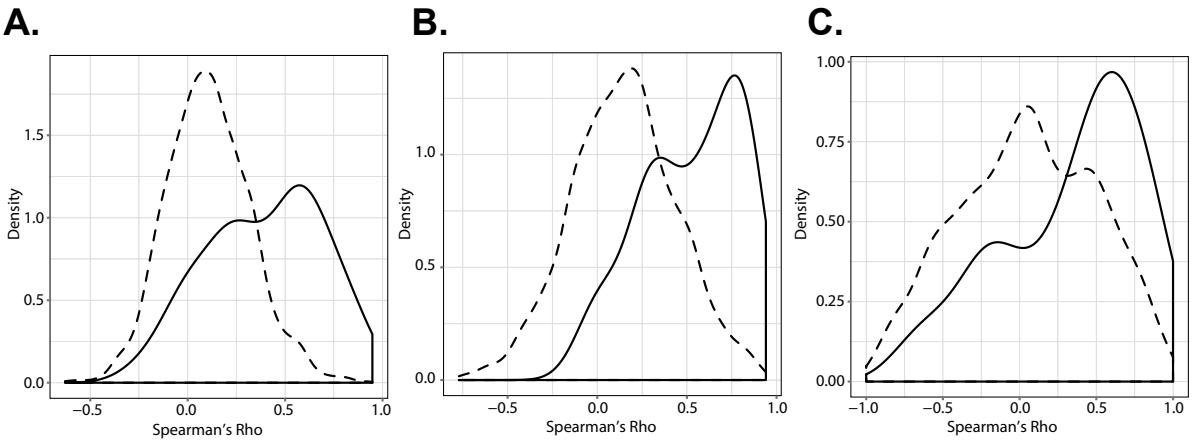
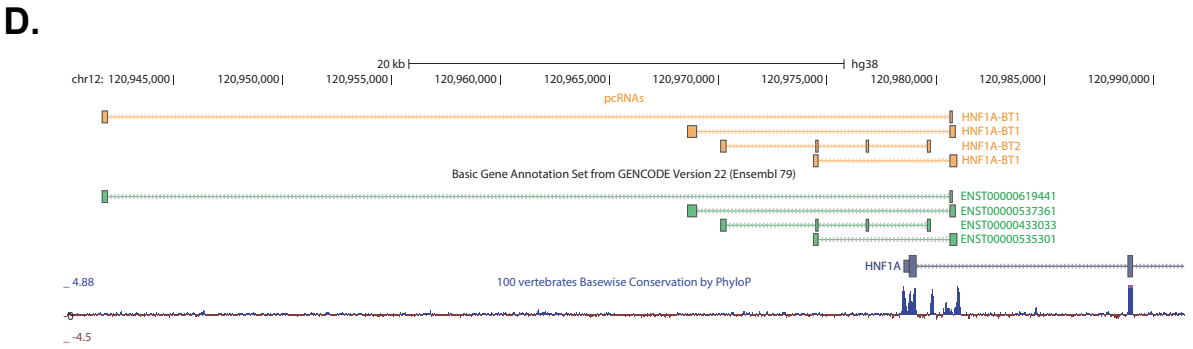
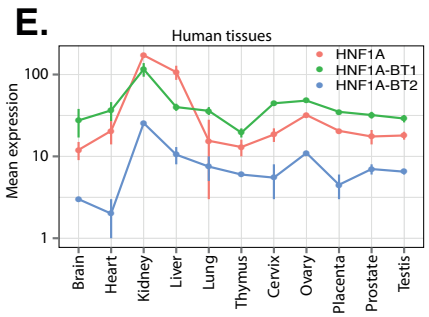
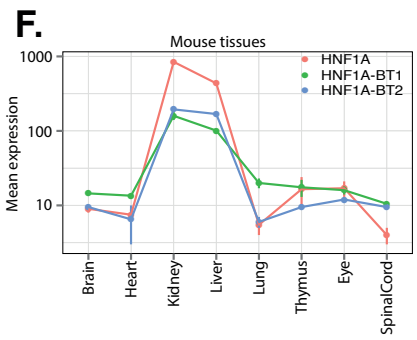


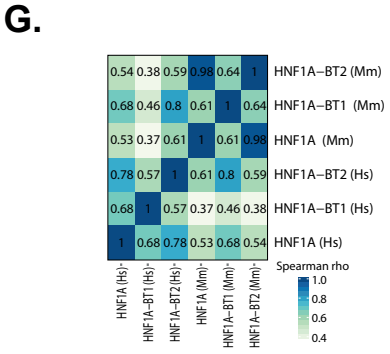
Figure 5: *Real time PCR data showing the expression of* SOX2, EVX1, HOXA5, TBX2, NR2F1 *(left) and associated pcRNAs (right) over 5 time-points of NT2 cells differentiation with retinoic acid (RA). The data is expressed relative to the expression of* B2M*; the error bars indicate the standard error of the mean (SEM) across two replicate experiments.*


A.
B.
C.
Density
Spearman's Rho


D.
20 kb
hg38
pcRNAs
HNF1A-BT1
HNF1A-BT1
HNF1A-BT2
HNF1A-BT1
Basic Gene Annotation Set from GENCODE Version 22 (Ensembl 79)
ENST00000619441
ENST00000537361
ENST00000433033
ENST00000535301
HNF1A
100 vertebrates Basewise Conservation by PhyloP


E.
Human tissues
HNF1A
HNF1A-BT1
HNF1A-BT2
Mean expression
Brain
Heart
Kidney
Liver
Lung
Thymus
Cervix
Ovary
Placenta
Prostate
Testis


F.
Mouse tissues
HNF1A
HNF1A-BT1
HNF1A-BT2
Mean expression
Brain
Heart
Kidney
Liver
Lung
Thymus
Eye
SpinalCord


G.
HNF1A–BT2 (Mm)
HNF1A–BT1 (Mm)
HNF1A (Mm)
HNF1A–BT2 (Hs)
HNF1A–BT1 (Hs)
HNF1A (Hs)
Spearman rho

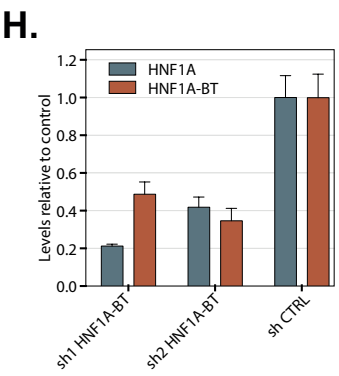

H.
HNF1A
HNF1A-BT
Levels relative to control
sh1 HNF1A-BT
sh2 HNF1A-BT
sh CTRL

Figure 6: ***A-B:*** *Density distribution of the Spearman's correlation coefficients calculated from Nanostring data for human (**A**) and mouse (**B**) pcRNAs and corresponding coding genes showing significant positive correlation (mean Spearman's rho 0.40 and 0.53 for human and mouse respectively, permutation test p-values $<10^{-6}$). The dotted line shows the background distribution of all pairwise Spearman's correlations between pcRNAs and pcRNA-associated coding genes.* ***C:*** *Density distribution of the Spearman's correlation coefficients on Nanostring data between human and mouse pcRNAs pairs, showing conserved expression profiles across species (mean Spearman's rho 0.33, permutation test p-value $<10^{-6}$).* ***D:*** *Illustration of the* HNF1A *locus modified from a screenshot of the UCSC genome browser. For clarity, only one representative isoform of the coding gene is displayed.* ***E,F:*** *Nanostring expression profiles of* HNF1A *and HNF1A-associated pcRNAs across human (**E**) and mouse (**F**) tissues. The plots report the mean value of two technical replicates, while the error bars report the value of each replicate.* ***G:*** *Heatmap showing Spearman's correlation coefficients between human and mouse* HNF1A, HNF1A–BT1 *and* HNF1A–BT2. ***H:*** *Real Time PCR data showing the expression of* HNF1A *and* HNF1A–BT *in HepG2 cells upon knock-down of* HNF1A–BT. *Sh1- and sh2- HNF1A–BT indicate two different, non-overlapping shRNAs designed against HNF1A–BT. The data is expressed relative to the expression of the control transfected with scrambled shRNAs; the error bars indicate the SEM across two replicate experiments.*

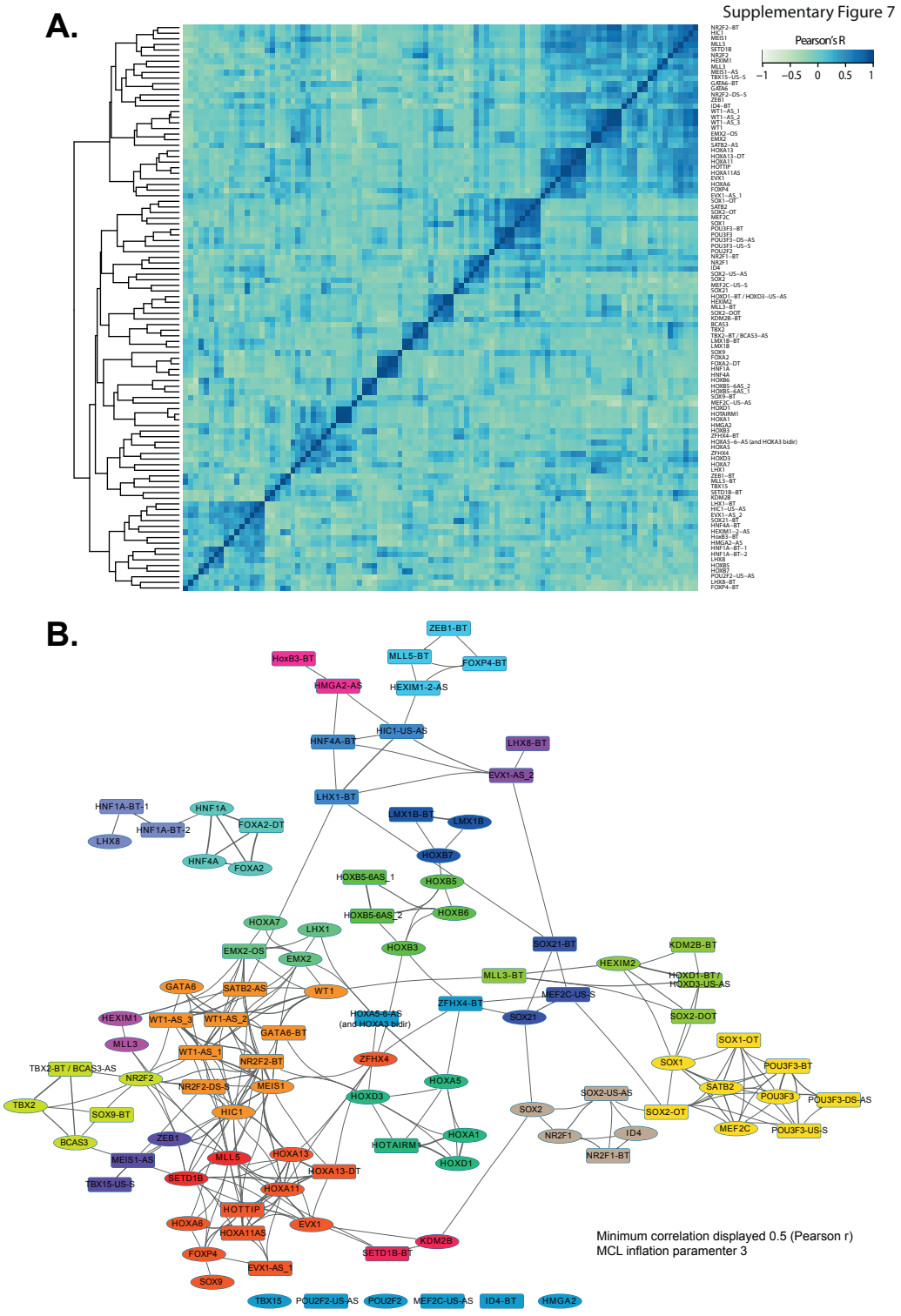


Figure 7: ***A****: Heatmap showing the pairwise Pearson correlation coefficients between all human transcripts included in the Nanostring experiment (both pcRNAs and pcRNA-associated coding genes).* ***B:*** *Network displaying all human transcripts included in the Nanostring experiment (nodes) and the Pearson correlation coefficient between their expression profiles (edges). Only edges with correlation coefficient higher than 0.5 are shown. The color coding of the nodes indicates the result of applying the Markov Clustering Algorithm to the matrix of correlation coefficients (see Supplementary Methods).*

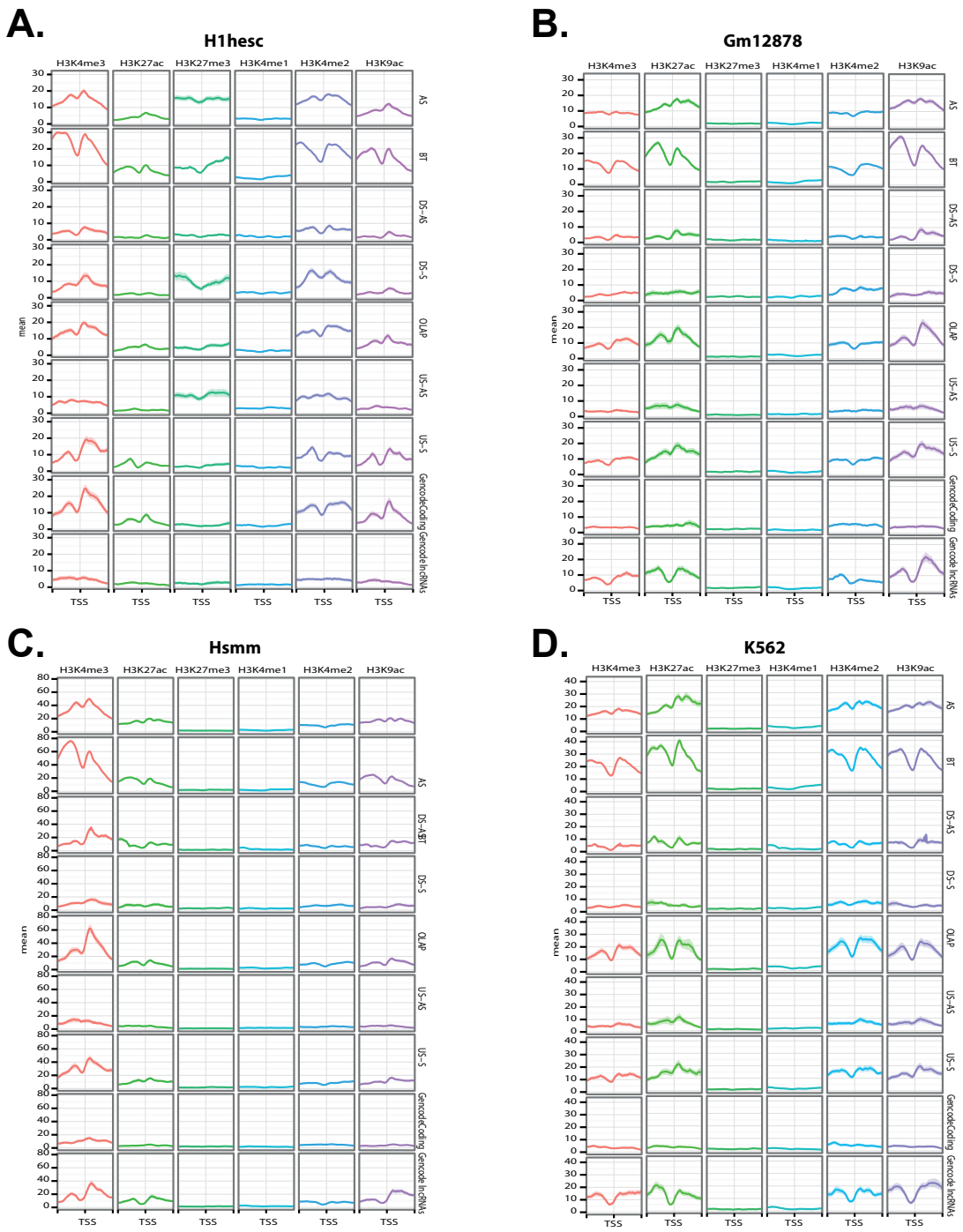


Figure 8: *Histone modification profiles of pcRNA promoters (split by their relative orientation), promoters of 1000 random Gencode lncRNAs and promoters of 1000 random Gencode coding genes based on ChIP-Seq data by the ENCODE project on H1-hESCs (**A**), GM12878 (**B**), HSMM (**C**) and K562 (**D**). The lines represent the mean ChIP-Seq coverage and the shaded area around the line represents the standard deviation of the mean.*

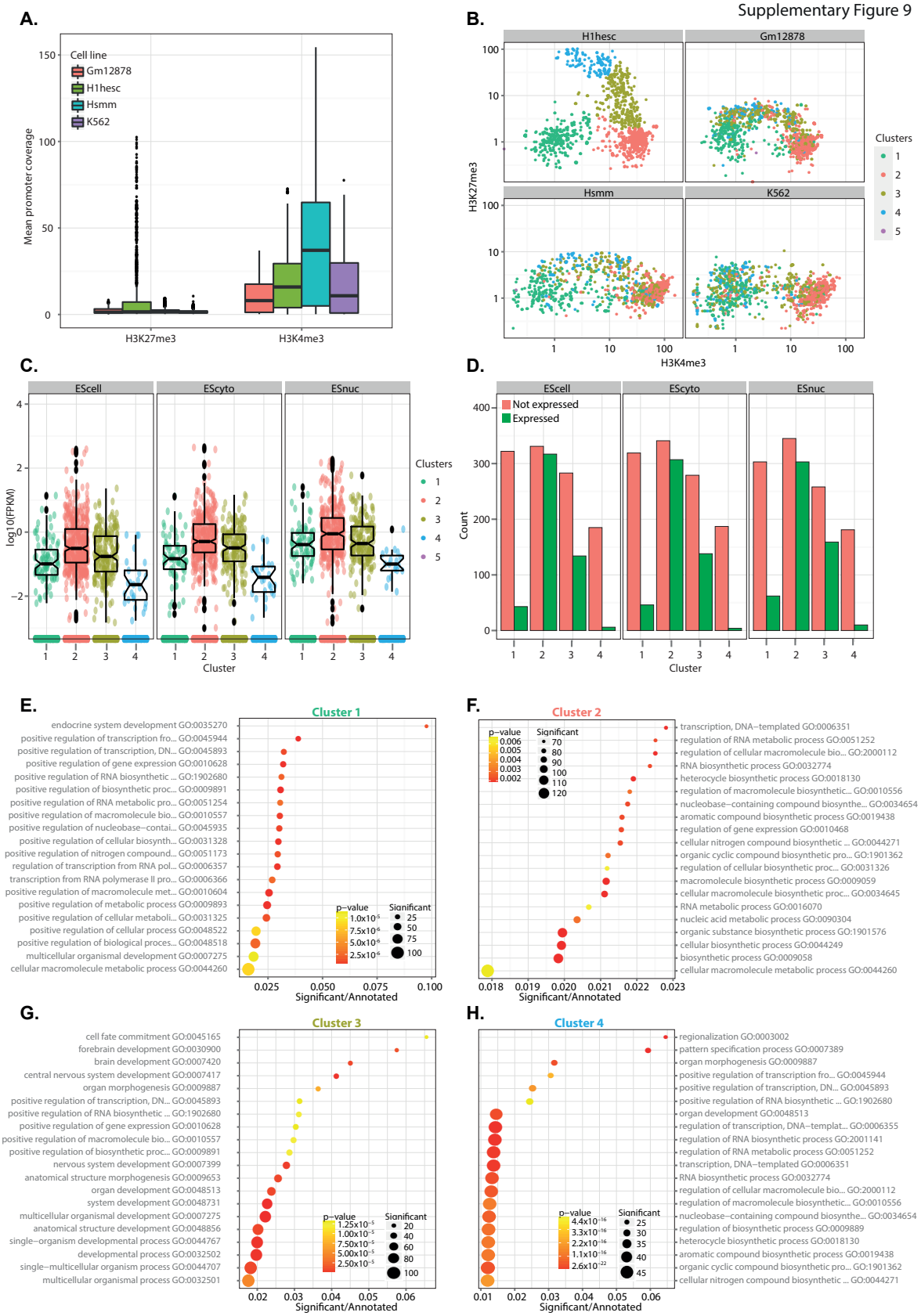

A.
Cell line
Gm12878
H1hesc
Hsmm
K562
Mean promoter coverage
H3K27me3
H3K4me3
B.
H1hesc
Gm12878
Hsmm
K562
H3K27me3
H3K4me3
Clusters
C.
EScell
EScyto
ESnuc
log10(FPKM)
Cluster
Clusters
D.
Not expressed
Expressed
Count
Cluster
E.
Cluster 1
endocrine system development GO:0035270
positive regulation of transcription fro... GO:0045944
positive regulation of transcription, DN... GO:0045893
positive regulation of gene expression GO:0010628
positive regulation of RNA biosynthetic ... GO:1902680
positive regulation of biosynthetic proc... GO:0009891
positive regulation of RNA metabolic pro... GO:0051254
positive regulation of macromolecule bio... GO:0010557
positive regulation of nucleobase−contai... GO:0045935
positive regulation of cellular biosynth... GO:0031328
positive regulation of nitrogen compound... GO:0051173
regulation of transcription from RNA pol... GO:0006357
transcription from RNA polymerase II pro... GO:0006366
positive regulation of macromolecule met... GO:0010604
positive regulation of metabolic process GO:0009893
positive regulation of cellular metaboli... GO:0031325
positive regulation of cellular process GO:0048522
positive regulation of biological proces... GO:0048518
multicellular organismal development GO:0007275
cellular macromolecule metabolic process GO:0044260
p−value
Significant
Significant/Annotated
F.
Cluster 2
transcription, DNA−templated GO:0006351
regulation of RNA metabolic process GO:0051252
regulation of cellular macromolecule bio... GO:2000112
RNA biosynthetic process GO:0032774
heterocycle biosynthetic process GO:0018130
regulation of macromolecule biosynthetic... GO:0010556
nucleobase−containing compound biosynthe... GO:0034654
aromatic compound biosynthetic process GO:0019438
regulation of gene expression GO:0010468
cellular nitrogen compound biosynthetic ... GO:0044271
organic cyclic compound biosynthetic pro... GO:1901362
regulation of cellular biosynthetic proc... GO:0031326
macromolecule biosynthetic process GO:0009059
cellular macromolecule biosynthetic proc... GO:0034645
RNA metabolic process GO:0016070
nucleic acid metabolic process GO:0090304
organic substance biosynthetic process GO:1901576
cellular biosynthetic process GO:0044249
biosynthetic process GO:0009058
cellular macromolecule metabolic process GO:0044260
p−value
Significant
Significant/Annotated
G.
Cluster 3
cell fate commitment GO:0045165
forebrain development GO:0030900
brain development GO:0007420
central nervous system development GO:0007417
organ morphogenesis GO:0009887
positive regulation of transcription, DN... GO:0045893
positive regulation of RNA biosynthetic ... GO:1902680
positive regulation of gene expression GO:0010628
positive regulation of macromolecule bio... GO:0010557
positive regulation of biosynthetic proc... GO:0009891
nervous system development GO:0007399
anatomical structure morphogenesis GO:0009653
organ development GO:0048513
system development GO:0048731
multicellular organismal development GO:0007275
anatomical structure development GO:0048856
single−organism developmental process GO:0044767
developmental process GO:0032502
single−multicellular organism process GO:0044707
multicellular organismal process GO:0032501
p−value
Significant
Significant/Annotated
H.
Cluster 4
regionalization GO:0003002
pattern specification process GO:0007389
organ morphogenesis GO:0009887
positive regulation of transcription fro... GO:0045944
positive regulation of transcription, DN... GO:0045893
positive regulation of RNA biosynthetic ... GO:1902680
organ development GO:0048513
regulation of transcription, DNA−templat... GO:0006355
regulation of RNA biosynthetic process GO:2001141
regulation of RNA metabolic process GO:0051252
transcription, DNA−templated GO:0006351
RNA biosynthetic process GO:0032774
regulation of cellular macromolecule bio... GO:2000112
regulation of macromolecule biosynthetic... GO:0010556
nucleobase−containing compound biosynthe... GO:0034654
regulation of biosynthetic process GO:0009889
heterocycle biosynthetic process GO:0018130
aromatic compound biosynthetic process GO:0019438
organic cyclic compound biosynthetic pro... GO:1901362
cellular nitrogen compound biosynthetic ... GO:0044271
p−value
Significant
Significant/Annotated

Figure 9: *We identified an embryonic stem cell specific signature of pcRNA promoters with high levels of both H3K27me3 and H3K4me3 (bivalent promoters) or high levels of tri-methylation of H3K27 (H3K27me3) and intermediate levels of H3K4me3 (**A,B**). The pcRNAs clustered in these groups show intermediate level or no expression in ES cells, respectively (**C,D**). Whereas both clusters are associated with developmental genes, the bivalent cluster is particularly enriched in central nervous system development (**E-H**). These results suggest these pcRNAs are targets of Polycomb and silenced or transcriptionally poised in undifferentiated pluripotent cells (Bernstein et al., 2006) consistent with roles in differentiation and development. (**A**) Boxplot showing the mean coverage of pcRNA promoters based on ChIP-Seq signal for H3K27me3 and H3K4me3 in GM12878, H1-hESCs, HSMM and K562. **(B)**: Scatter plot reporting the signal intensities of H3K4me3 (x-axis) and H3K27me3 (y-axis) in the promoters of pcRNAs. The four subplots represent data from H1-hESCs, GM12878, HSMM and K562. The colour coding reports the hierarchical clustering results. A single pcRNA had 0 H3K4me3 signal in H1hESCs and fell alone in a fifth cluster (not shown). (**C**) Boxplot showing the expression (log10 FPKM) of pcRNAs based on RNA-Seq data on ES cells (left total cells; middle, cytoplasm; right, nucleus) and split by the cluster determined by applying hierarchical clustering to the H3K27me3 and H3K4me3 ChIP-Seq data (See Supplementary Methods). (**D**) Histograms showing the number of expressed pcRNAs based on RNA-Seq data on ES cells (left total cells; middle, cytoplasm; right, nucleus) and split by the cluster determined by applying hierarchical clustering to the H3K27me3 and H3K4me3 ChIP-Seq data (See Supplementary Methods). pcRNAs with FPKM higher than 0.1 were considered expressed. **(E-H):** GO enrichment analysis of coding genes associated to pcRNAs in each of the clusters determined by applying hierarchical clustering to the H3K27me3 and H3K4me3 ChIP-Seq data (See Supplementary Methods). The x-axis shows the enrichment score, calculated as the number of pcRNA-associated genes in a given GO category divided by the total number of genes in the category. The size of the points indicates the absolute number of pcRNA-associated genes in the given GO category. The color-coding indicates the adjusted p-value.*

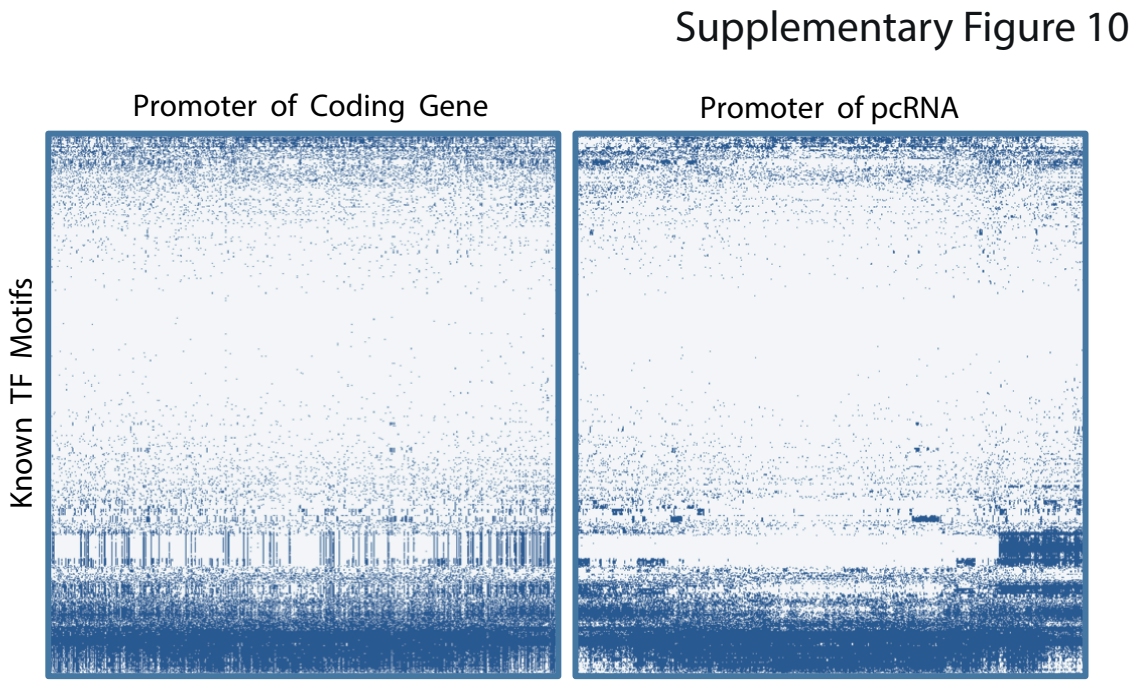


Figure 10: *Heatmap indicating what transcription factor binding motifs are found in pcRNA promoters (right) and promoters of pcRNA-associated coding genes (left). The known motifs are from JASPAR (freeze 2014-12-10, 263 motifs) (Kheradpour and Kellis, 2014) (2,065 motifs) and (Jolma et al., 2013) (843 motifs).*

Supplementary Figure 11

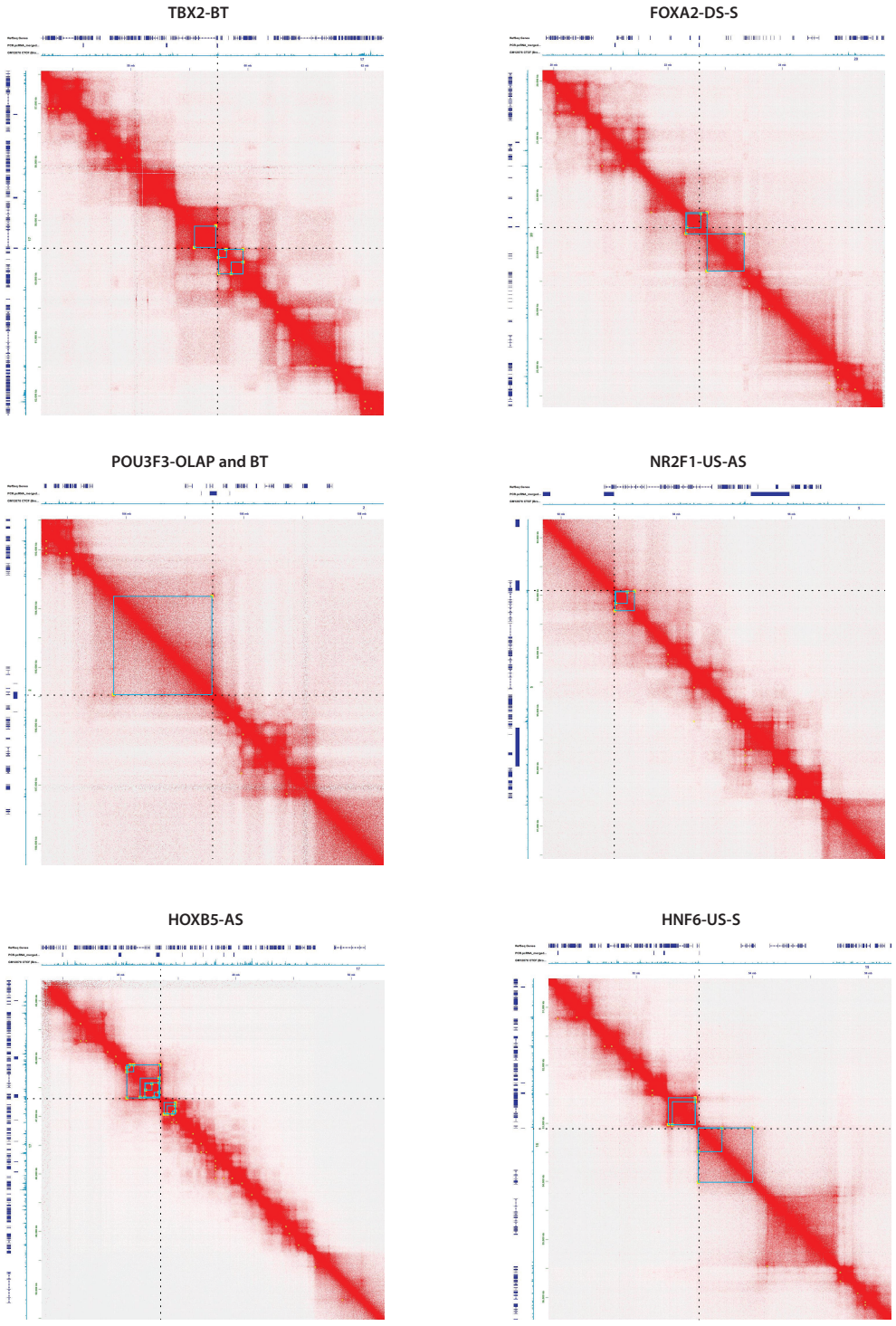


Figure 11: *Examples of HiC heat-maps, showing TADs (blue rectangles), loop anchor points (yellow dots), and tapRNA positions (black dotted lines). HiC data from GM12878 cells (Rao et al., 2014), with the first, second and third tracks corresponding to RefSeq coding genes, tapRNAs and CTCF ChiP-seq signal in GM12878, respectively.*

Supplementary Figure 12

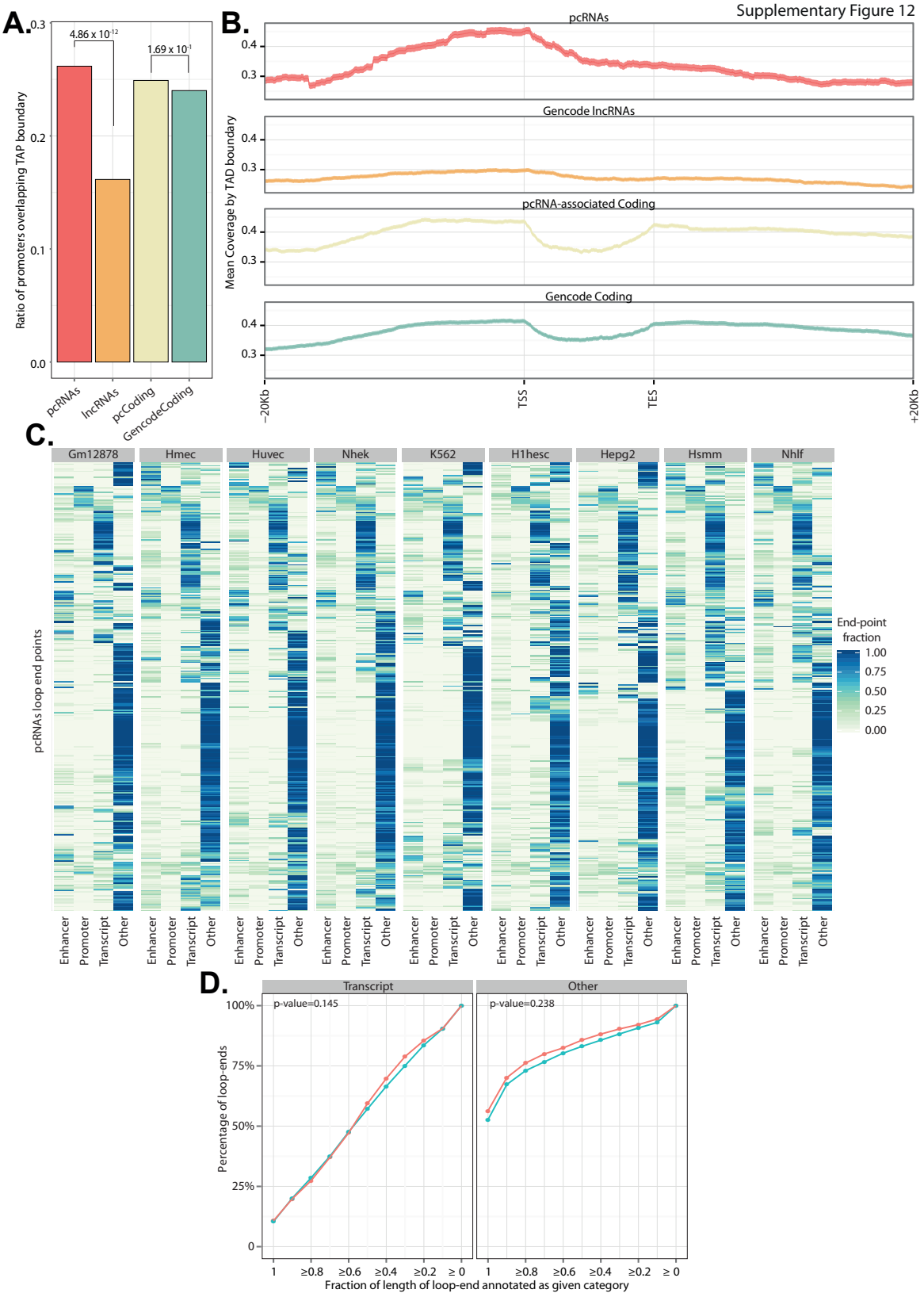


Figure 12: ***A**: Bar chart showing the proportion of pcRNAs, pcRNA-associated coding genes, Gencode lncRNAs and Gencode coding genes with a TAD boundary overlapping their promoter. The p-values reported were calculated with hypergeometric tests. **B**: TAD boundary coverage of loci of pcRNAs, pcRNA-associated coding genes, Gencode lncRNAs and Gencode coding genes. The plots report the loci from 20kb upstream of the transcription start site (TSS) to 20kb downstream of the transcription end site (TES). For visualization purposes these profiles show the coverage of a random sample of 5000 Gencode lncRNAs and 5000 random Gencode coding genes (**C**) Heatmap showing the proportion of each distal genomic region in contact with pcRNA promoter annotated is each genomic category derived from the ENCODE chromatin segmentation data (see Supplementary Methods). (**D**) Cumulative distribution plot showing the percentage of distal genomic regions in contact with pcRNA promoters (y-axis) as a function of the fraction of length of loop-end annotated as Transcript (left) or Other (right) according to the ENCODE chromatin segmentation data (see Supplementary Methods).*

Supplementary Figure 13

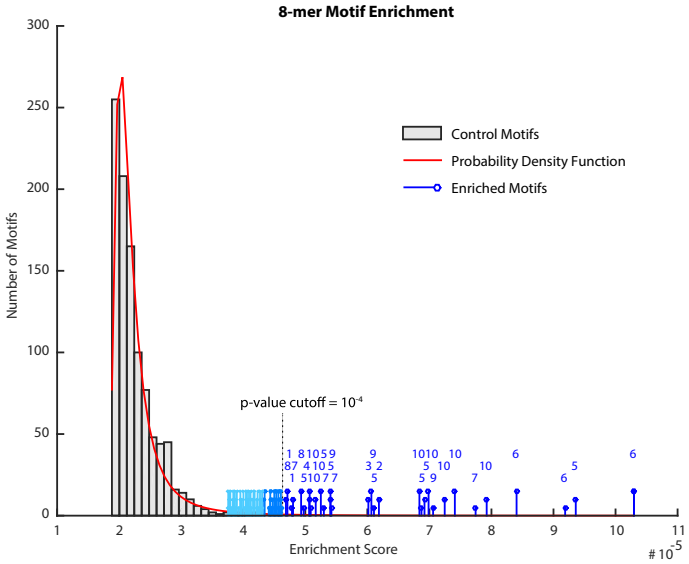


Figure 13: *Significantly enriched 8-mer motifs in conserved domains. Probability density function of Monte Carlo simulation results are shown in bar graph. The motifs have p-value <= 10^{-4} are considered as enriched motifs (shown in blue). The numbers on the enriched 8-mer motif stems are the consensus motif numbers as in* ***Figure 5D***.

Supplementary Figure 14

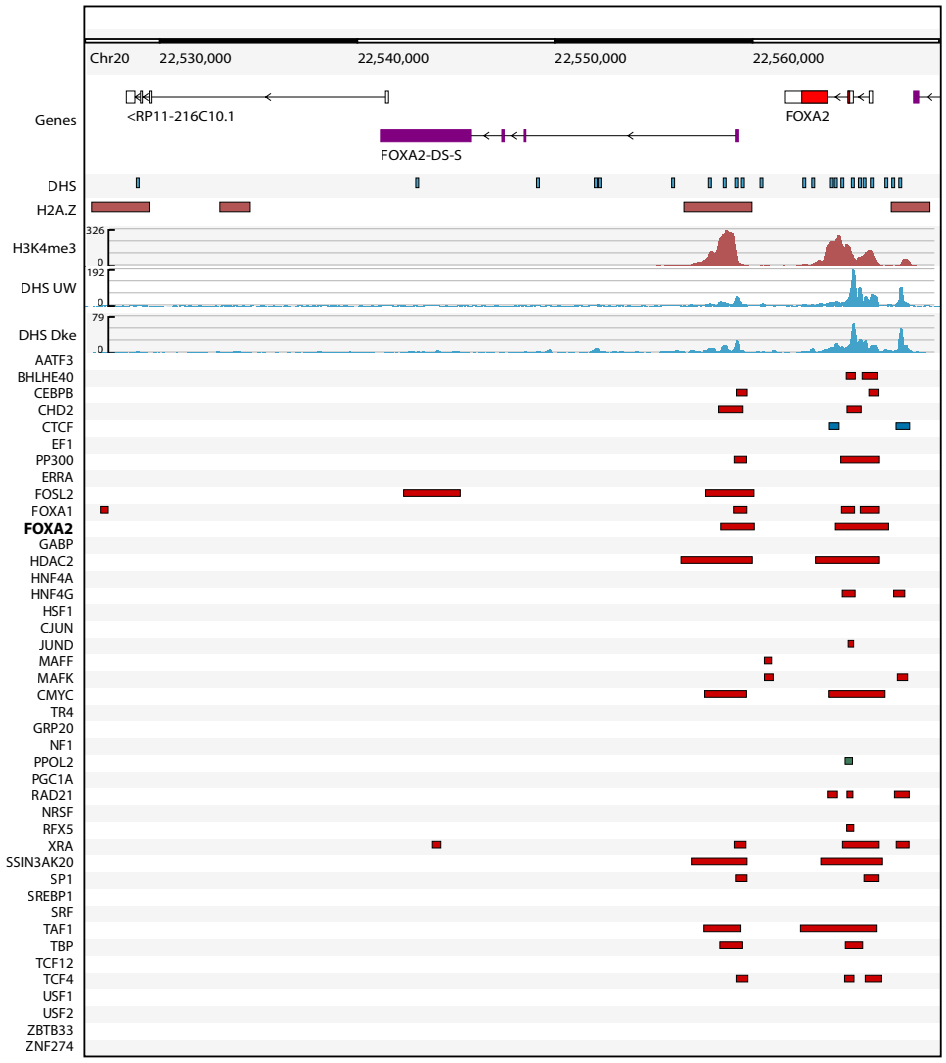


Figure 14: *Screenshot from the Dalliance genome browser (Down et al., 2011) showing the* FOXA2 *locus with tracks displaying coverage data for several ChIP-Seq experiments performed by the ENCODE project on HepG2 cells.*

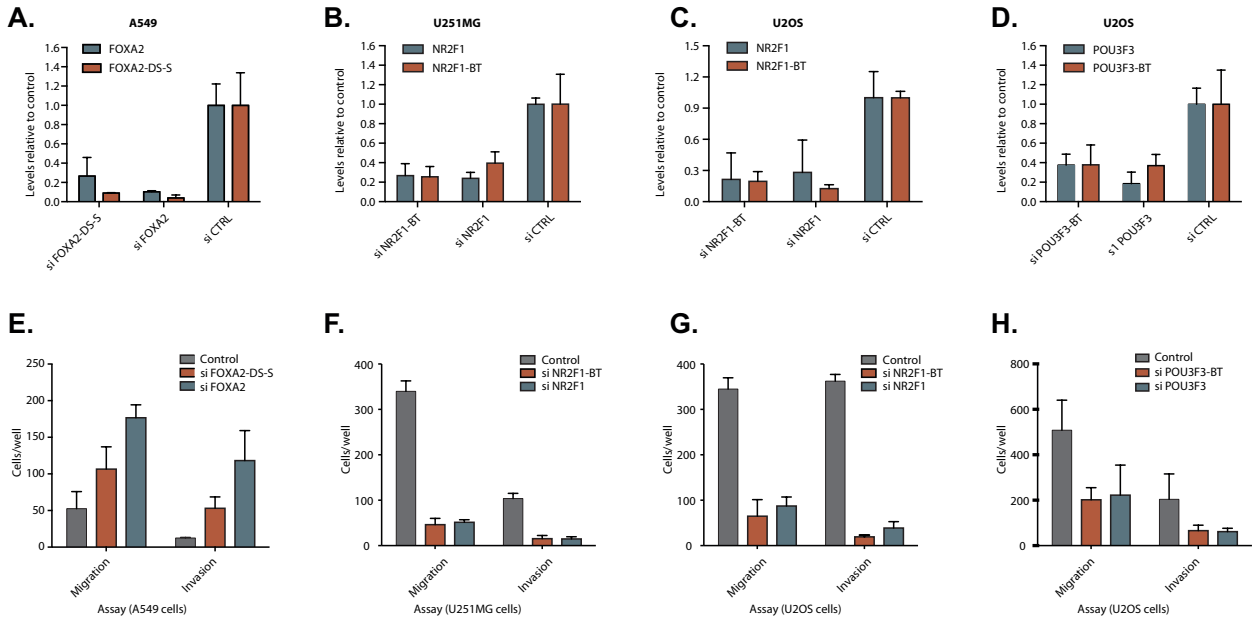


Figure 15: ***A-D****: Real time PCR data showing the expression of pcRNAs and associated coding genes upon knock-down of* FOXA2-DS-S *(**A**),* NR2F1 *(**B, C**),* POU3F3 *(**D**) and their associated pcRNAs. (**E-H**) Invasion and migration assay analysis upon knock-down of* FOXA2 *(**E**),* NR2F1 *(**F,G**),* POU3F3 *(**H**) and their pcRNAs compared to negative control siRNA.*

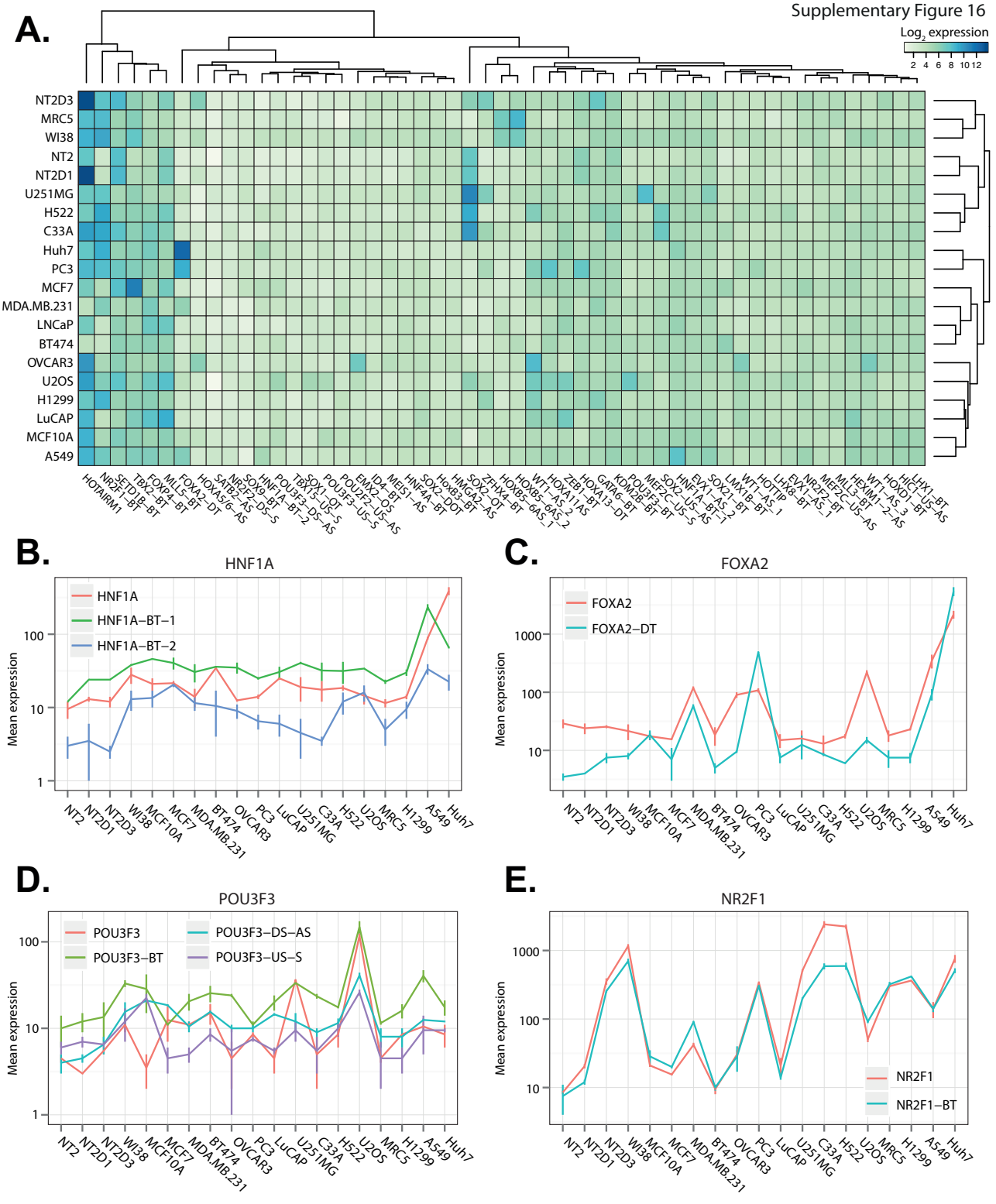


Figure 16: ***A****: Heatmap showing the Nanostring expression profiles of human pcRNAs across all the cancer cell lines included in the assay. (**B-E**) Nanostring expression profiles of human pcRNAs HNF1A (**B**), FOXA2 (**C**), POU3F3 (**D**) and NR2F1 (**E**) across all the cancer cell lines included in the assay.*

2 Supplementary Table Legends

Supplementary Table S1: List of RNA-Seq datasets used in the study.

Supplementary Table S2: Annotation of pcRNAs.

Supplementary Table S3: GO enrichment of pcRNA-associated protein coding genes.

Supplementary Table S4: GO enrichment protein coding genes associated with pcRNAs in each possible orientation.

Supplementary Table S5: Nanostring data. **A**: Annotation of probes used in the assay. **B:** List of samples tested in the assay. **C:** Normalised expression of tested human and mouse pcRNAs and associated coding genes.

Supplementary Table S6: Metanalysis of pcRNA expression across different cancer studies. **A:** Samples used in the meta-analysis. **B:** Expression of pcRNAs. **C:** Expression of pcRNA-associated coding genes. **D:** Summary table with the number of expressed pcRNAs and coding genes in each study.

Supplementary Table S7: Oligonucleotides, clones and cell lines used in this study.

Supplementary Table S8: ChIP-seq experiment data from the ENCODE Project.

3 Supplementary Methods

3.1 Human and Mouse reference genomes

The reference genomes for human (hg38) and mouse (mm10) where downloaded from the UCSC FTP server (Rosenbloom et al., 2015) in 2bit format and converted to fasta format using the twoBitToFa tool from the UCSC genome browser. The fasta files where indexed using samtools faidx (v1.2).

The Bowtie index for both genomes where built with bowtie2-build (v2.1.0) (Langmead and Salzberg, 2012).

3.2 Human and Mouse reference transcriptomes

The reference gencode transcriptomes for human and mouse (version 21 for human and version M4 for mouse) were obtained from the Gencode website in GTF format (Harrow et al., 2012).

3.3 Genbank all RNAs

The annotation of mouse Genbank mRNAs was obtained from the "Mouse mRNAs from GenBank" track of the UCSC genome browser using the Table Browser (Karolchik et al., 2004).

3.4 RNA-Sequencing data analysis

In order to obtain comprehensive transcriptomes for human and mouse as well as to quantify pcRNA abundance, we integrated the reference Gencode transcriptomes with RNA-Seq data on human and mouse tissues and cell lines. We used RNA-Seq from six matched human and mouse tissues (Brain, Cerebellum, Heart, Kidney, Liver and Testis) as well as data produced by the ENCODE project from similar human and mouse cell lines (**Supplementary Table S1**).

3.4.1 Mapping

The RNA-Seq datasets where mapped to the reference human and mouse genomes (hg38 and mm10 respectively) using Tophat2 (v2.0.10, bowtie2 v2.1.0 (Kim et al., 2013; Langmead and Salzberg, 2012)) with the options `--b2-sensitive --zpacker pigz` and the Gencode comprehensive GTF files as reference transcriptomes (v21 and M4 for human and mouse respectively). The reference transcriptomes were built with an independent Tophat run without fastq files and then provided to all subsequent mapping runs through the option `--transcriptome-index`

The `--library-type` option was set to "fr-unstranded" for unstranded datasets and "fr-firststrand" for stranded datasets.

For two very deep (>120mln reads each), single end 45nt reads mouse datasets (SRR549335 and SRR549339, see Supplementary Table S1) Tophat by default tried to identify splice junction by coverage search, but stopped at the stage "Searching for junctions via segment mapping" probably due to the very high number of reads. To overcome this problem, we disabled only for these two samples the coverage-search functionality (option `--no-coverage-search`) as suggested by Tophat's standard error.

3.4.2 Assembly

The transcriptomes where assembled independently for each RNA-Seq dataset using Cufflinks (v2.2.1) with the following options:

`--library-type` "fr-unstranded" for unstranded datasets and "fr-firststrand" for stranded datasets.

`-F 0.05`

`--multi-read-correct`

`--frag-bias-correct` , pointing to the genome fasta file

`-M` a masking GTF files to exclude ribosomal transcripts and mitochondrial transcripts. This file was produced by selecting from the Gencode GTF files the lines that matched "Mt_" or "rRNA" in field 14.

`-g` exon-cds-filtered reference transcriptome GTF. This file was produced by selecting only exon and CDS features from the Gencode reference GTF files (field 3), therefore excluding the "gene" and "transcript" entries. Such a filtered GTF file contains all the information needed by Cufflinks and provides a significant speed up in cufflinks' running time.

The Cufflinks assembled transcriptomes for each sample were then merged using Cuffmerge (with the same exon-cds reference transcriptome used for Cufflinks) and converted to BED12+ format using the gtfToBed tool (Kent et al., 2002) with the option "`-a gene_id,oId,class_code`" to preserve Gene ID, Gencode ID and Cufflinks class codes as additional fields.

3.4.3 Abundance estimation and expression normalisation

The human and mouse merged transcriptomes (merged.gtf) where then quantified against each BAM file using Cuffquant (v2.2.1) with the following options:

`--library-type` (see 3.4.2)

`--multi-read-correct`

`--frag-bias-correct`

`-M` Reference masked regions (see 3.4.2)

Finally, the Cuffquant binary output files were normalised with Cuffnorm (v.2.2.1) to produce the human and mouse expression matrices. Cuffnorm was run with the following options:

`--output-format cuffdiff`

`--use-sample-sheet`

`--library-type fr-unstranded`

3.5 Identification of pcRNAs

3.5.1 Human Data preparation

The purpose of this data preparation step is to produce an annotation of reference and novel non-coding transcripts from which we will later identify pcRNAs.

1) Annotation of coding transcripts and CDS

From the Gencode BED annotation we selected transcripts containing an annotated CDS. We then used the getCoding tool of pinstripe to obtain a BED annotation of only the coding portion (CDS) of each coding transcript.

2) Reference non coding RNAs

We filtered the Gencode V21 BED file in the following way:

1. We used awk to select all transcripts without an annotated ORF and composed of more than one exon.

2. We used overlapSelect (UCSC genome browser tool (Kent et al., 2002)) to exclude all transcripts that had more than 20bp of sense overlap with the CDS region of a coding transcript.

3) Novel non coding RNAs

We filtered the merged RNA-Seq transcriptome BED file in the following way:

1. We used awk to remove single exon transcripts as well as transcripts that don't map to the primary assemblies of the autosomes or sex chromosomes.

2. We used overlapSelect to discard transcripts with more than 20bp of sense overlap with the CDS region of a coding transcript.

3. We used bedtools intersect (v2.24.0) to discard transcripts with more than 50% sense exonic overlap with reference non-coding transcripts (previous step).

4. We used Pinstripe dedup (version v1.0.4554.32000, with option --exEncomp) to remove redundant transcripts.

5. We used CPAT (v1.2, (Wang et al., 2013)) to calculate the coding potential of each transcript and only retained transcripts with score $<$0.364 (see CPAT documentation for information on the threshold).

Finally, we combined *the Reference non coding RNA* annotation and the *Novel non coding RNAs* annotation and we used bedtools intersect to remove all transcripts with more than 50% sense exonic overlap with coding transcripts (although in the previous step we had already filtered out CDS-overlapping transcripts, this step ensures that we do not have transcripts that have more than 50% overlap with the UTR of coding genes).

The file that we obtain is a comprehensive annotation of all reference and novel human non-coding RNAs and we will hereafter refer to it as *know+novel ncRNAs*.

3.5.2 Mouse Data preparation

The purpose of this data preparation step is to produce an annotation of reference and novel non-coding transcripts from which we will later identify pcRNAs.

1) Annotation of coding transcripts and CDS

From the Gencode BED annotation we selected transcripts containing an annotated CDS. We then used the getCoding tool of pinstripe to obtain a BED annotation of only the coding portion (CDS) of each coding transcript.

2) Reference non coding RNAs

We filtered the Gencode M4 BED file in the following way:

1. We used awk to select all transcripts without an annotated ORF and composed of more than one exon.

2. We used overlapSelect to exclude all transcripts that had more than 20bp of sense overlap with the CDS region of a coding transcript.

3) Novel non coding RNAs

We filtered the merged RNA-Seq transcriptome BED file in the following way:

1. We used awk to remove single exon transcripts as well as transcripts that don't map to the primary assemblies of the autosomes or sex chromosomes.

2. We used overlapSelect to discard transcripts with more than 20bp of sense overlap with the CDS region of a coding transcript.

3. We used bedtools intersect (to discard transcripts with more than 50% sense exonic overlap with reference transcripts (previous step).

4. We used Pinstripe dedup (with option --exEncomp) to remove redundant transcripts.

5. We used CPAT to calculate the coding potential of each transcript and only retained transcripts with score $<$0.44 (see CPAT documentation for information on the threshold).

6. We removed all transcripts with more than 50% sense exonic overlap with a coding transcript to remove UTR overlapping RNAs.

4) *Genbank non coding RNAs*

Given the lower number of lncRNAs annotated by Gencode in mouse (6951 in Gencode M4 vs 15877 in human Gencode V21), we also incorporated in our analysis Genbank non coding RNAs.

To identify them, we downloaded the "all mRNAs" GTF from the UCSC genome browser and processed it in the following way:

1. We used the gffread tool (v2.2.1, part of the Cufflinks suite) to exclude all transcripts with non-canonical splice sites (i.e. not GT-AG, GC-AG or AT-AC) and with introns shorter than 4nt, then we converted the filtered GTF file to BED with Pinstripe gtfToBed.
2. We retained only transcripts with more than one exon.
3. We discarded transcripts with more than 20bp of sense overlap with the CDS region of a coding transcript (overlapSelect).
4. We used CPAT to calculate the coding potential of each transcript and only retained those with score $<$0.44 (see CPAT documentation for information on the threshold).
5. We removed all transcripts with more than 50% sense exonic overlap with a coding transcript to remove UTR overlapping RNAs.

3.5.3 Identification of conserved promoters

For each transcript in the human *know+novel ncRNAs* annotation (see 3.5.1) we produced a BED file of their promoter regions by extending their TSSs of 500bp in each direction, then we obtained their FASTA sequence from the reference genome using the Pinstripe getDna tool.

In order to make a blast database of the mouse genome we first used the ncbi-blast convert2blastmask tool (v2.2.30+, options `-masking_algorithm repeat -masking_options "repeatmasker and tandem repeats from UCSC" -outfmt maskinfo_asn1_bin`) to extract masking information from the soft masked genome fasta file, and then the makeblastdb tool (v2.2.30+, `options -mask_data path-to- convert2blastmask --out -dbtype nucl`).

We then used the following command line to align human ncRNA promoters to the mouse genome with blast (v2.2.30+):

```
blastn -task blastn --db path/to/db -out path/to/out -query path/to/promoters/fasta -outfmt
6 -evalue 0.001 -num_threads n-processors -db_soft_mask 40 -lcase_masking
```

Finally, we processed the blastn output file with awk to only retain alignments longer 100nt and with E-value $<10^{-10}$ and convert the blast coordinates (1 based) into BED coordinates (0 based).

3.5.4 Non-coding to coding positional annotation

We next aimed to associate each ncRNA identified in human and mouse to its closest protein coding transcripts. To this end we used Pinstripe "closest", which returned – for each input ncRNA – the closest upstream, downstream and overlapping coding transcript. We then processed each entry and compared the non-coding and coding coordinates to annotate their TSS-to-TSS distance as well as the orientation of the non-coding relative to the coding in the following way:

- If the coding and non-coding intervals overlapped we defined the coding-non-coding pair as OLAP if on the same strand or AS if on different strands.
- If there was no overlap and the non-coding was upstream of the coding (relative to the strand of the coding), we defined the pair as US-S if coding and non-coding were on the same strand, otherwise US-AS if the TSS-to-TSS distance was $>$2000bp or BT if $<=$2000.
- If there was no overlap and the non-coding was downstream of the coding (relative to the strand of the coding), we defined the pair as DS-S S if coding and non-coding were on the same strand, otherwise DS-AS.

We then matched each human and mouse coding transcript to their corresponding Ensembl Gene Ids, and for each non-coding/coding gene pair in a given orientation we only retained the closest coding transcript.

3.5.5 Human-mouse positional comparison

To identify mouse ncRNAs arising from conserved human ncRNA promoters we extended each region in the mouse genome that resulted from blasting human ncRNA promoter (see 3.5.3) of 500nt in each direction, and then we intersected these regions with the 5' exon of each mouse ncRNA.

This step allowed us to obtain pairs of human/mouse ncRNAs that have a conserved promoter. We then selected those pairs for which at least one coding neighbour of the human non-coding (where neighbouring means the closest upstream, downstream and overlapping as defined in the previous step) was orthologous to at least one coding neighbour of the mouse non-coding

To identify orthologous genes between mouse and human we programmatically downloaded from Ensembl Biomart (v80) a table that associates each human gene_id to the gene_id of the orthologous gene in mouse.

The resulting annotation contains human and mouse ncRNAs whose promoter is conserved and whose neighbouring gene(s) is(are) orthologous.

We further filtered this annotation by removing all human/mouse ncRNA pairs that were in opposite orientations relative to the coding genes in the two species (i.e. DS-S, US-S or OLAP in one species and AS, BT, DS-AS, US-AS in the other).

In numerous cases we could not univocally associate each ncRNA to a single coding gene, since the same ncRNA can have multiple neighbouring coding genes orthologous and in the same orientation in mouse and human. To resolve these ambiguities and univocally assign each ncRNA to a unique coding genes we applied the following criteria:

1) In case any of the possible coding genes were either AS or OLAP in human we retained the closest (TSS-to-TSS) of those.

2) In all other cases we retained the coding with shortest TSS-to-TSS distance in human.

3.5.6 Annotation of pcRNA genomic characteristics

To annotate pcRNAs that overlapped Gencode lncRNAs we intersected the human pcRNA annotation with the Gencode annotation of lncRNAs considering all exonic sense overlaps.

To annotate pcRNAs that overlapped miRNAs we queried the UCSC genome browser MySQL server for all transcripts containing the string "miR" in the geneName field.

To annotate pcRNA promoters we extended each pcRNA TSS by 2000bp in each direction and merged the resulting promoter regions that overlapped (bedtools mergeBed).

3.6 Characterisation of pcRNA features and expression analysis

To produce human and mouse expression matrices we matched the Ensemble Transcript IDs of human and mouse pcRNAs with the "oID" identifiers reported by Cuffmerge; we then used the corresponding Cuffmerge IDs to track pcRNAs in the isoforms FPKM tracking files reported by Cuffnorm.

For human and mouse coding genes we used a similar approach to extract the FPKM transcript information for all transcripts of each coding gene, and then summed the FPKMs to obtain a single expression measure at the gene level.

For the expression analysis all FPKM values below 10^{-3} were set 0 and all transcripts with 0 FPKMs in all samples were excluded.

3.6.1 pcRNA expression heatmaps

The expression heatmaps for human and mouse pcRNAs were produced with the function heatmap.2 of the gplots package. The rows and columns were clustered with the default methods. For visualisation purposes

in order to calculate the log2 of the FPKMs the smallest FPKM value was added to each value. The vertical sidebar reports the tissue specificity score calculated as indicate below.

3.6.2 pcRNA expression distance heatmaps

The heatmaps showing the Euclidean distance between pcRNA expression profiles have been realized by calculating the matrix of pairwise Euclidean distances between all pcRNAs using the *dist()* function in R. The heatmap was produced with the heatmap.2 function the gplots package using the default methods for clustering rows and columns. The horizontal sidebar reports the tissue specificity score calculated as indicate below. The vertical sidebar reports the tissue where a given pcRNA has maximal expression.

3.6.3 GO enrichment of pcRNA-associated coding genes

The GO enrichment of pcRNA-associated genes was obtained using the TopGO package of Bioconductor. The ontology mapping used was provided by the package org.Hs.eg.db. The background set of coding genes consisted of all human protein coding genes with an annotated mouse ortholog. GO nodes with less than 10 annotated terms were excluded from the analysis. The p-values were calculated using the "default" method of TopGO and Fisher's Exact test. P-values were corrected using the Benjamini-Hochberg method as implemented in the *p.adjust(method="BH")* function in R. For the GO enrichments of pcRNA-associated genes divided by relative orientation, we used as background the set of all pcRNA-associated coding genes. P-values were calculated as described above but were not corrected for multiple hypothesis testing.

3.6.4 Correlation of expression between pcRNAs and coding genes and between human and mouse pcRNAs

To calculate the Spearman's rank correlation coefficients between human pcRNAs and coding genes we first calculated a matrix of coefficients between each pcRNA and each coding gene, where the diagonal represented the coefficients between each pcRNA and its associated coding gene.

To test whether the mean correlation coefficient was higher than expected by chance we performed a permutation test: we selected 10^6 samples of random coefficients from the entire matrix, and calculated how many times the mean of the random sample was higher or equal to the mean of the diagonal of the matrix. We reported a p-value$<10^6$ when none of the random samples' means was higher or equal to the mean of the diagonal.

The correlation coefficients were calculated in R with the function *cor(method="spearman")*.

The correlation of expression between human and mouse pcRNAs was calculated in the same way. When the same human pcRNA was associated to multiple mouse pcRNAs we calculated the correlation between all pairs.

3.6.5 Tissue specificity score and GO enrichment by tissue

The tissue specificity score for human and mouse pcRNAs and coding genes was based on the square root of the Jensen-Shannon divergence as in (Cabili et al., 2011). The p-value for the difference between pcRNAs and coding genes was calculated with the Wilcoxon test as implemented in the wilcox.test function in R. To verify whether the difference of tissue specificity score between pcRNAs and coding genes was due to their different expression levels, we used the MatchIt R package (*method="subclass", subclass=5, sub.by="control"*) to subdivide pcRNAs and coding genes in 5 classes so that each class had similar distributions of maximal FPKM. We then calculated the tissue specificity score distribution for each of the 5 classes.

The GO enrichment of pcRNA by tissue of maximal expression was done by selecting the coding genes associated with pcRNAs with maximal expression in the given tissue and with a tissue specificity score above the mean of all specificity scores. The GO enrichment was performed in R using the TopGO package. The background set of coding genes consisted of all pcRNA-associated coding genes. GO nodes with less than 20 annotated terms were excluded from the analysis. The p-values were calculated using the "default" method of TopGO and Fisher's Exact test. P-values were not corrected for multiple hypothesis testing.

3.6.6 Human-mouse conservation analysis

To calculate the sequence conservation of human and mouse pcRNAs we aligned the human and mouse pcRNA sequences using the Needleman–Wunsch algorithm (Needleman and Wunsch, 1970) . In case multiple mouse pcRNA isoforms were associated with the same human pcRNAs we performed all possible pairwise alignments and only retained those with the highest sequence identity. Similarly, to calculate the sequence conservation of pcRNA-associated protein coding genes we performed pairwise alignments (Needleman–Wunsch algorithm) between all transcripts of the human gene and all transcripts of the mouse gene, and retained the alignment with the highest sequence identity.

3.7 Nanostring analysis

For the nanostring experiment we designed probes to detect 50 pairs of pcRNAs and corresponding coding genes in human and mouse. The probes were designed according to the Nanostring guidelines and to maximize their specificity (Supplementary Table S5) and included 9 house-keeping genes for normalization (*ALAS1, B2M, CLTC, GAPDH, GUSB, HPRT, PGK1, TDB, TUBB*).

The raw count data were first normalized by Nanostring Technologies with the nSolver software using a two-step protocol. First, data were normalized to internal positive controls, then to the geometric mean of house-keeping genes. The normalised data was then imported into R for further analysis. The correlation of expression between pcRNAs and coding genes was calculated with the *cor()* function in R after averaging replicate samples. To test whether the mean correlation coefficient between pcRNAs and coding gene as well as between human and mouse pcRNA pairs was higher than expected by chance, we used a permutation test as described for the RNA-Seq analysis.

To cluster pcRNAs and coding genes based on their expression profiles with first used the mcxarray tool of MCL (van Dongen, 2008) to produce a graph where nodes represented human pcRNAs and corresponding coding genes, and edges connected nodes with a Pearson correlation coefficients higher than 0.5. We then run MCL on such graph with the inflation parameter set to 3 to identify clusters of pcRNAs and coding genes.

3.8 FOXA2-DS-S knock-down microarray analysis

RNA samples were amplified using the TotalPrep 96-RNA amplification kit from Ambion (Applied Biosystems). Briefly, the RNA was converted into cDNA, and amplified by In Vitro Transcription (IVT) to generate biotin-labeled cRNA. The cRNA was then hybridized to the *HumanHT-12 Expression Chips, version 4* following the Direct Hybridization assay.

The data obtained was imported into R and analyzed with the beadarray Bioconductor package (Dunning et al., 2007) and the illuminaHumanv4.db annotation package. We summarized the data for each array using the *summarize()* function of beadarray with default parameters (log2 transformation, removal of outliers with a 3 median absolute deviation cutoff) and removed all probes without a quality score or with a "Bad" quality score in the annotation package. We then normalized the data with the *normaliseIllumina()* function with the quantile method and retrieved Ensembl IDs for each array probe using biomaRt. We then performed the differential expression analysis using limma with the model formula ~0+Condition, where Condition identifies the Control samples, FOXA2-KD samples and FOXA2-DS-S samples. We also supplied to the *lmFit()* function a weight for each array estimated using the *arrayWeights()* function. The GO enrichment analysis was performed separately on significantly up-regulated (adjusted p-value$<$0.05 and log2 fold change$>$0) and down-regulated (adjusted p-value$<$0.05 and log2 fold change$<$0) genes using the TopGO package. As background set we used all probes in the array with an Ensembl gene id. The GO enrichment was performed with the classic algorithm and p-values calculated with the fishes exact test. P-values were corrected for multiple hypothesis testing using the Benjamini-Hochberg method as implemented in the *p.adjust()* function. All GO terms with less than 20 annotated genes were excluded from the analysis.

3.9 Microarray meta-analysis

Probe set sequences of GPL570 platform (Affymetrix Human Genome U133 Plus 2.0 Array) were retrieved from Affymetrix website (`http://www.affymetrix.com/`), and aligned against hg38 human genome assembly using Blat (Kent, 2002). We next removed probe sets with less than 90% identity and coverage, and cross-referenced the remaining ones against the pcRNAs genomic coordinates (BED12 format) and the protein coding genes (Gencode version 23) using BEDtools (Quinlan, 2014). Probe sets were annotated as pcRNA or as coding gene if at least 70% of its sequence mapped to the reference transcript sequence.

We then download from GEO database (`http://www.ncbi.nlm.nih.gov/geo`) 63 microarray studies of GPL570 platform, which contained tumor and non-tumour tissue samples (Figure 6A, Supplementary Table S6). For each study, raw data (CEL files) were processed and normalized using RMA algorithm, and samples were manually classified as "normal" or "tumor" according to the description provided by authors. We used student t-test (fold-change > 1.25 and p-value < 0.005) to identify transcripts differentially expressed in either tumor or normal samples. Spearman correlation was used to compare the expression between the pcRNA and its associated coding gene. Plots were generated using the following R packages: gplots and ggplot2.

3.10 pcRNA histone modification profiles

To produce histone modification maps of pcRNA promoters we downloaded the normalised bigWig files from the EBI ENCODE respository for 14 ChIP-Seq experiments (Control, Ctcf, H2az, H3k27ac, H3k27me3, H3k36me3, H3k4me1, H3k4me2, H3k4me3, H3k79me2, H3k9ac, H3k9me1, H3k9me3, H4k20me1) in GM12878, H1-hESC, HSMM and K562 cells. We then converted the data to bedGraph format (with bigWigToBedGraph) and used liftOver to convert the coordinates from hg19 to hg38. Then, we used bedtools mergeBed to merge the overlapping intervals and converted the resulting files back to bigWig format (bedGraphToBigWig).

For each cell line we then used the computeMatrix of the Deeptools package (reference point TSS) to calculate the coverage in each ChIP-Seq experiment of each pcRNA as well as of 100 random Gencode lncRNAs and 100 random Gencode coding Genes (random coding genes and lncRNAs selected with Bedtools sample with seed 383847). The resulting matrix was then loaded in R to produced the profile plots.

3.11 Analysis of H3K27me3 in ESCs

To study the H3K27me3 profiles of pcRNA promoters, the bedGraph files (in hg38 coordinates, see above for details of conversion from hg19) of H3K27me3 and H3K4me3 in GM12878, H1-hESC, HSMM and K562 have been mapped to the promoters of pcRNAs (defined as TSS +/- 1Kb) using the mapBed tool of bedtools to calculate the mean coverage of each promoter in each cell line. The data was then loaded into R and the data for H1hesc was subjected to hierarchical clustering using the hclust function (default parameters) on the Euclidean distances matrix (dist function) between the base 10 logarithms of the mean promoter coverage for H3K27me3 and H3K4me3 (the log10 was calculated after adding 0.01 to each value). The GO enrichment for the coding genes associated to the pcRNAs in each cluster was performed using TopGO (classic algorithm) using the Fisher Exact Test to compute p-values. P-values correction and background set were the same as described for "*GO enrichment of pcRNA-associated coding genes*".

3.12 ENCODE ChIP-seq data analysis

We downloaded 2,216 ChIP-seq experiment data from the ENCODE Project. The list of the data is in Supplementary Table S8. The data were lifted over from hg19 to hg38. We found overlapping peaks on four different categories: (1) 500bp upstream the promoter region of pcRNA-associated coding genes, (2) 500bp upstream the promoter region of pcRNAs, (3) pcRNA genomic loci, and (4) pcRNA genomic loci but not overlapping with promoter region. To understand the correlation of TF binding patterns in the four categories, we made a binary matrix per category that consists of rows of TFs and columns of pcRNA/coding genes. Hence, the matrix contains connections between TF and pcRNA/associate coding genes. The matrix of category 2 is clustered by Euclidian Distance. To directly compare the TF binding patterns between each category, the other three

matrices were sorted by the same order of the clustered matrix. We used the MatLab function *corr2* to calculate r-value between category (1) and (2). We performed Monte Carlo simulation to calculate the p-value and test the significance of the r-value.

3.13 Known TF-binding motif data analysis

We downloaded known TF-binding motifs from JASPAR (freeze 2014-12-10, 263 motifs) (Kheradpour and Kellis, 2014) (2,065 motifs) and (Jolma et al., 2013) (843 motifs). We applied same analytical procedures on these datasets as described in previouse section (ENCODE ChIP-seq data analysis).

3.14 Identification of CTCF binding sites in pcRNA promoters

To identify CTCF binding sites within pcRNA promoters we downloaded the TFBS clusters (V3) from the ENCODE portal at the UCSC Genome Browser (wgEncodeRegTfbsClusteredWithCellsV3.bed.gz) and filtered the file for CTCF sites. We then converted the CTCF binding sites to hg38 coordinates using the liftOver tool and calculated how many pcRNA promoters overlapped HiC loops by (1) extending the TSS of each pcRNA by 2kb in both directions, (2) merging overlapping promoters (bedtools merge) and (3) intersecting the promoters with the CTCF sites. We repeated the same procedure for pcRNA-associated genes, Gencode coding genes and Gencode lncRNAs. To test whether pcRNA promoter were significantly enriched in CTCF binding sites compared to Gencode lncRNAs we performed a hypergeometric test as implemented in the phyper function in R.

3.15 Identification of HiC loops that overlap pcRNAs

We obtained the annotation of HiC loops by downloading the loops list files for HMEC, HUVEC, NHEK, K562, HeLa, KBM7, IMR90 and GM12878 cells deposited on GEO (GSE63525) and converted the intervals into Hg38 coordinates using liftOver. To calculate how many pcRNA promoters overlapped HiC loops we first extended the TSS of each pcRNA by 2kb in both directions, merged overlapping promoters (bedtools merge) and intersected the promoters with the loop coordinates. We also repeated the same procedure for all pcRNA-associated coding genes, Gencode coding genes and Gencode lncRNAs. To test whether pcRNA promoters are significantly enriched in HiC peaks compared to Gencode lncRNAs we performed a hypergeometric test as implemented in the *phyper()* function in R.

We applied the same strategy to identify TAD boundaries overlapping pcRNAs promoters. However, because TAD boundaries are single nucleotides rather then intervals, we extended each boundary of 10kb in each direction.

To analyze the end-points of the loops that overlap pcRNA promoters we downloaded the ENCODE Broad HMM data (Ernst et al., 2011) from the UCSC repository for GM12878, H1-hESC, HEPG2, HMEC, HSMM, HUVEC, K562, NHEK and NHLF cells. After converting the coordinated to Hg38 with liftOver we intersected each HMM dataset with the coordinates of the end points of the loops that overlapped pcRNAs or – as a control – all Gencode lncRNAs.

The data were then loaded into R for further analysis. First, we simplified the data by reducing the number of HMM categories in the following way: Strong and Weak Enhancer categories were grouped as Enhancer; Active, Weak and Poised promoter were grouped as Promoter; Txn_Elongation, Txn_Transition and Weak_Txn were grouped as Transcript; everything else was grouped as Other.

Then, for each end-point in each cell line we calculated the fraction covered by each HMM category and plotted these data as a heatmap using the heatmap.2 function of the gplots package. To determine whether pcRNA-loop end-points were enriched in any specific HMM category we calculated for each HMM category x the fraction of end-points annotated as x for at least $y\%$ of their length, where y ranged from 1 to 0 in steps of 0.1. Finally, we compared this distribution to the distribution obtained for all Gencode lncRNAs using the Kolmogorov-Smirnov test (as implemented in the *ks.test()* function in R).

3.16 TAD/Loop Boundary Enrichment Analysis

To check whether pcRNAs are localized at the boundary of TAD/Loop, we generated a density plot that shows cumulated count of pcRNA appearance across TAD/Loop regions (for each TAD including 10% proximity regions outside the TADs). Firstly, we extended the TSS of each pcRNA by 2kb in both directions, merged overlapping promoters (bedtools merge). Secondly, we extended TAD regions by 10% in both directions. We then intersected the promoters with the Loop coordinates and the extended TAD coordinates (bedtools intersect).

To visualize the cumulated count as a density plot, we only cumulated 10-bp window centered by TSS of each overlapping pcRNA to show precise localization of pcRNA TSS. We used all lncRNAs in the Gencode database (excluding pcRNAs) as a control. We then performed Kolmogorov-Smirnov test (as implemented in the kstest2 function in MatLab) to check how the enrichment is significant.

3.17 PhastCons Conservation Analysis

To understand the general conservation level of pcRNAs, we used phastCons scores for multiple alignments of 99 vertebrate genomes to the human genome. We downloaded wigFix format files from UCSC database to analyze at single base-pair resolution (`http://hgdownload.cse.ucsc.edu/goldenPath/hg38/phastCons100way/hg38.100way.phastCons/`). We fetched phastCons scores spanning the regions of pcRNA exon. The fetched scores were summed up per pcRNA and then divided by the total length of exons per pcRNA. We also repeated the same procedure for Gencode coding genes and lncRNAs. We then used Kernel smoothing function estimate (as implemented in the ksdensity function in MatLab) to plot the density of the normalized phastCons scores per pcRNAs. To test whether pcRNA are significantly conserved than Gencode lncRNAs, we performed Kolmogorov-Smirnov test as implemented in the kstest2 function in MatLab.

3.18 Conserved domain search

To identify conserved domains, we aligned transcribed sequences of human pcRNAs against their counterpart mouse pcRNAs. We took two alignment approaches: sliding-window and exon-by-exon. For the first approach, we made a 200nt-long window on each human pcRNA sequence and shifted the window by 40nt to align against the whole length of the transcribed mouse pcRNA sequence. For the second approach, we took each exon of human pcRNAs and aligned them against the whole length of the transcribed mouse pcRNA sequences. In both approaches, we used MATLAB function localalign, which returns local optimal and suboptimal alignments between two sequence. We found highly concordant search results in the results of both approaches and further analyzed the alignments by applying the following steps: (1) retain alignments only if the alignment score is greater than 100 or the ratio of identical matches is greater than 80%, (2) remove duplicate alignments among isoforms of pcRNAs based on merged isoforms of pcRNAs list (3) remove alignments if the aligned regions in human and mouse pcRNAs are not in same order of exons on their transcribed sequences, and (4) retain the best alignment if there are multiple alignments for the same region. Regarding the merged isoforms, we extracted only exonic regions of each pcRNA, and then merged the regions by using the bedtools merge function. The merged isoform allowed us to search once per a given genomic region that prevented multiple counting of same conserved domain and motif.

Through this process we generated a heatmap of conserved domains in human pcRNAs and clustered it by using the MATLAB function kmeans, which performs k-means clustering to partition the observations of a given matrix into k clusters. We used kmeans on the squared Euclidean distance measure and the k-means++ algorithm for cluster center initialization. We found 16 clusters and merged them into four larger clusters (**Figure 4A**). To annotate the functional category for each of the four clusters, we used the DAVID functional annotation tool with default settings (Huang da et al., 2009).

3.18.1 Motif search in conserved domains

To determine which regulatory motifs are over-represented in conserved domains with respect to background non-conserved regions, we identified all possible ungapped 8-mers in conserved domains and computed their frequency. An 8-mer is considered over-represented if its frequency in conserved domain is significantly higher

than the frequency in background non-conserved region. In the list of over-represented motifs, we found the presence of repeats that are consisted of a single nucleotide or dimer repeated for the entire 8-mer. This phenomenon is common in genomic sequences and generally is associated with non-functional components, and thus, these were filtered out.

To assess the statistical significance of the computed frequency for the over-represented motifs, we generated random sequences according to the nucleotide composition of the original sets of sequences. The frequencies for the random 8-mers were computed, and the distribution of the frequencies was approximated by the extreme value distribution. We used the MATLAB function gevfit to compute the maximum likelihood estimation of the extreme value distribution. We then overlaid a scaled version of its probability density function, computed using MATLAB function gevpdf, with the histogram of the frequency of the random 8-mer sequences. We repeated this process 100 times for bootstrapping and calculated the p-value. We concluded that the over-representation of the 8-mer motifs in conserved domain is statistically significant if the p-value estimate is less than $1x10^{-4}$ (**Supplementary Figure S13**).

3.18.2 Consensus motifs and De novo motif discovery

To identify consensus motifs, 32 enriched 8-mers were phylogenetically clustered into 10 groups. We used the MATLAB function seqlinkage to construct phylogenetic tree from pairwise distances. We then used the MATLAB function seqlogo to identify consensus motif and its weight matrix for the clustered 8-mer(s) in each group. We then found known transcription factors per 10 identified consensus that have aligned part of sequence with the consensus by using the MEME suite (Figure 5C).

3.18.3 Enriched motif search in enhancer region of the other end of loop anchor points

We checked whether the 32 enriched motifs found in conserved pcRNA domains are also over-represented in enhancer regions on the other end of the loop anchor points. The definitions for enhancer region and loop anchor points are described in previous Method section, "Identification of HiC loops that overlap pcRNAs".

We found pcRNA that overlaps with loop anchor points by using Bedtools intersect. The 32 enriched motifs were searched in both pcRNA transcribed sequence as well as enhancer region of the other end of overlapping loop anchor points. We counted a given motif only if the motif was found in both pcRNA and enhancer region. We also searched non-enhancer region of the other end of the loop anchor points as a control set. The counts were normalized by the total length of enhancer or non-enhancer region accordingly (**Figure 4D**).

4 Supplementary References

1. Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125, 315-326.

2. Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes & development 25, 1915-1927.

3. Down, T.A., Piipari, M., and Hubbard, T.J. (2011). Dalliance: interactive genome viewing on the web. Bioinformatics (Oxford, England) 27, 889-890.

4. Dunning, M.J., Smith, M.L., Ritchie, M.E., and Tavare, S. (2007). beadarray: R classes and methods for Illumina bead-based data. Bioinformatics (Oxford, England) 23, 2183-2184.

5. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473, 43-49.

6. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome research 22, 1760-1774.

7. Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols 4, 44-57.

8. Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-binding specificities of human transcription factors. Cell 152, 327-339.

9. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. Nucleic acids research 32, D493-496.

10. Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome research 12, 656-664.

11. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome research 12, 996-1006.

12. Kheradpour, P., and Kellis, M. (2014). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic acids research 42, 2976-2987.

13. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology 14, R36.

14. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods 9, 357-359.

15. Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48, 443-453.

16. Quinlan, A.R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al] 47, 11 12 11-11 12 34.

17. Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159, 1665-1680.

18. Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., et al. (2015). The UCSC Genome Browser database: 2015 update. Nucleic acids research 43, D670-681.

19. van Dongen, S. (2008). Graph Clustering Via a Discrete Uncoupling Process. SIAM Journal on Matrix Analysis and Applications 30, 121-141.

20. Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.P., and Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic acids research 41, e74.