nature portfolio

https://doi.org/10.1038/s41587-021-00936-1

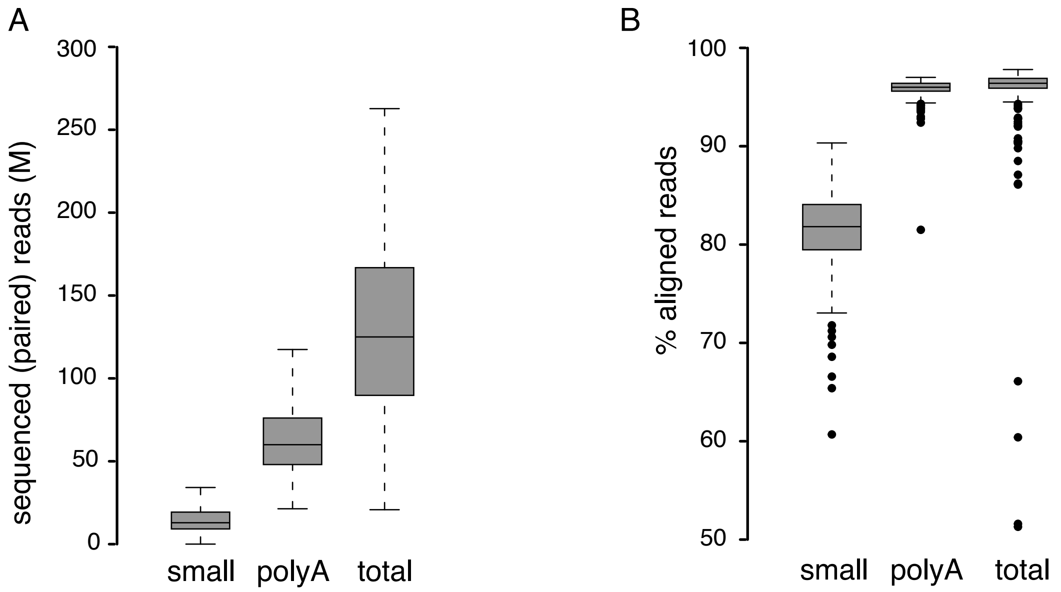
Supplementary information

The RNA Atlas expands the catalog of human non-coding RNAs

In the format provided by the authors and unedited

Supplementary Material

A. Supplemental Figures



C

Coverage of known transcriptome by blind meta-assembly

	GENCODE v33 protein-coding and lncRNA genes
Bases covered	67%
Exons with exact match	58%
Exons with overlap	86%
Exact matched transcripts	13%
Near matched isoforms	78%
Transcripts with other overlaps	5%
Genes with shared isoforms	43%
Genes with near matched isoforms	32%
Genes with other overlaps	13%
Overlapping transcripts	96%
Overlapping genes	88%

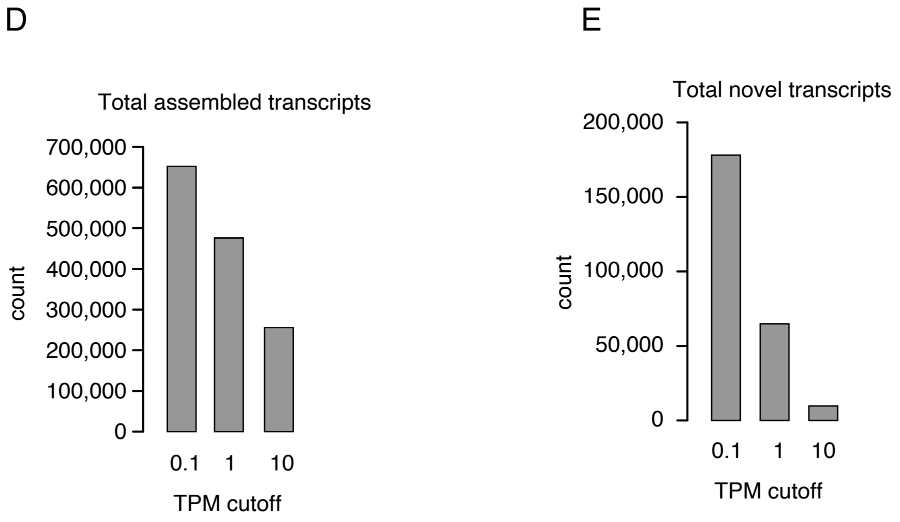


Figure S1. **Quality of the RNA sequencing data and coverage of known transcriptome**. (**A**). Distribution of total sequenced reads (small RNA-seq) or read pairs (polyA and total RNA-seq) across samples for the different data sets (small: 298 samples, polyA: 295 samples, total: 296 samples) and (**B**) the corresponding percentage of reads that were aligned to the human genome. Boxes extend from the 25th to the 75th percentiles, with the center lines at the median values; whiskers extend to the largest and smallest values no further than 1.5 times the inter-quartile range from the upper and lower box hinges, respectively. "Outlier" samples beyond the limits of whiskers are plotted as individual points. (**C**) Percentage of known transcriptome covered at base, exon, transcript and gene level by a 'blind' transcriptome assembly from RNA Atlas RNA-seq data. To evaluate coverage of the known transcriptome by samples used in this study, we produced a meta-assembly by merging individual assemblies across polyA and total RNA libraries without including any information on annotated transcripts (i.e. 'blind'). The generated transcript models overlapped 88% of the genes and 96% of the transcripts in GENCODE[1] v33 protein-coding and lncRNA loci. This analysis was performed as a quality control. To generate the primary RNA Atlas transcriptome, we did incorporate information on known transcripts (see Figures 1 and S2 and Methods). (**D and E**) To evaluate the impact of the minimum TPM required for transcripts to be included by StringTie[2] merge during the meta-assembly, we performed a transcriptome-size comparison for assemblies using different TPM cutoff values. When increasing the cutoff ten-fold from 0.1 to 1 and from 1 to 10, the number of assembled transcripts decreases 1.4-fold and 1.9-fold, respectively (**D**). We observed that while the coverage of the annotated transcriptome is high and similar for all 3 cutoffs (data not shown), the main differences between the different TPM cutoffs reside in the newly assembled transcripts (**E**). Our selection criterium of 1 TPM was based on our aim of being stringent enough as to avoid the assembly of too many unreliable transcripts models, while allowing for the detection of novel genes.

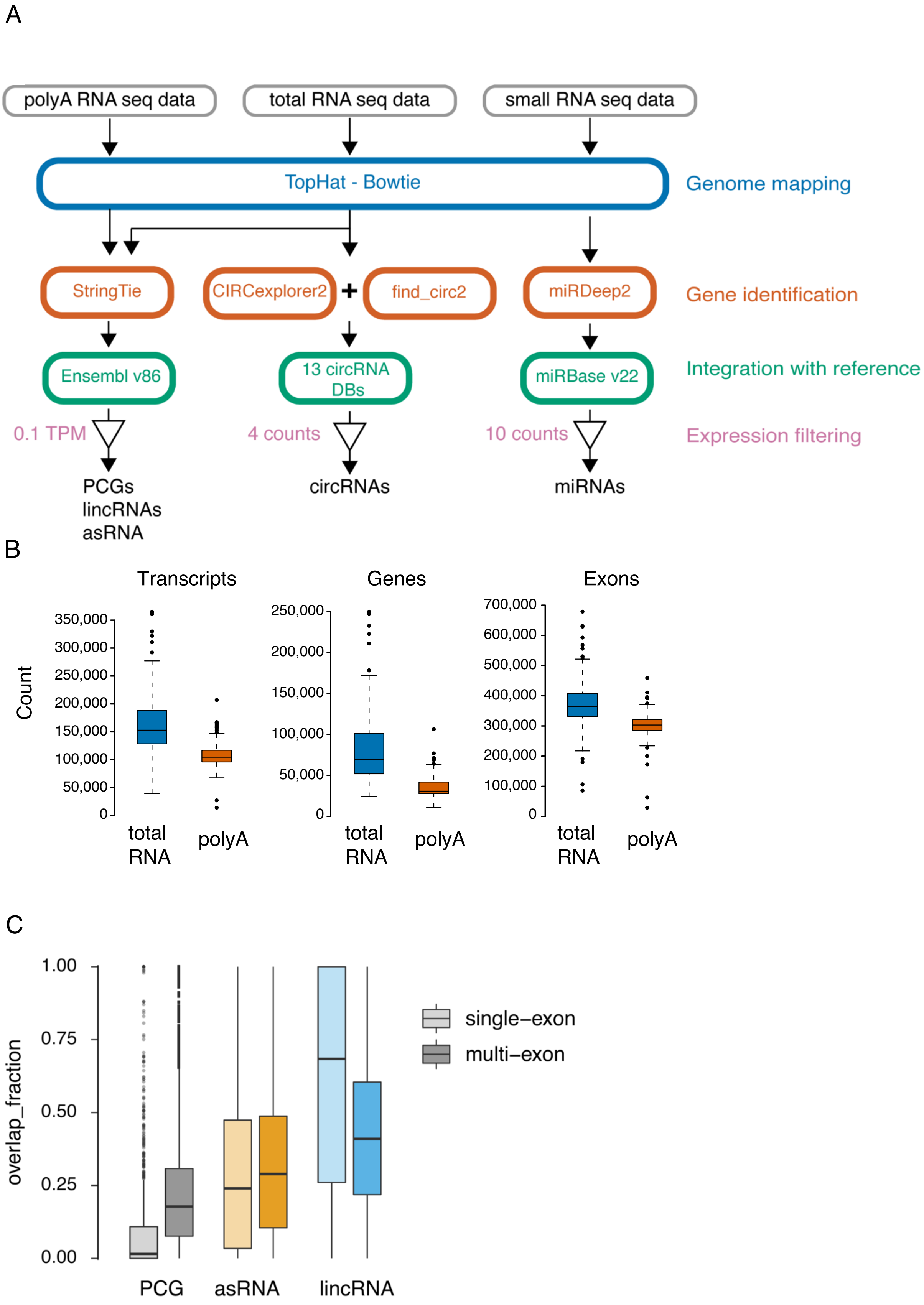

A
polyA RNA seq data
total RNA seq data
small RNA seq data
TopHat - Bowtie
Genome mapping
StringTie
CIRCexplorer2
+
find_circ2
miRDeep2
Gene identification
Ensembl v86
13 circRNA DBs
miRBase v22
Integration with reference
0.1 TPM
4 counts
10 counts
Expression filtering
PCGs
lincRNAs
asRNA
circRNAs
miRNAs
B
Transcripts
Genes
Exons
Count
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
250,000
200,000
150,000
100,000
50,000
0
700,000
600,000
500,000
400,000
300,000
200,000
100,000
0
total RNA
polyA
total RNA
polyA
total RNA
polyA
C
overlap_fraction
1.00
0.75
0.50
0.25
0.00
PCG
asRNA
lincRNA
single−exon
multi−exon

Figure S2. **Assembly of the RNA Atlas transcriptome**. (**A**) Overview of the workflow used. (**B**) Distribution of the number of transcripts, genes, and exons assembled in each sample in polyA (n=295) and total (n=296) RNA libraries (see Supplemental Table 2). (**C**) After expression filtering of PCGs, lincRNAs and asRNAs, overlap with repetitive elements was calculated to further filter assembled transcripts, as described in Methods. The plot shows the distribution of the fraction of exon sequence overlap with repeats aggregated at gene level. For each RNA biotype, the left and right boxes correspond to exons from single-exon genes (PCG=1,148, asRNA=1,405, lincRNA=37,382) and multi-exon genes (PCG=19,094, asRNA=7,091, lincRNA=13,196), respectively. (**B,C**) Boxes extend from the 25th to the 75th percentiles, with the center lines at the median values; whiskers extend to the largest and smallest values no further than 1.5 times the inter-quartile range from the upper and lower box hinges, respectively. "Outlier" genes beyond the limits of whiskers are plotted as individual points.

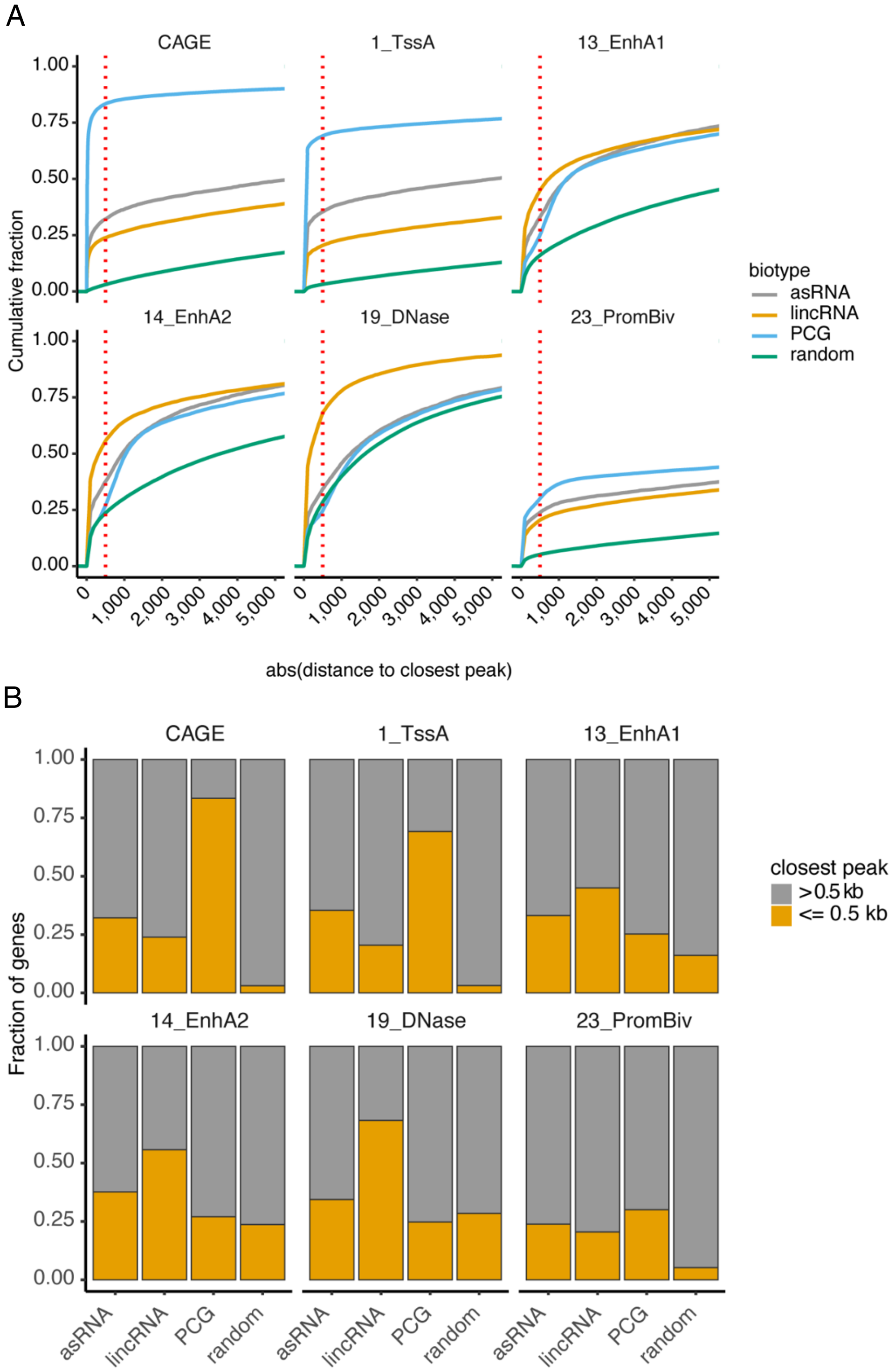
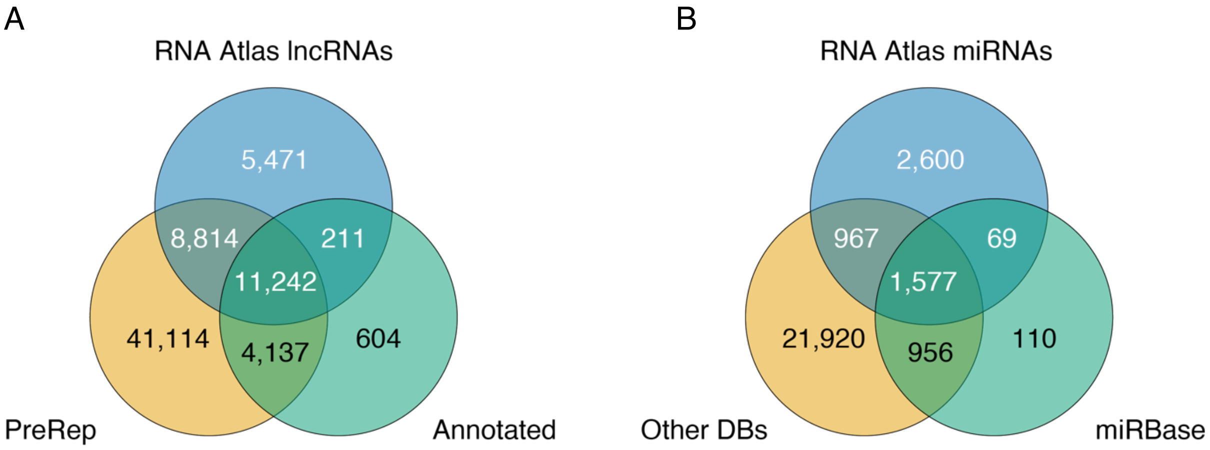


Figure S3. **Distance distribution for CAGE and various epigenomic marks used to define the RNA Atlas transcriptome.** (**A**) Cumulative distributions of closest peak's absolute distance for those molecular marks that were associated to at least 25% of RNA Atlas genes (these include CAGE[3] and the chromatin states[4] DNase, TssA, EnhA2, EnhA1, and PromBiv, see Methods section *Selection of the RNA Atlas transcriptome*) relative to the TSS of the different RNA biotypes and randomly selected genomic positions across non-repetitive DNA regions. Red dotted lines indicate the 500 bp distance cutoff applied to select the RNA Atlas transcriptome. (**B**) The fraction of genes that were positive for each of the 6 molecular features used. We note that these are molecular features of genes and are widely studied for this reason. Our results suggest that these features are indeed enriched in the presence of coding and non-coding genes. As expected, some features (CAGE and TSS_A) are more strongly associated with coding genes or highly expressed genes. We note that the dataset available to download from FANTOM5[3] is pre-filtered at a cutoff of 3 TPM, which is largely above the mean expression value of newly assembled lncRNA genes. Therefore, we expected lower CAGE-peak association rates with a large fraction of lncRNAs which are typically low expressed and tissue-specific genes. However, some features, including DNase and enhancer marks, were enriched for lncRNA-associated regions. We concluded that the presence of these features helped enrich our predicted candidate pool for true genes. We note that novel lincRNAs showed the highest enrichment with the EnhA1 and EnhA2 features, and more than 50% of these genes associated with a proximal EnhA2 peak. This too was expected as proximity to enhancer regions has been shown to have a significant predictive value for lncRNA function[5].



C

Summary of genes included in each dataset used

Dataset	Number of protein-coding genes	Number of lncRNAs
Ensembl v86 (equivalent to GENCODE v25)	20,343	14,707
Ensembl v99 (equivalent to GENCODE v33)	20,365	16,892
RefSeq Curated (v200)	19,364	5,160
PreRep (the union of CHESS, MiTranscriptome, FANTOM5,	17,227	58,855

BIGTranscriptome and model RefSeq records)		

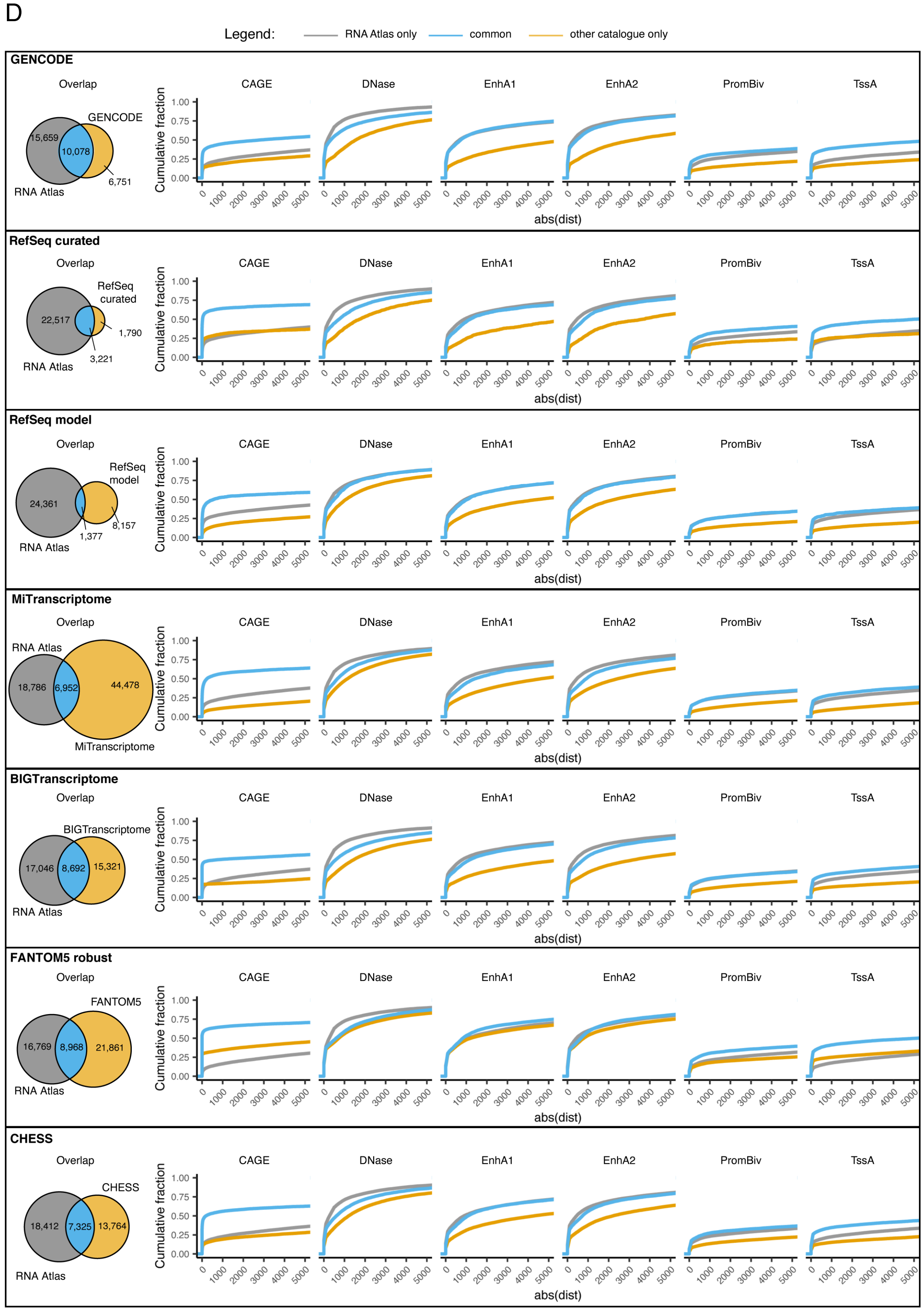
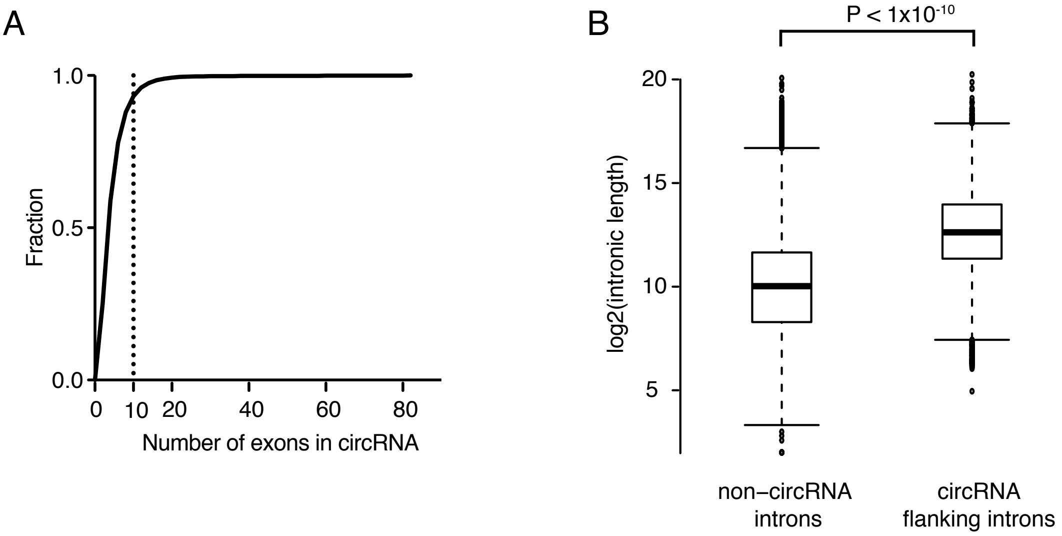

D
Legend:
RNA Atlas only
common
other catalogue only
GENCODE
Overlap
GENCODE
15,659
10,078
6,751
RNA Atlas
RefSeq curated
Overlap
RefSeq curated
22,517
1,790
3,221
RNA Atlas
RefSeq model
Overlap
RefSeq model
24,361
8,157
1,377
RNA Atlas
MiTranscriptome
Overlap
RNA Atlas
18,786
6,952
44,478
MiTranscriptome
BIGTranscriptome
Overlap
BIGTranscriptome
17,046
8,692
15,321
RNA Atlas
FANTOM5 robust
Overlap
FANTOM5
16,769
8,968
21,861
RNA Atlas
CHESS
Overlap
CHESS
18,412
7,325
13,764
RNA Atlas
CAGE
DNase
EnhA1
EnhA2
PromBiv
TssA
Cumulative fraction
0.00
0.25
0.50
0.75
1.00
0
1000
2000
3000
4000
5000
abs(dist)

Figure S4. Comparison of RNA Atlas lncRNAs and mature miRNAs with various up to date reference annotations. (**A**) Overlap between RNA Atlas lncRNAs, lncRNAs profiled by other efforts (combined transcriptome across FANTOM5[6], MiTranscriptome[7], BIGTranscriptome[8] and CHESS[9], named PreRep (Previously Reported) RNAs, see Methods), and lncRNAs annotated in GENCODE[1] and RefSeq[10] (named Annotated RNAs, see Methods). (**B**) Overlap between the set of candidate mature miRNAs studied in RNA Atlas (these include miRBase[11] miRNAs that are expressed in RNA Atlas samples and miRDeep2[12] predicted candidates from RNA Atlas small RNA-seq data) and different miRNA databases. "Other_DBs" include the union of the candidate miRNA set of FANTOM5[13] project, the miRCarta[14] annotation and the MirGeneDB[15] annotation (see Methods for details). (**C**) Table providing summary of number of protein-coding and lncRNAs included in each reference annotation or combined annotation included in this study for primary transcriptome assembly (Ensembl[16] v86, equivalent to GENCODE v25), and for comparison and final annotation of the derived RNA Atlas transcriptome (Ensemble v99 -equivalent to GENCODE v33- and RefSeq curated set, and the combination of FANTOM5 stringent set, CHESS, MiTranscriptome, BIGTranscriptome and model RefSeq, named PreRep). (**D**) Pairwise comparisons between RNA Atlas lncRNAs and other lncRNA catalogs and association to different marks for the subset of genes in common or unique for each dataset. Similar to the analysis shown in Figure S3, here we show comparisons of the distance distribution to 6 relevant molecular marks for each subset of genes resulting from the individual overlaps between RNA Atlas lncRNAs and lncRNAs from other lncRNA catalogs. The results show that, in all cases, lncRNAs common to both annotations under comparison are more enriched for CAGE peaks and TSS-proximal promoter chromatin state (TssA) relative to genes found in either annotation alone, whereas common lncRNAs and those exclusively found in RNA Atlas show higher association with enhancer chromatin states and DNase compared to lncRNAs absent in RNA Atlas.



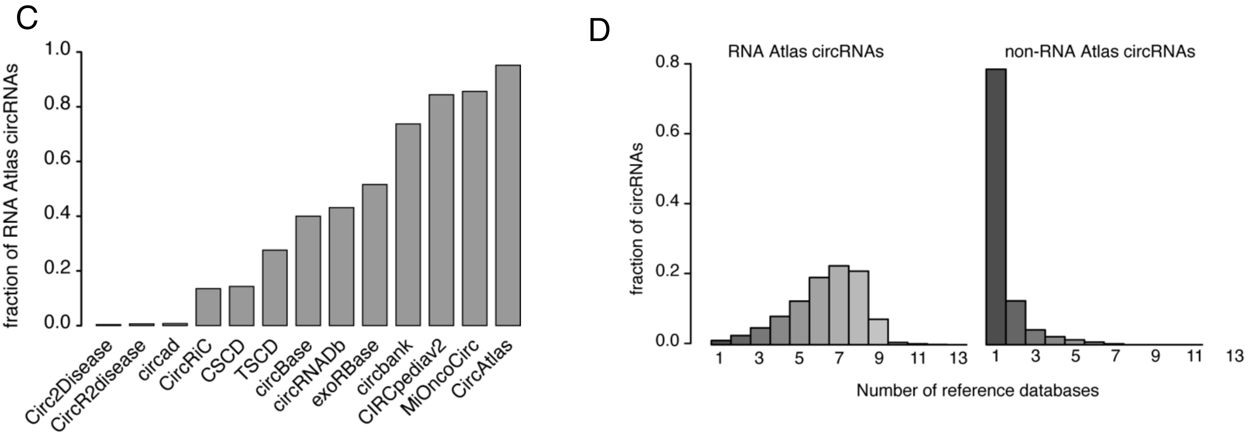


Figure S5. **Properties of circRNAs**. (**A**) Cumulative fraction of number of exons spanned per circRNA. (**B**) Intron length distributions for introns flanking circRNAs (n=65,534) and introns not flanking any circRNA (n=499,999). Statistical significance of the difference was assessed with a two-sided Wilcoxon rank-sum test. Boxes extend from the 25th to the 75th percentiles, with the center lines at the median values; whiskers extend to the largest and smallest values no further than 1.5 times the inter-quartile range from the upper and lower box hinges, respectively. "Outlier" intron lengths beyond the limits of whiskers are plotted as individual points. (**C**) Overlap between RNA Atlas circRNAs and circRNAs reported in other efforts. A total of 98% of RNA Atlas circRNAs were also predicted elsewhere, but RNA Atlas identified 446 circRNAs that were not previously annotated or predicted. (**D**) Most RNA Atlas circRNAs were identified in multiple other databases.

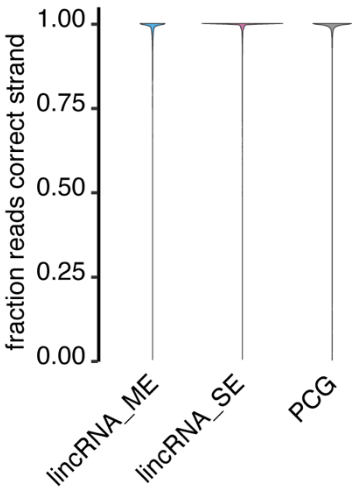


Figure S6. **Analysis of strandedness of single- and multi-exon genes**. Fraction of reads mapping to the correct genomic strand in the sample with maximum expression for exons from multi- and single-exon lincRNAs (lincRNA_ME, and lincRNA_SE, respectively) and exons from PCGs.

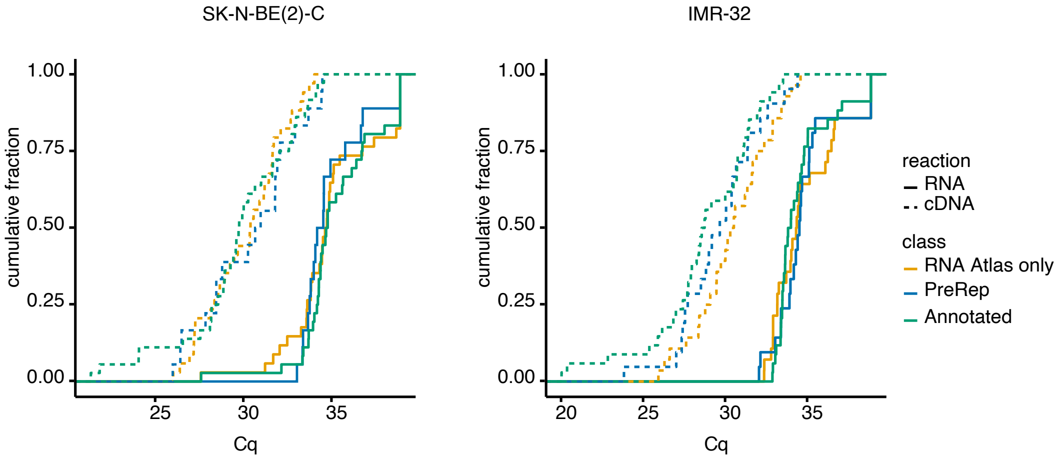


Figure S7. **qPCR validation of single-exon genes**. For each of the cell lines used, two qPCR reactions were performed, one using total RNA, and one using cDNA as template. Cumulative fractions of Cq-values obtained for each reaction are shown separately for RNA Atlas only (orange), PreRep (blue) and Annotated (green) single-exon genes.

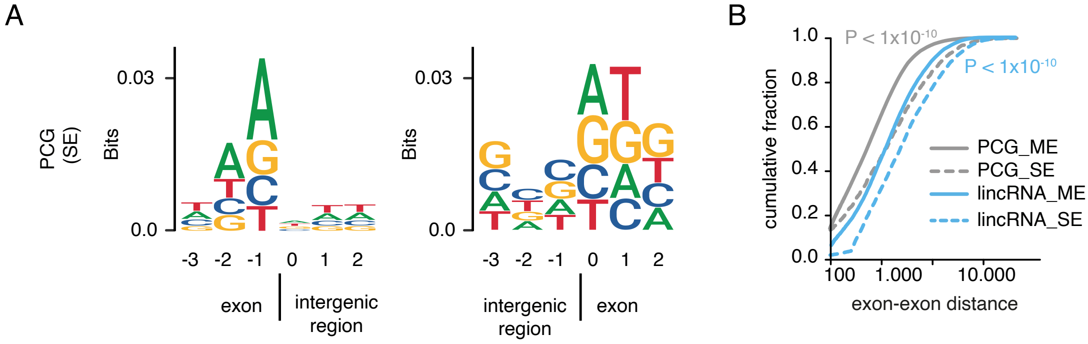


Figure S8. **Expected properties of single-exon genes.** (**A**) Nucleotide frequencies at intergenic/exonic boundaries for single-exon PCGs. (**B**) Cumulative fraction of the distance to the closest exon for multi-exon and single-exon PCG exons (grey-colored lines, PCG_ME and PCG_SE, respectively), and multi-exon and single-exon lincRNA exons (light blue-colored lines, lincRNA_ME and lincRNA_SE, respectively). p-values were calculated using two-sided Wilcoxon rank-sum tests.

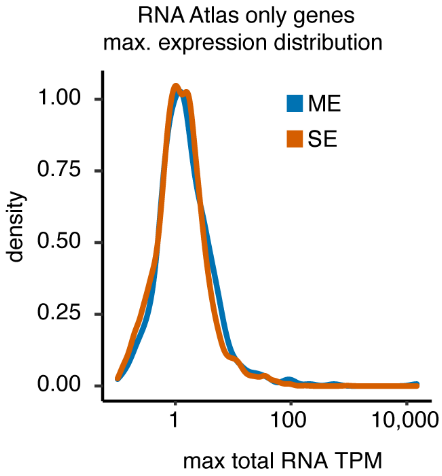


Figure S9. Comparison of distributions of maximum expression across samples between RNA Atlas only multi- and single-exon genes. The maximum total RNA TPM across all samples was retrieved for each RNA Atlas only gene. Density distributions were plotted separately for multi-exon and single-exon genes. The observed almost identical profiles suggest that low expression is unlikely to explain the absence of junction reads overlapping RNA Atlas only single-exon genes.

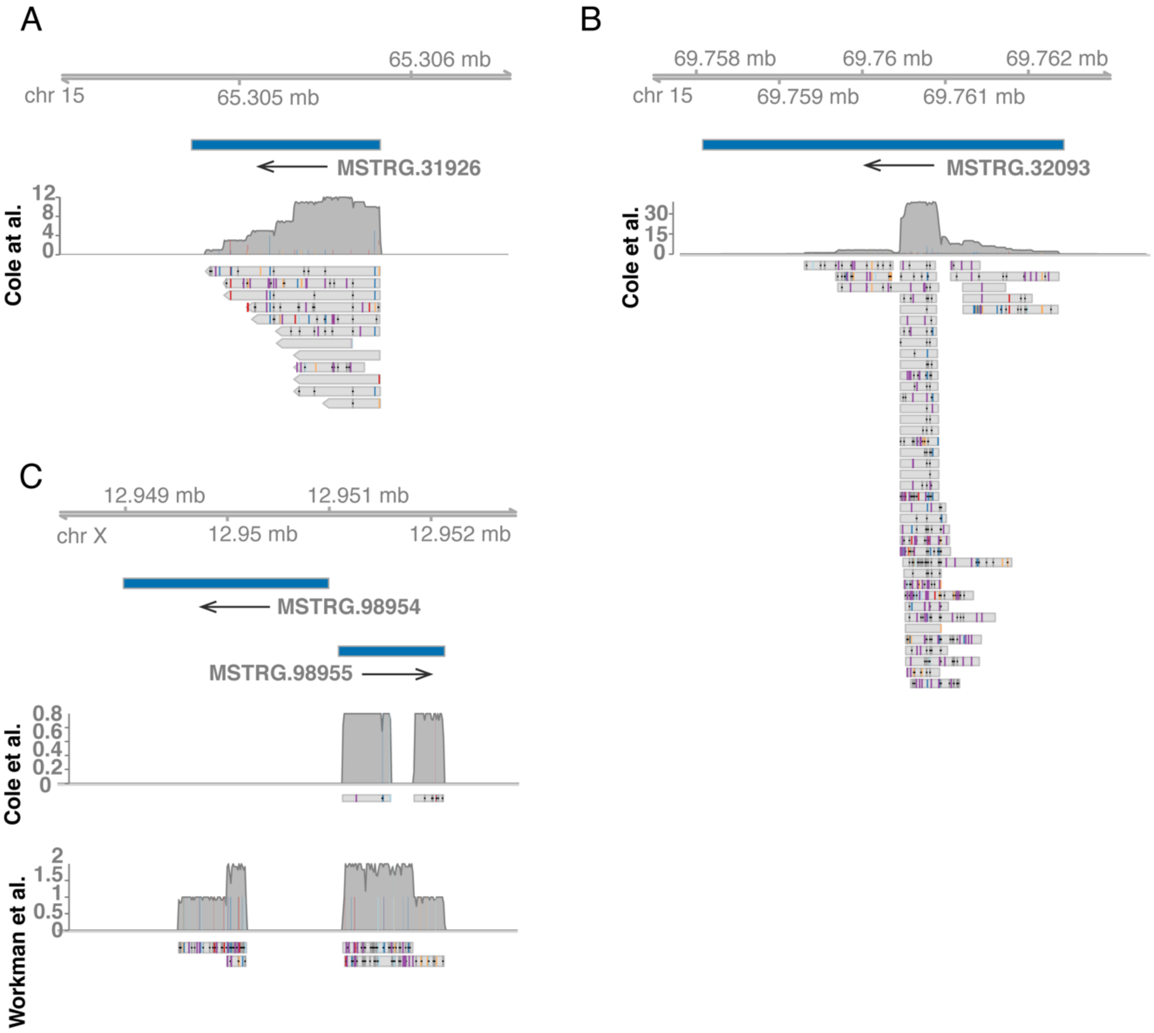


Figure S10. Examples of single-exon RNA Atlas only genes overlapped by ONT sequencing reads. Uniquely mapped and strand-matched ONT reads overlapping RNA Atlas only single exon genes were retrieved from 4 public ONT datasets (see Methods). All identified overlapping reads were non-spliced, supporting the single-exon status of these genes. Examples shown here illustrate the overlaps found for 4 RNA Atlas only single-exon genes. (**A**) Gene MSTRG.31926 was overlapped by 12 reads in the Cole et al.[17] dataset. (**B**) Gene MSTRG.32093 was overlapped by 46 reads in the Cole et al.[17] dataset. (**C**) Genes MSTRG.98954 and MSTRG98955, transcribed in opposite strands, were overlapped by 2 reads each in the Workman et al.[18] dataset. Additionally, MSTRG.98955 was overlapped by 2 reads in the Cole et al.[17] dataset. Note that, in all cases, ONT reads match the direction of the overlapped gene.

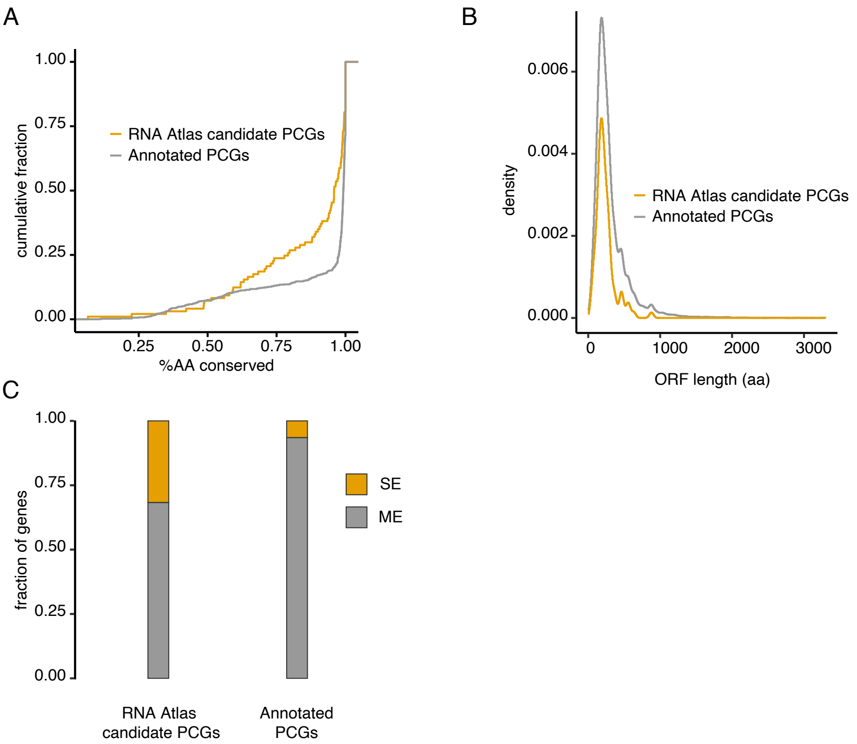


Figure S11. Properties of RNA Atlas candidate protein-coding genes compared to Annotated protein-coding genes. (**A**) Empirical cumulative distribution of the percentage of amino acids conserved between human and chimpanzee (see Methods) (**B**) ORF length distribution. The length distribution for the largest predicted ORF from each candidate protein-coding gene was compared to the mean ORF length across isoforms from Annotated protein-coding genes in the RNA Atlas set. (**C**) Fractions of single-exon (SE) and multi-exon (ME) genes for RNA Atlas candidate coding genes and Annotated coding genes.

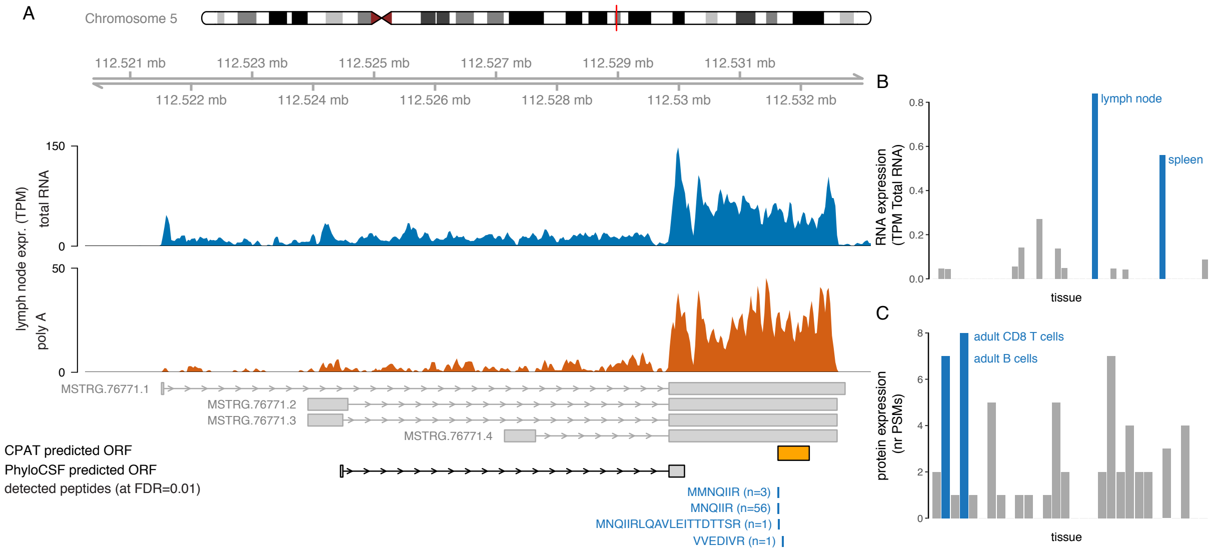


Figure S12. **Example of a candidate new protein-coding gene with matching peptides from public mass spectrometry data**. (**A**) The different tracks show genomic coordinates of the identified gene, coverage profiles from total and polyA

RNA-seq, transcript structure of the 4 assembled isoforms, predicted ORFs by CPAT and PhyloCSF and detected peptides matching the CPAT predicted ORF. Expression of this gene is enriched in immune related samples (blue bars) both at RNA (**B**, expression for isoform 3 is shown) and peptide (**C**) level.

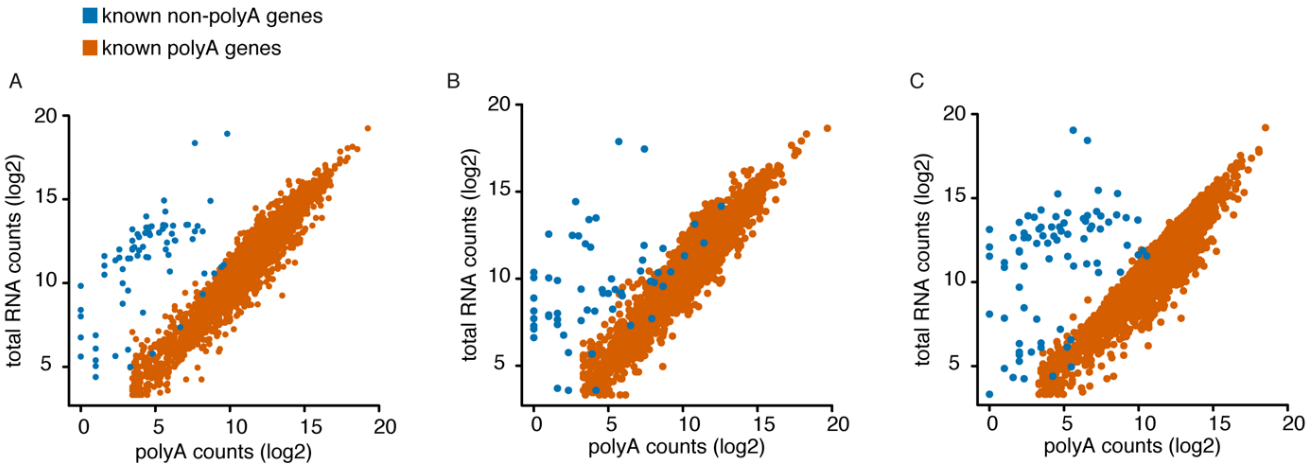


Figure S13. **Correlations between polyA and total RNA-seq counts for known polyadenylated and non-polyadenylated genes**. Scatter plots between log2 counts from total RNA-sequencing and polyA RNA-sequencing for known polyadenylated (orange) and non-polyadenylated (blue) genes in a cell type (**A**: human umbilical vein endothelial cell), a tissue (**B**: distal colon), and a cell line (**C**: K562).

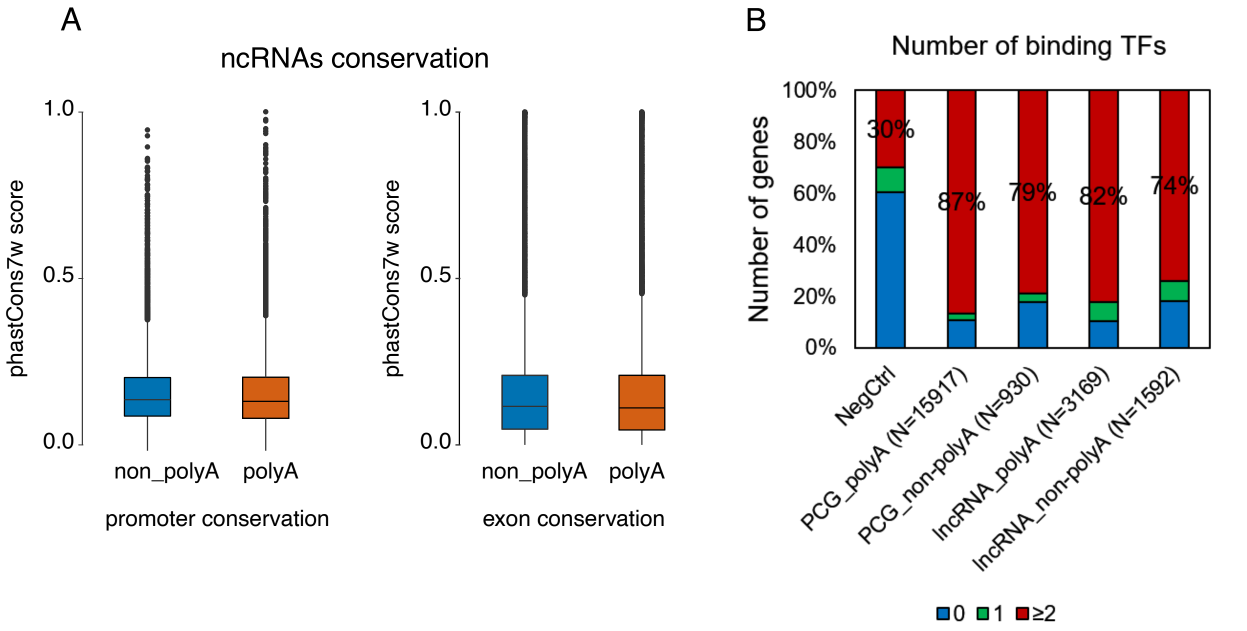


Figure S14. **Conservation and TF occupancy for polyadenylated and non-polyadenylated lncRNAs**. (**A**) Conservation was evaluated and compared between polyadenylated (n=12,104) and non-polyadenylated (n=9,909) lncRNAs by retrieving the mean phastCons7way scores across gene promoter regions (1 kb upstream and 1 kb downstream the TSS of the most abundant transcript) and across unique continuous exonic regions of genes (i.e. if two exons from different transcripts of the same gene overlap partially they are collapsed into a single exonic region). Polyadenylated and non-polyadenylated lncRNAs show similar cross-species conservation scores at both promoter (**left**) and exonic (**right**) levels. Boxes extend from the 25th to the 75th percentiles, with the center lines at the median values; whiskers extend to the largest and smallest values no further than 1.5 times the inter-quartile range from the upper and lower box hinges, respectively. "Outlier" genes beyond the limits of whiskers are plotted as individual points.

(**B**) We also compared transcription factor occupancy of 2kb promoters of non-polyadenylated and polyadenylated PCG and lncRNA RNA Atlas genes with functional evidence (see section "*Evidence for transcriptional and post-transcriptional regulation by long noncoding RNAs*" in Results) to that of their 5' flanking regions used as negative control (NegCtrl, these flanking regions are 10 kb upstream from the TSSs and their lengths are 2 kb) by mapping these regions to ENCODE-identified binding sites (see Methods section *Transcription factor binding analyses*). The results showed a very significant—with p-value calculations exceeding machine precision—2.5-fold increase in the binding of multiple transcription factors in both polyadenylated and non-polyadenylated RNA Atlas lncRNA promoters versus 5'-end flanking regions. This suggests that predicted proximal promoters for both polyadenylated and non-polyadenylated lncRNA genes are dramatically more likely to be targeted by transcription factors than their flanking regions.

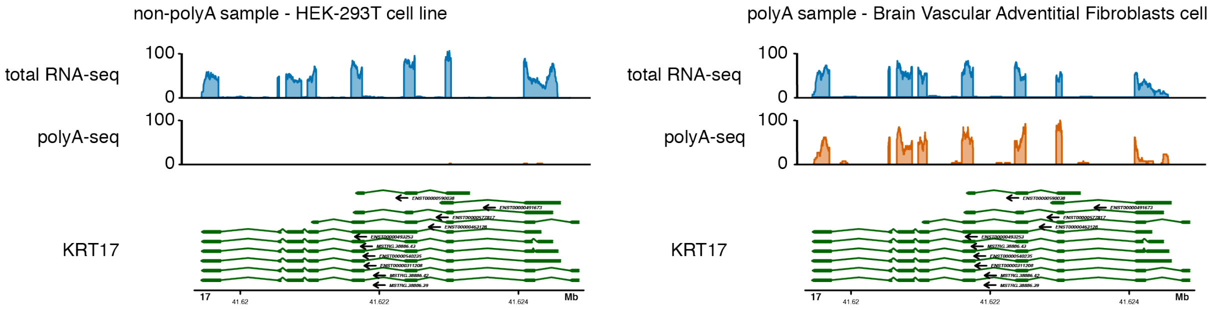


Figure S15. **Example of a gene (KRT17) with variable polyadenylation across samples**. In this case, in contrast to the one shown in Figure 3 H, no clear differences between the expression patterns at transcript level for polyadenylated and non-polyadenylated samples are observed. Coverage profiles from total RNA-sequencing and polyA-sequencing are shown for a non-polyadenylated sample (**left**) and a polyadenylated sample (**right**).

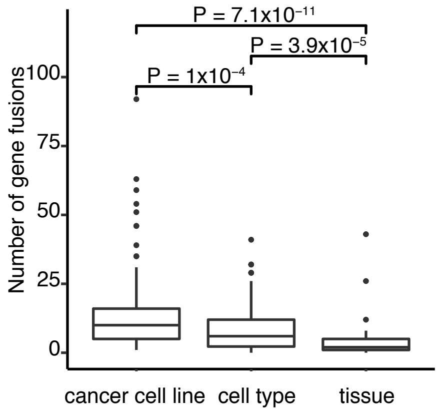


Figure S16. **Numbers of predicted fusion genes per sample.** Distribution of the number of predicted fusion genes identified per sample are shown for the different sample types (cancer cell lines: n= 87, cell types: n= 158, tissues: n=44). The significance of differences in number of fusion genes between sample types was assessed with two-sided Wilcoxon rank-sum tests. Boxes extend from the 25th to the 75th percentiles, with the center lines at the median values; whiskers extend to the largest and smallest values no further than 1.5 times the inter-quartile range from the upper and lower box hinges, respectively. "Outlier" intron lengths beyond the limits of whiskers are plotted as individual points.

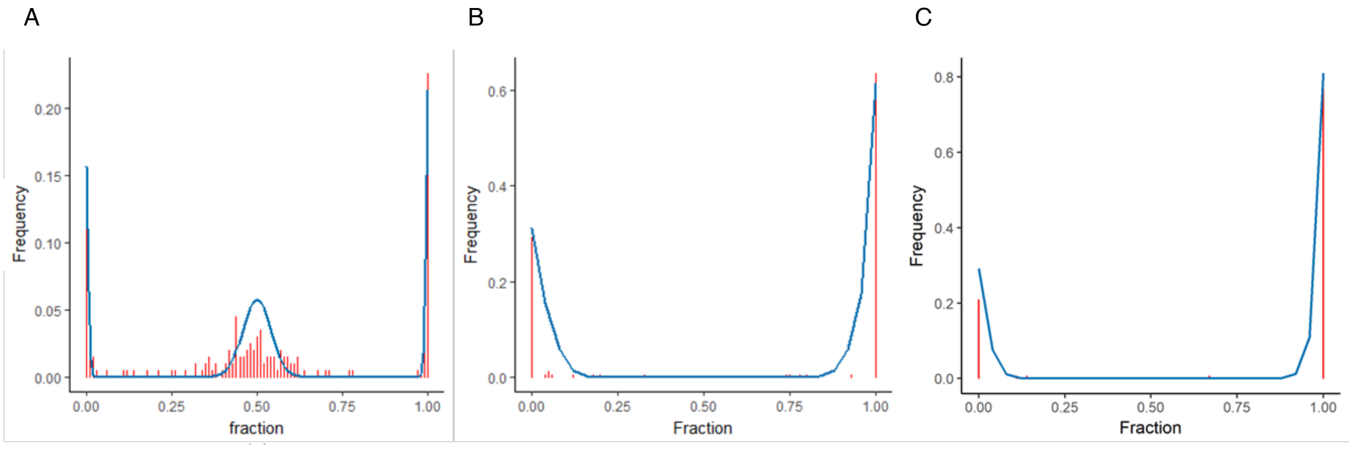


Figure S17. **Imprinting analyses.** Observed (**red**) and expected (**blue**) frequencies of alternative alleles for two SNP positions. (**A**) An example of a non-imprinted SNP in FAM20C (rs36138803, degree of imprinting (i) = 0). A clear heterozygous peak is present, i.e. a set of samples with an alternative allele fraction of roughly 50%. (**B**) A significantly imprinted SNP in PLAGL1 (rs17073273, adj. p-value = 9.2x10^{-49} and i = 0.97). Here the heterozygous peak is eliminated and virtually no heterozygous samples are observed. (**C**) A significantly imprinted SNP in MIR381HG (rs35844276, adj. p-value = 7.5x10^{-34} and i = 0.99). For each evaluated SNP, a likelihood-ratio-test was used to assess the significance of imprinting; p-values were adjusted for multiple testing using the Benjamini–Hochberg correction.

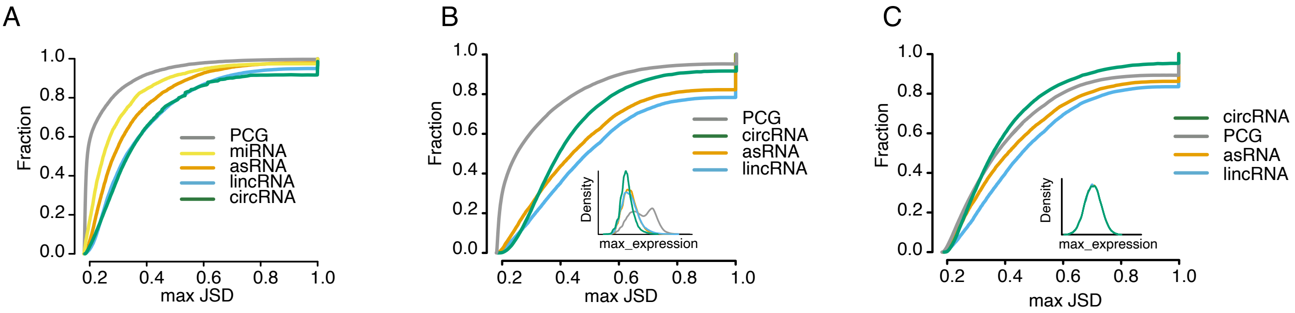


Figure S18. **Cell-type specificity scores for the different RNA biotypes before and after correcting for expression distributions.** In all cases, the specificity score for a given feature was calculated as the maximum Jensen-Shannon divergence across cell-types. (**A**) Cumulative fraction of the specificity score at gene level for all RNA biotypes without correcting for differences in abundance. (**B**) Cumulative fraction of the specificity scores for all forward-spliced junctions from PCGs, lincRNAs and asRNAs, and back-spliced junctions from circRNAs. The inset plot shows the density distributions for the maximum expression across samples for the different biotypes. The expression was calculated as junction counts scaled by library size. (**C**) Similar to (B), but correcting for differences in abundance distribution by performing a directed subsampling of junctions from the different biotypes to match a common expression distribution, as shown in the inset plot.

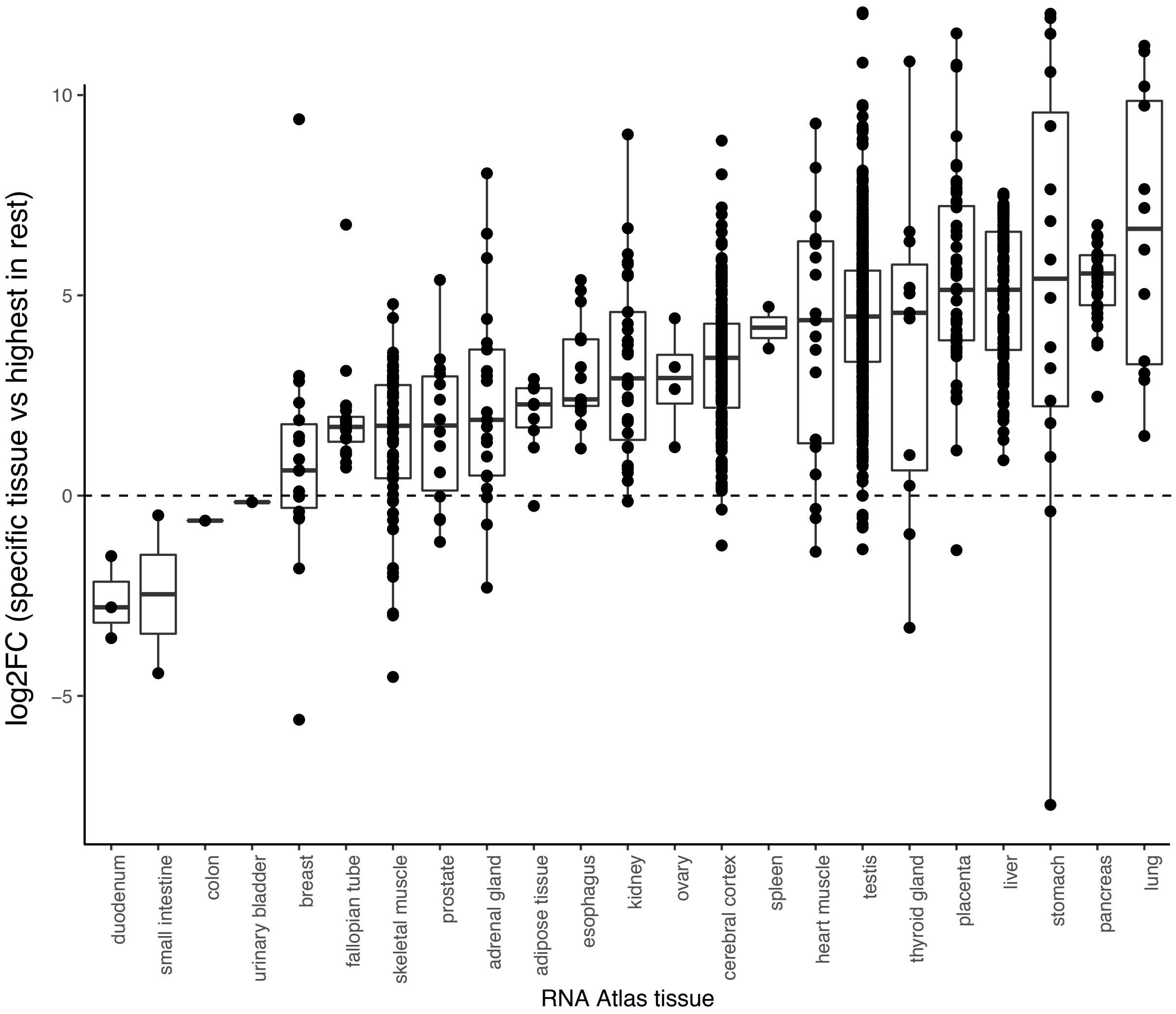


Figure S19. **Cross validation of tissue-specific markers selected from the Human Protein Atlas**. The y-axis shows the log2 fold-change between the expression of the tissue-specific marker in the matching RNA Atlas tissue and its highest expression among the remaining 22 tissues; duodenum (n=3), small intestine (n=2), colon (n=1), urinary bladder (n=1), breast (n=18), fallopian tube (n=20), skeletal muscle (n=65), prostate (n=14), adrenal gland (n=21), adipose tissue (n=10), esophagus (n=15), kidney (n=31), ovary (n=4), cerebral cortex (n=203), spleen (n=2), heart muscle (n=19), testis (n=704), thyroid gland (n=11), placenta(n=52), liver (n=120), stomach (n=16), pancreas (n=29), lung (n=12). Boxes extend from the 25th to the 75th percentiles, with the center lines at the median values; whiskers extend to the largest and smallest values no further than 1.5 times the inter-quartile range from the upper and lower box hinges, respectively. Fold changes for all markers are plotted as individual points.

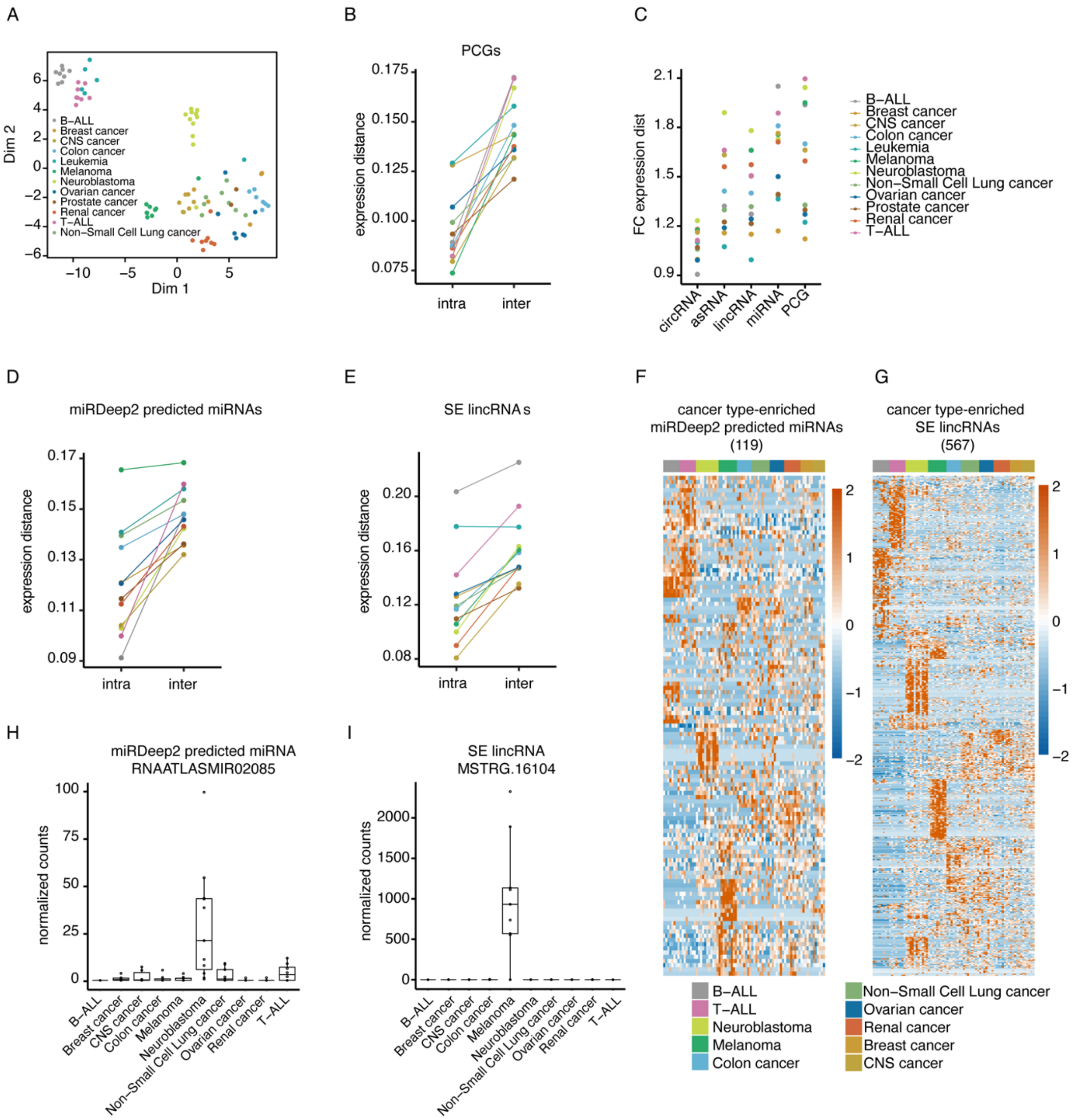


Figure S20. **The association between cancer type and expression-distance.** (**A**) t-SNE plot of the RNA Atlas cancer cell lines based on PCG expression. Samples were colored according to the cancer type. (**B**) For each cancer type, we showed the median PCG expression-based distances between pairs of samples from the same a cancer type (intra-distances) and across cancer types (inter-distances). (**C**) The distribution of fold changes between median inter- and intra-distances calculated based on expression of the different RNA biotypes. (**D**) Median intra- and inter-distances based on expression of mirDeep2 predicted miRNAs. (**E**) Median intra- and inter-distances based on expression of single-exon lincRNAs. (**F**) Expression heatmap for mirDeep2 predicted miRNAs significantly upregulated in each of the cancer types. (**G**) Expression heatmap for single-exon lincRNAs significantly upregulated in each of the cancer types. (**H,I**) Examples of an Neuroblastoma-specific miRDeep2 predicted candidate miRNA (**H**) and an Melanoma-specific single-exon lincRNA (**I**). Expression distributions are shown as boxplots of normalized counts as computed by the DESeq function for each cancer type: B-ALL (n=8), Breast cancer (n=6), Central Nervous System cancer (n=6), Colon cancer (n=7), Melanoma (n=9), Neuroblastoma (n=11), Non-Small Cell Lung cancer (n=9), Ovarian cancer (n=7), Renal cancer (n=8), T-ALL (n=8). The boxes extend from the 25th to the 75th percentiles, with the center lines at the median values; the whiskers extend to the largest and smallest values no further than 1.5 times the inter-quartile range from the upper and lower box hinges, respectively. "Outlier" samples beyond the limits of whiskers are plotted as individual points.

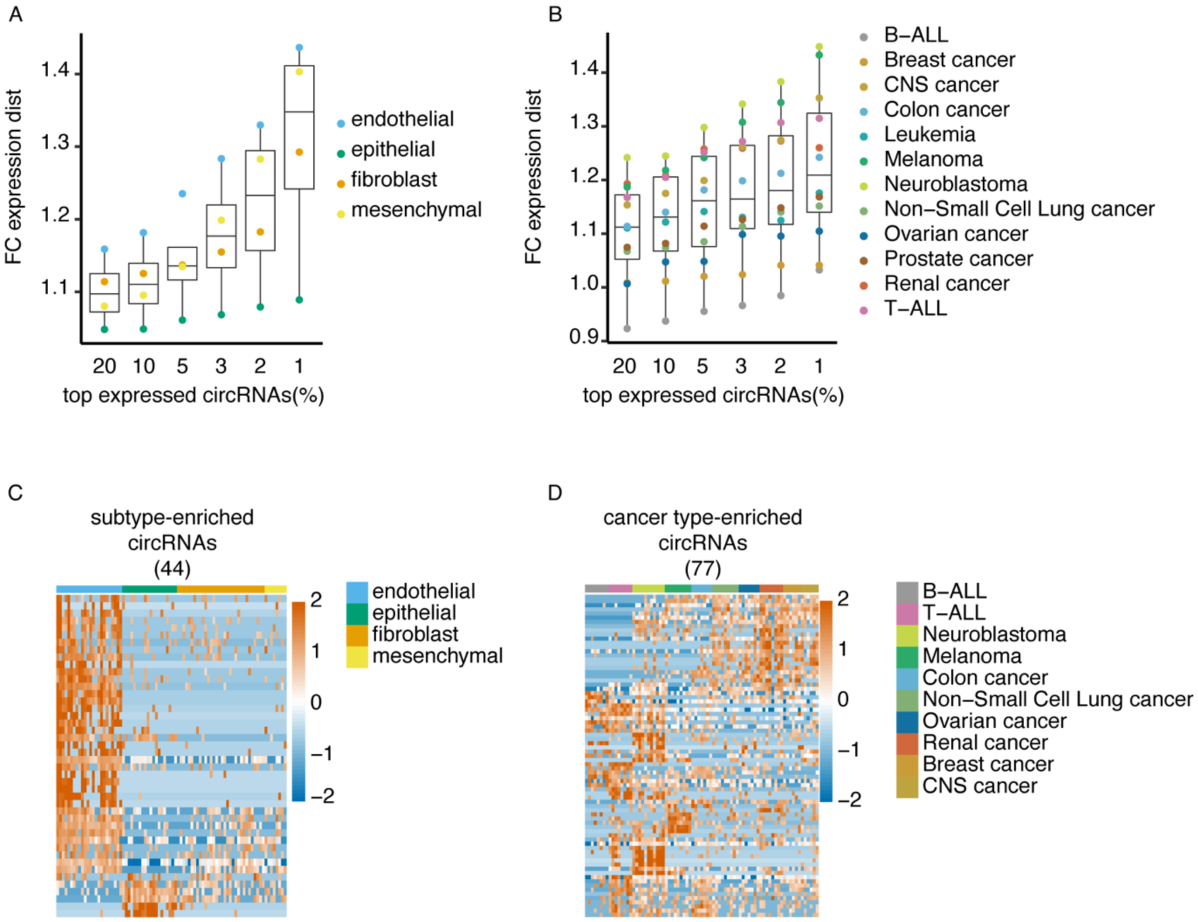


Figure S21. **Analyses of circRNA intra- and inter- expression distances for biological subtypes and cancer types.** (**A and B**) Impact of selecting subsets of abundant circRNAs across samples over the fold change in expression distances within and between cell subtypes (**A**), and within and between cancer types (**B**). Boxes extend from the 25th to the 75th percentiles, with the center lines at the median values; whiskers extend to the largest and smallest values no further than 1.5 times the inter-quartile range from the upper and lower box hinges, respectively. All fold changes are plotted as individual points. (**C and D**) Heatmaps of identified differentially expressed circRNAs for each of the 4 subtypes studied (**C**) and for each of the cancer types (**D**).

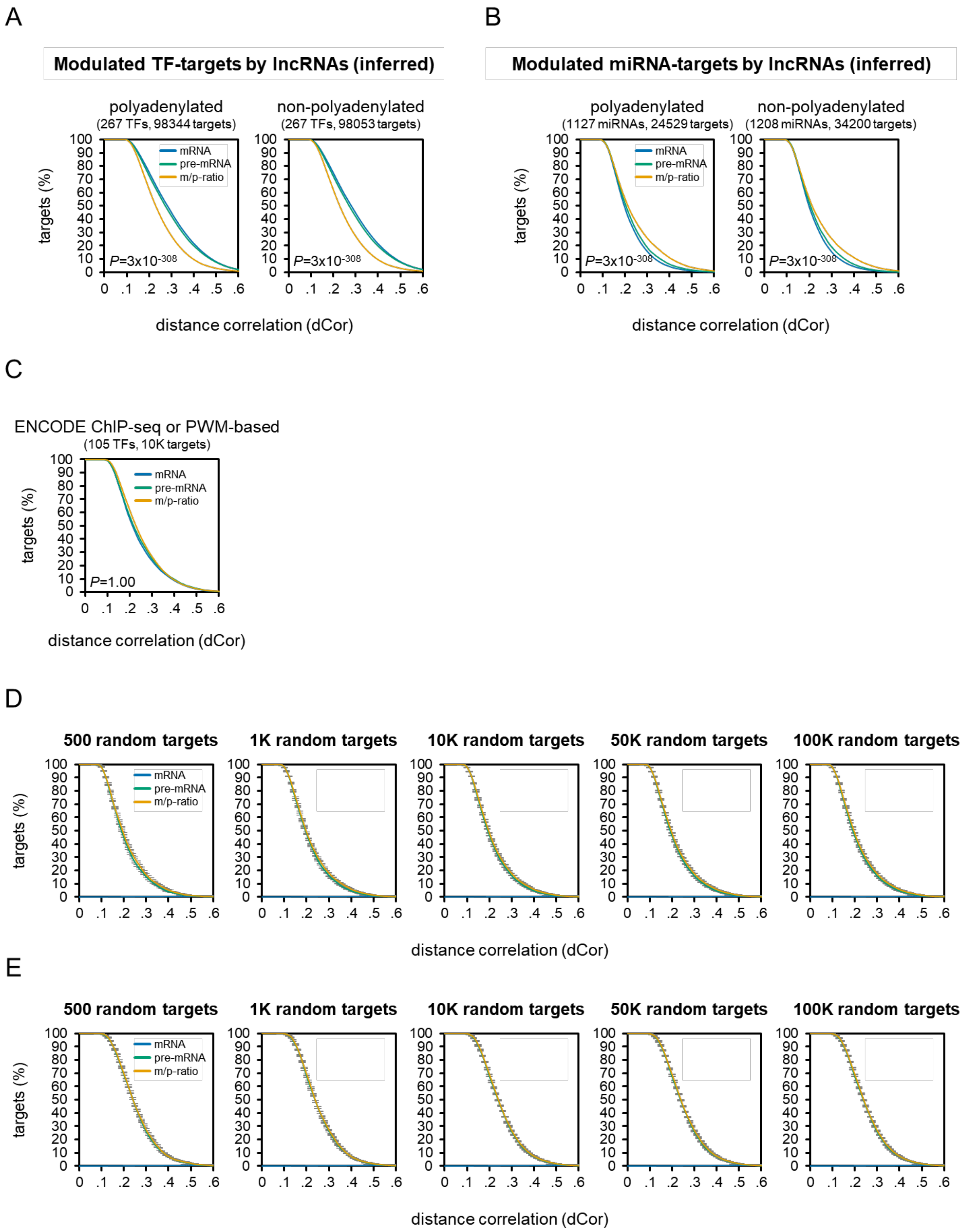


Figure S22. **Extended comparisons of distance correlations between regulators and target mRNA, pre-mRNA, or m/p ratio expression profiles.** As in Figures 6 B and 6C for multiple lncRNA biotypes, when considering predicted transcriptional (**A**) and post-transcriptional (**B**) targets for either polyadenylated or non-polyadenylated lncRNAs, we found similar and significant correlation deviations between regulators and their target pre-mRNA, mRNA, and m/p ratio profiles when predicted to be modulated by lncRNAs. (**C**) In contrast, a list of 10K predicted TF-target pairs with no supporting evidence from expression were randomly drawn and did not show correlation deviations between TF profiles and profiles of their target pre-mRNA and m/p-ratio. Note that the p-values in (A-C) are the geometric mean of the one-sided p-values, estimated by the paired Student’s T test, for the differences between pre-mRNA - m/p-ratio and mRNA - m/p-ratio correlations. We used permutation testing to generate randomized regulator-target pair sets, including 500, 1K, 10K, 50K, and 100K pairs for TFs (**D**) and miRNAs (**E**) individually. This process was repeated 10 times for each setand data are presented as mean values across replicates +/- SEM. The results suggested that the differences in distance correlation between regulator expression and target mRNA, pre-mRNA, or m/p-ratio estimates are independent of the number of regulator-target pairs, and therefore an increase in this number does not affect distance correlations.

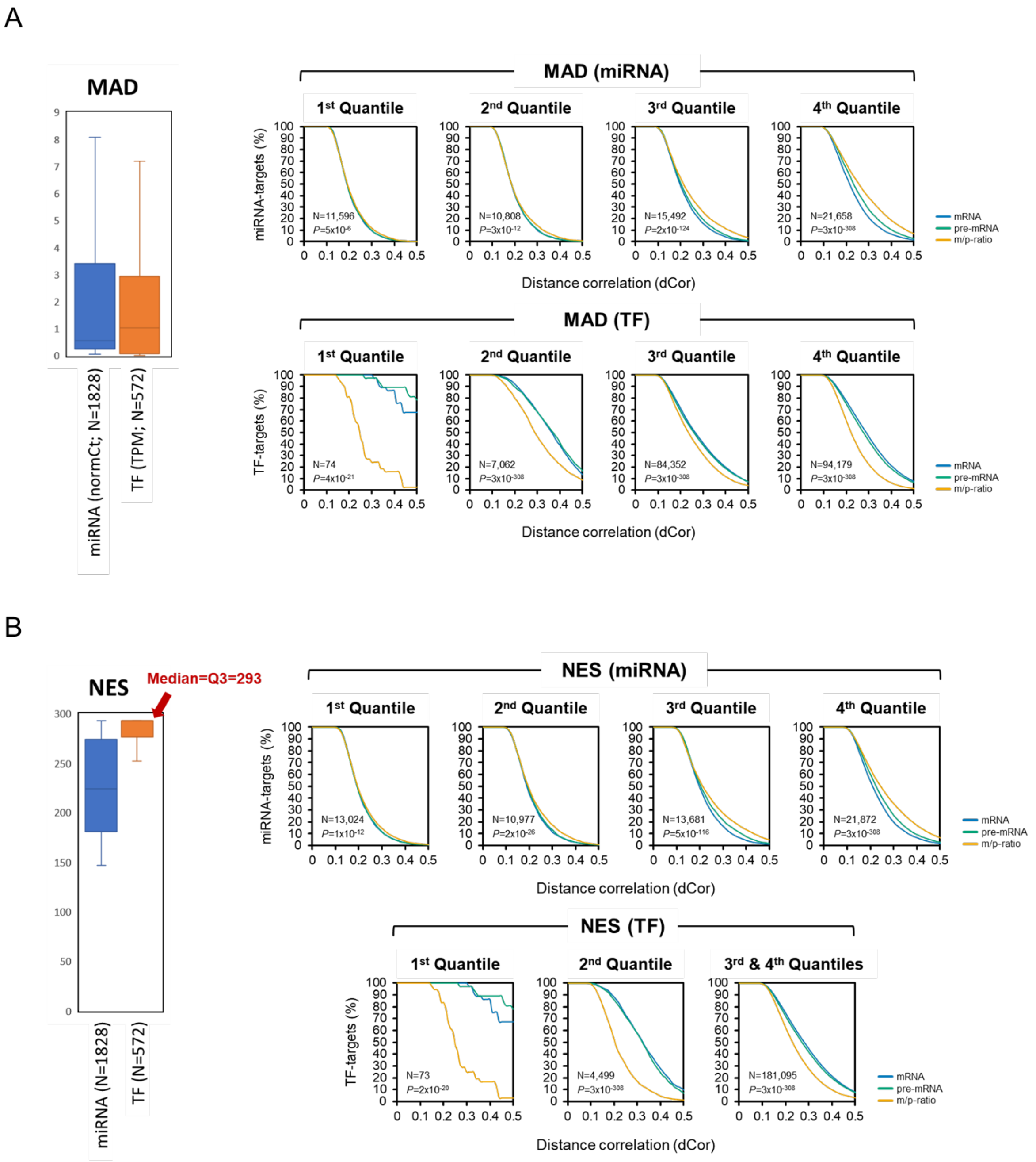


Figure S23. Comparisons of correlation deviations for regulators across expression feature quantiles based on target mRNA, pre-mRNA, or m/p-ratio expression profiles. LongHorn-predicted regulator-target interactions were partitioned into 4 quantiles with low (1st quantile) to high (4th quantile) regulator expression variability based on median absolute deviation (MAD) (**A**) or with lower (1st quantile) to higher (4th quantile) number of expressed samples (NES) that expressed regulators (**B**). On the left, boxplots display the data quantiles for regulators, either TFs or miRNAs, while different expression features, including MAD and NES, were applied. The number of regulators is given in parentheses. In general, miRNAs exhibited more expression variability and were expressed in fewer RNA Atlas samples than TFs. Because more than half of the TFs were expressed in all 293 RNA Atlas

samples, the median and the Q3 of their NES were the same; therefore, TFs fell into these two NES quantiles were combined. On the right, we plotted cumulative descending distributions of distance correlations for regulator-target interactions in each quantile using target mRNA, pre-mRNA, or m/p ratio expression profiles. The number of interactions and the p-values of correlation deviations were indicated within each panel. Note that these p-values are the geometric mean of the one-sided p-values, estimated by the paired Student's T test, for the differences between pre-mRNA - m/p-ratio and mRNA - m/p-ratio correlations. The y-axis shows the percentages of regulator-target interactions had greater distance correlations than the cutoffs indicated in the x-axis. TFs had fewer predicted targets in the 1st MAD and NES quantiles. Only regulators expressed at a sufficiently high level (MAD>0) were considered. The results confirmed that TFs and miRNAs, whose regulation were evidenced by LongHorn-predicted lncRNA modulators, were respectively less and more correlated with their target's m/p expression ratios across MAD and NES quantiles. In addition—and as expected—the results suggested that higher- and wider-expressed miRNAs showed these trends more strongly than poorly-expressed miRNAs and miRNAs that were identified in fewer samples, respectively. In general, higher- and wider-expressed miRNAs were better correlated with their target expression profiles and were expected to more tightly regulate each individual target, as explicitly shown by Mukherji et al. (2011)[19]. Boxes extend from the 25th to the 75th percentiles, with the center lines at the median values; whiskers extend to the largest and smallest values no further than 1.5 times the inter-quartile range from the upper and lower box hinges, respectively.

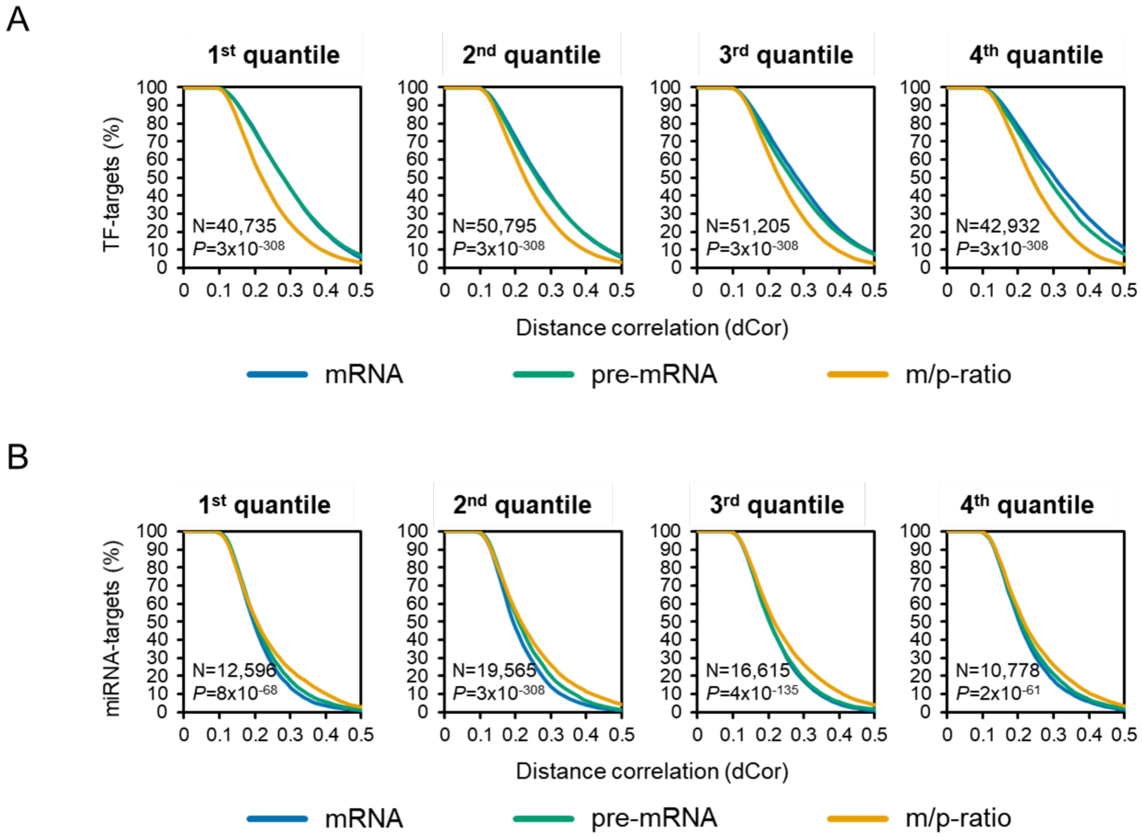


Figure S24. Comparisons of correlation deviations for targets in different MAD quantiles using their mRNA, pre-mRNA, or m/p-ratio expression profiles. Similar to Figure S23, LongHorn-predicted regulator-target interactions were partitioned into 4 quantiles with low (1st quantile) to high (4th quantile) target expression variability based on median absolute deviation (MAD). We plotted cumulative descending distributions of distance correlations for TF-target (**A**) and miRNA-target (**B**) interactions in each quantile using target mRNA, pre-mRNA, or m/p ratio expression profiles. All tested targets were expressed at a sufficiently high level (MAD>0). Note that the represented p-values are the geometric mean of the one-sided p-values, estimated by the paired Student's T test, for the differences between pre-mRNA - m/p-ratio and mRNA - m/p-ratio correlations. In addition, each tested target in this analysis was required to have sufficient exonic and intronic reads in all 293 RNA Atlas samples (see Methods), and >99% of them whose gene-level expressions fell into the 4th quantile while the dataset was partitioned according to the number of samples that expressed them. Namely, most tested targets were expressed in all 293 samples. Despite the imbalance of target counts across NES quantiles, the observations were still similar to those shown in Figures 5 and 6.

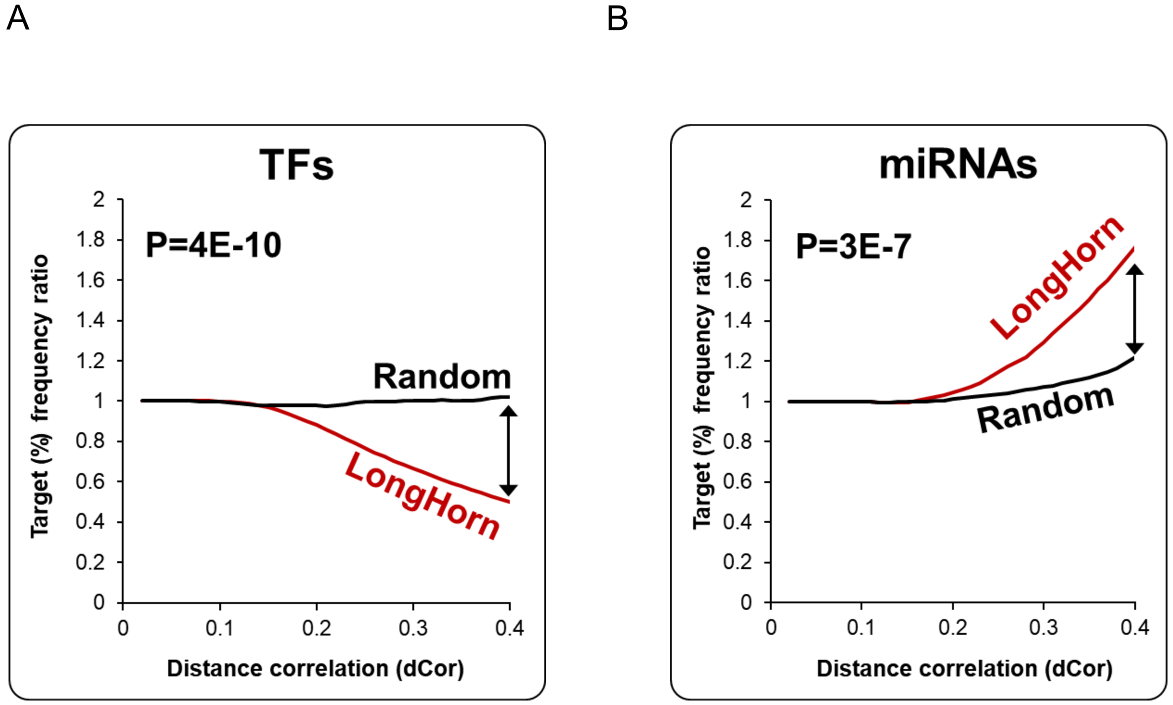


Figure S25. LongHorn-predicted regulator-target interactions showed significantly greater correlation deviations than expected. For each dCor cutoff *x* in the x-axis, the y-axis gives the ratio of target frequencies—the percentage of predicted regulator-target interactions with a dCor value greater than *x*—using target m/p-ratio relative to pre-mRNA expression profiles. Compared to random, LongHorn-predicted TF-target (**A**) and miRNA-target (**B**) interactions were increasingly more likely to have lower and higher dCor with regulator using target m/p-ratio expression profiles, respectively, as a function of dCor. We generated random regulator-target interactions by selecting predicted interactions with no consideration for lncRNA regulation and while maintaining regulator expression; see “The RNA Atlas lncRNA-target set” in Methods for details. Two-sided p-values were calculated by the paired Student’s T test.

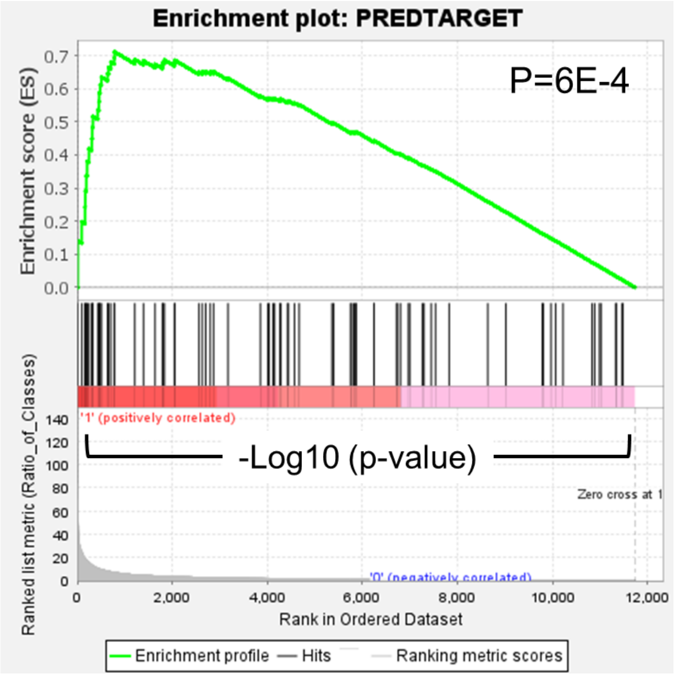


Figure S26. GSEA enrichment analysis of LongHorn-predicted EMX2OS targets for dysregulation following FANTOM6 ASO-mediated silencing of EMX2OS in human primary dermal fibroblast (HDF) cells. All expressed genes (TPM>0) profiled by RNA-Seq were sorted in descending order by their log10-transformed dysregulation p values according to FANTOM6. GSEA used weighted enrichment statistics and ratio of classes, with p-values computed using 10k gene-set permutations.

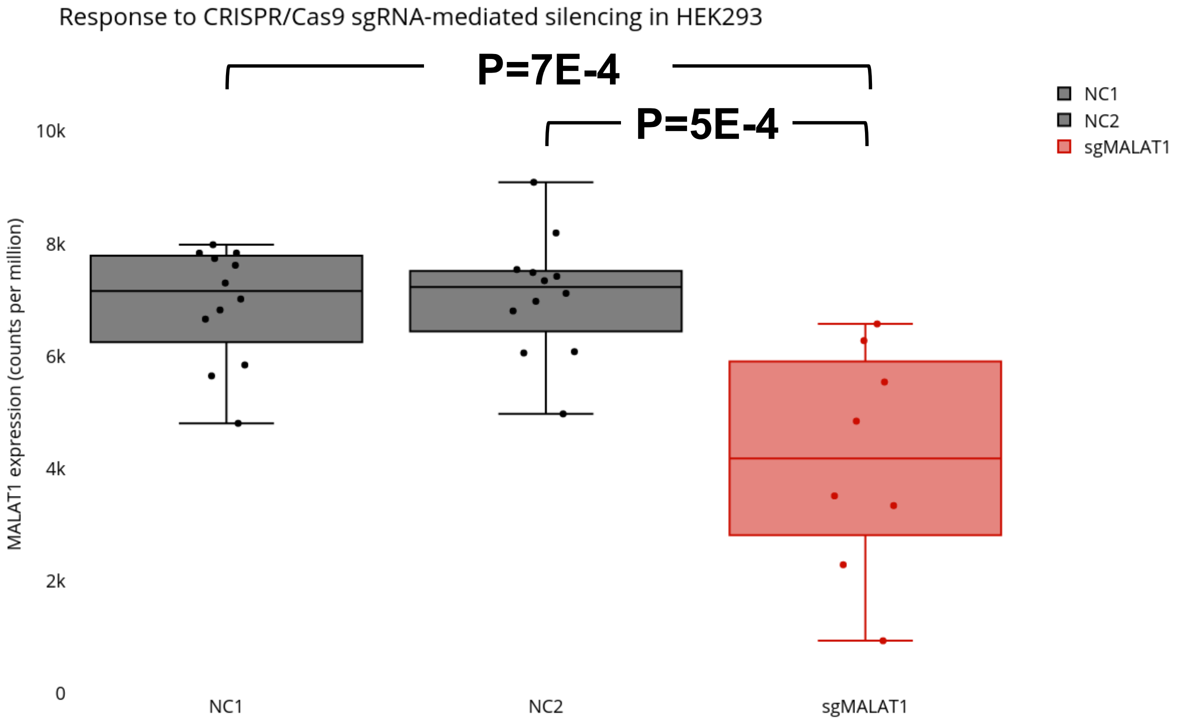


Figure S27. CRISPR/Cas9 sgRNA-mediated silencing of MALAT1 in HEK293 cells. sgMALAT1 transfections significantly downregulated MALAT1 by >40% (red), compared to the non-targeting controls NC1 and NC2 (black). MALAT1 expressions are measured in counts per million. Boxes extend from the 25th to the 75th percentiles, with the center lines at the median values; whiskers extend to the largest and smallest values no further than 1.5 times the inter-quartile range from the upper and lower box hinges, respectively. NC1, NC2, and sgMALAT1 have 12, 12, and 8 biologically independent samples, respectively. Each point in the boxplot represents the MALAT1 expression measurement in an individual sample. P-values are calculated by the two-sided Student’s T test. ***: P<1E-3.

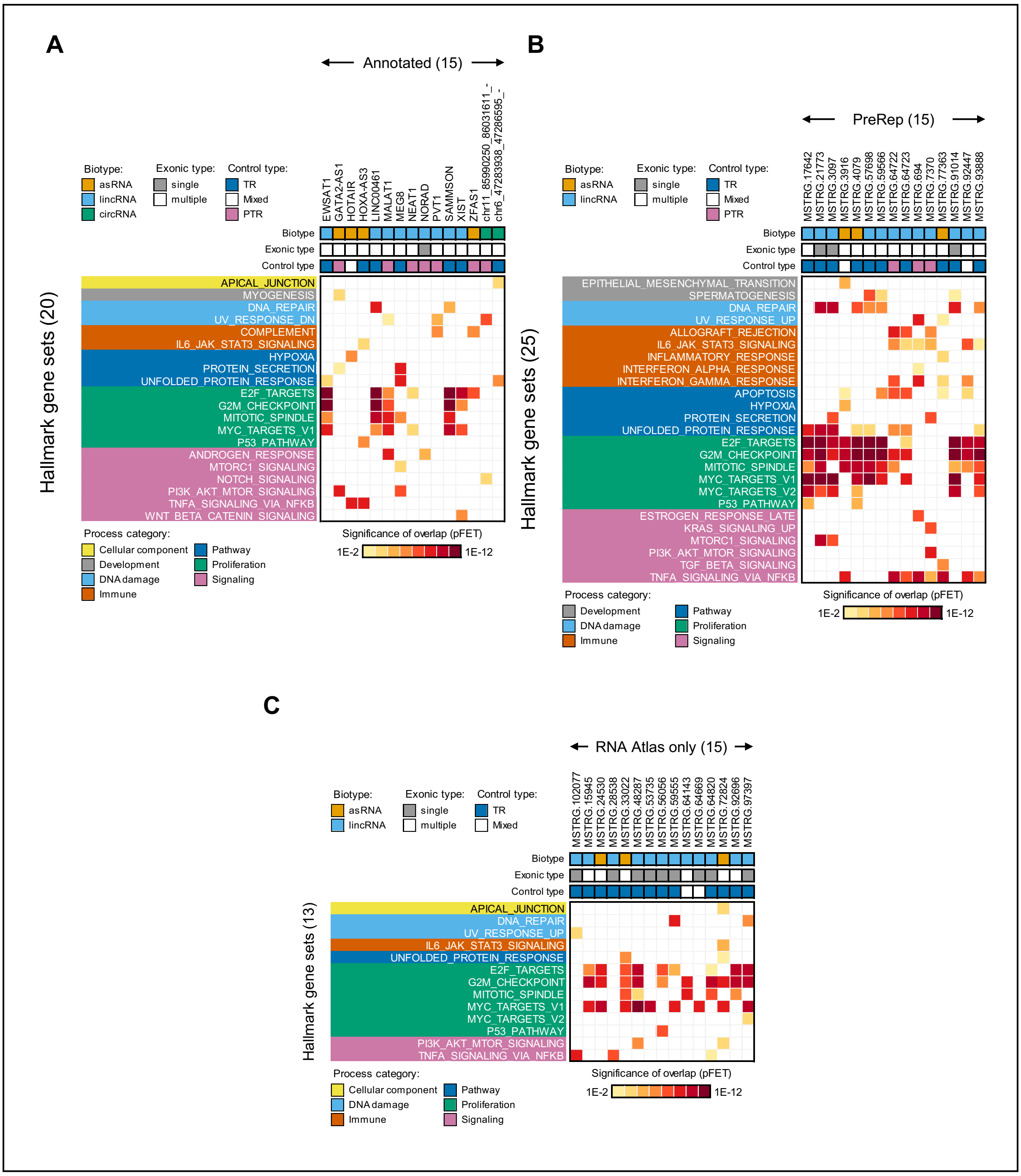


Figure S28. **Enrichment of predicted functional lncRNA regulators in Hallmark gene sets.** Enrichments for the predicted targets of fifteen representative (**A**) Annotated, (**B**) PreRep, and (**C**) RNA Atlas only lncRNAs in MsigDB[20] hallmark pathways. lncRNAs were cataloged based on biotype, whether they are single- and multi-exon, and regulatory modality.

B. Supplementary Tables description

All supplemental tables are available in either csv or tsv format in the online version of the manuscript.

Table S1. RNA Atlas samples.
samplenames: unique sample identifier (equivalent to “samplenames” field in R2 platform)
name: sample descriptive name
type: type of sample (tissue, cell type, cell line or cancer cell line)
biological_source: tissue or cell of origin or cancer type
organ_system: organ system of origin
UBERON ontology: id of the associated UBERON term (integrated cross-species anatomy ontology; https://www.ebi.ac.uk/ols/ontologies/uberon/terms)
available_RNA-seq_data: what RNA-seq data is available for the specific sample. For most samples, total RNA, polyA and small RNA-seq data are available, but in some cases, data from one or more methods may not be available because of fail in sequencing or library preparation.

Table S2. Summary information for polyA and total RNA-seq per-sample assemblies.
samplenames: unique sample identifier (equivalent to “samplenames” field in R2 platform)
seq_type: RNA-seq method (polyA or total RNA)
n_transcripts: number of total assembled transcripts
n_genes: number of total assembled genes
n_exons: number of total assembled exons
median_tr_length: median transcript length
mean_tr_length: mean transcript length
median_exon_length: median exon length
mean_exon_length: mean exon length
median_exons_per_tr: median exons per transcript
mean_exons_per_tr: mean exons per transcript
single_exon_transcripts: number of single-exon transcripts
single_exon_genes: number of single-exon genes

Table S3. Annotation of the filtered primary RNA Atlas transcriptome, including PCG, lincRNA and asRNA genes resulting from the merging of polyA and total RNA individual assemblies. This table includes both genes in the RNA Atlas transcriptome and candidate genes that were not included in RNA Atlas.
gene_name: unique gene identifier. Derived either from an Ensembl v86 gene name, or a from the name assigned by StringTie-merge in case of newly assembled genes (“MSTRG” naming).
chr: chromosome in which the gene is located
start: gene start coordinate, leftmost exon start position
end: gene end coordinate, rightmost exon end position
strand: genomic strand orientation

biotype: gene biotype (lincRNA, asRNA, PCG)
primary_assembly: an indication of whether the gene was known in the reference used to guide the assembly, Ensembl v86 (Ensembl.86 label), or newly assembled from RNA-seq data (newly_assembled label)
exonic_type: either single-exon (all transcripts from the gene have one exon only) or multi-exon (the gene has at least one transcript with more than one exon)
n_exons: total number of exons across all gene's transcripts
n_transcripts: number of different transcripts or isoforms
evidence_cat: evidence category based on the proximity of the gene (within 500 bp) to either a CAGE peak (RNA), one of various chromatin states peaks (DNA), or both of these (both), or evidence based on supra-median expression levels (expr_level), or no supporting evidence found from any of the analyzed features (none), see Methods for details.
final_set: the RNA Atlas transcriptome includes genes with evidence categories RNA, DNA, both or expr_level, genes from evidence category none are regarded as excluded candidate.
polyadenylation: polyadenylation status classification for RNA Atlas genes based on majority vote across samples (detailed in Methods). Levels: polyadenylated, non-polyadenylated, bimorphic or undetermined.
class: classification of RNA Atlas genes in Annotated, PreRep, or RNA Atlas only based on comparisons against other up to date gen sets annotations (see Methods).
in_FANTOM5_robust: weather the gene overlaps genes in the FANTOM5 robust set

Table S4. Annotation of candidate circRNAs identified across all total RNA libraries.
circRNA_id: circRNA unique identifier based on hg38 genomic coordinates
chr: chromosome in which the circRNA is located
start: position of the rightmost splice site involved in the back-splicing reaction (splice acceptor if "+" strand, splice donor if "-" strand)
end: position of the leftmost splice site involved in the back-splicing reaction (splice donor if "+" strand, splice acceptor if "-" strand)
strand: genomic strand orientation
SA_gene: names of gene(s) overlapped by the splice acceptor site, "." if there is no overlap with any gene
SD_gene: names of gene(s) overlapped by the splice donor site, "." if there is no overlap with any gene
consensus_gene: consensus gene(s) overlapped by both splicing sites. If multiple overlapping genes are annotated at the splice acceptor and/or donor site, the consensus (intersect) between both sites is taken.
consensus_gene_biotype: gene biotype of the consensus gene
consensus_gene_class: gene class (Annotated, PreRep or RNA Atlas only) of the consensus gene
circexplorer_total_junction: total number of reads overlapping the back-splice junction found across all samples by CIRCexplorer2
circexplorer_total_SA: total number of reads overlapping the splice acceptor site found across all samples by CIRCexplorer2
circexplorer_total_SD: total number of reads overlapping the splice donor site found across all samples by CIRCexplorer2

circexplorer_n_samples: total number of samples in which reads overlapping the backs-plice junction are found by CIRCexplorer2
findcirc_total_junction: total number of reads overlapping the back-splice junction found across all samples by find_circ2
findcirc_total_SA: total number of reads overlapping the splice acceptor site found across all samples by find_circ2
findcirc_total_SD: total number of reads overlapping the splice donor site found across all samples by find_circ2
findcirc_n_samples: total number of samples in which reads overlapping the backs-plice junction are found by find_circ2
n_samples: total number of samples in which reads overlapping the back-splice junction are found by either tool
mean_counts: mean read count overlapping the back-splice junction across samples. For each sample, the average counts between both tools was used.
final_set: the circRNA overlaps either genes in the RNA Atlas transcriptome or gene candidates that were excluded from RNA Atlas
in_other_DBs: weather the circRNA is annotated in any of 13 public circRNA resources analyzed

Table S5. Annotation of miRBase and miRDeep2 predicted mature candidate miRNAs identified across all small RNA libraries with 10 or more counts.
chr: chromosome in which the miRNA is located
start: miRNA start coordinate
end: miRNA end coordinate
strand: genomic strand orientation
miRNA_id: unique miRNA identifier. Note that this id is unique for the miRNA sequence, the same id is assigned to unique miRNA sequences that are found in multiple locations in the genome.
miRNA_extended_id: extended unique miRNA identifier. Note that this id is unique for both the miRNA sequence and the location in the genome.
class: RNA Atlas miRNA candidates are either annotated in miRBase or miRDeep2 predicted
arm: the precursor arm from which the mature miRNA is processed
precursor_id..pre.miRNA_extended_id.: unique identifier for the pre-miRNA
sequence: genomic sequence of miRNA
in_other_DBs: wether the mature miRNA overlaps miRNAs in other miRNA databases (including FANTOM5 miRNA set, miRCarta and MiRGeneDB)
Columns A to F: sequential restrictive conditions that were applied as criteria to select the final set of miRNAs with stringent functional evidence:

A: expressed (MAD>0)
B: at least one LongHorn-inferred target
C: at least one LongHorn-inferred target with adequate pre-mRNA and mRNA expressions
D: multiple LongHorn-inferred targets with adequate pre-mRNA and mRNA expressions
E: at least one LongHorn-inferred target with higher correlation between miRNA and target m/p-ratio

F: LongHorn-inferred targets have significantly higher correlations between miRNA and target m/p-ratio

Table S6. Annotation of miRBase and miRDeep2 predicted candidate miRNA precursors identified across all small RNA libraries.
pre.miRNA_extended_id: unique precursor miRNA identifier (including unique genomic coordinates)
pre.miRNA_id: unique precursor miRNA sequence identifier (some are duplicated in multiple genomic locations)
chr: chromosome in which the miRNA is located
start: pre-miRNA start coordinate
end: pre-miRNA end coordinate
strand: pre-miRNA genomic strand orientation
class: RNA Atlas pre-miRNA candidates are either annotated in miRBase or miRDeep2 predicted
sequence: genomic sequence of miRNA
N_expressed_arms: whether expression of mature miRNAs from one or both precursor arms was found across RNA Atlas samples
genomic_overlap: whether the pre-miRNA overlaps other RNA Atlas genes in exonic or intronic regions or is intergenic

Table S7. qPCR validation experiment with 110 single-exon genes.
gene_name: unique gene identifier
transcript_id: transcript used for primer design
forward_primer_sequence: DNA sequence of the design forward primer used in the qPCR reaction
reverse_primer_sequence: DNA sequence of the design reverse primer used in the qPCR reaction
biotype: gene biotype (lincRNA, asRNA, PCG)
class: classification of RNA Atlas genes in Annotated, PreRep, or RNA Atlas only based on comparisons against other up to date gen sets annotations (see Methods).
expr_profile: expression profile of the gene, whether it is expressed across many samples (ubiquitous) or specifically expressed in one of the cell lines used (IMR-32 or SK-N-BE(2)-C)
SK-N-BE(2)-C_cDNA.1: Cq values obtained with the qPCR reaction performed over the reversed-transcribed SK-N-BE(2)-C RNA samples (RT-qPCR), replicate 1
SK-N-BE(2)-C_cDNA.2: Cq values obtained with the qPCR reaction performed over the reversed-transcribed SK-N-BE(2)-C RNA samples (RT-qPCR), replicate 2
SK-N-BE(2)-C_RNA.1: Cq values obtained with the qPCR reaction performed over the SK-N-BE(2)-C RNA samples (qPCR), replicate 1
SK-N-BE(2)-C_RNA.2: Cq values obtained with the qPCR reaction performed over the SK-N-BE(2)-C RNA samples (qPCR), replicate 2
SK-N-BE(2)-C_water.1: negative control for SK-N-BE(2)-C reactions, Cq values obtained with the qPCR reaction performed over a control mastermix cotaining the forward and reverse primers, but no RNA nor cDNA, replicate 1
SK-N-BE(2)-C_water.2: negative control for SK-N-BE(2)-C reactions, Cq values obtained with the qPCR reaction performed over a control mastermix cotaining the forward and reverse primers, but no RNA nor cDNA, replicate 2

IMR-32_cDNA.1: Cq values obtained with the qPCR reaction performed over the reversed-transcribed IMR-32 RNA samples (RT-qPCR), replicate 1
IMR-32_cDNA.2: Cq values obtained with the qPCR reaction performed over the reversed-transcribed IMR-32 RNA samples (RT-qPCR), replicate 2
IMR-32_RNA.1: Cq values obtained with the qPCR reaction performed over the IMR-32 RNA samples (qPCR), replicate 1
IMR-32_RNA.2: Cq values obtained with the qPCR reaction performed over the IMR-32 RNA samples (qPCR), replicate 2
IMR-32_water.1: negative control for IMR-32 reactions, Cq values obtained with the qPCR reaction performed over a control mastermix cotaining the forward and reverse primers, but no RNA nor cDNA, replicate 1
IMR-32_water.2: negative control for IMR-32 reactions, Cq values obtained with the qPCR reaction performed over a control mastermix cotaining the forward and reverse primers, but no RNA nor cDNA, replicate 2

Table S8. Summary of overlap of RNA Atlas only single-exon genes with ONT reads across four public datasets
transcript_id: unique transcript identifier
gene_name: unique gene identifier
transcript_length: transcript length (bp)
transcript_chr_id: transcript chromosome id
transcript_strand: transcript strand
transcript_start: transcript start coordinate (one-based)
transcript_stop: transcript end coordinate
Columns Workman.et.al., Gleeson.et.al., Leger.et.al., Cole.et.al.: whether any overlapping reads were found in the corresponding dataset[17,18,23,24]

Tables S9-S12. Detailed information on ONT reads overlapping single-exon RNA Atlas only genes in each analyzed public dataset, Workman et al.[18] (Table S23), Gleeson.et.al.[23] (Table S24), Leger.et.al.[24] (Table S25) and Cole.et.al[17] (Table S26).
transcript_id: unique transcript identifier
transcript_length: transcript length (bp)
transcript_chr_id: transcript chromosome id
transcript_strand: transcript strand
transcript_start: transcript start coordinate (one-based)
transcript_stop: transcript end coordinate
read_id: ONT read id
read_length: ONT read length (bp)
read_start: ONT read start coordinate (one-based)
read_end: ONT read end coordinate
coverage: percentage of transcript length covered by the overlapping ONT read

Table S13. Candidate new protein-coding genes.
gene_name: unique gene identifier
chr: chromosome in which the gene is located
start: gene start coordinate, leftmost exon start position
end: gene end coordinate, rightmost exon end position
strand: genomic strand orientation

evidence_cat: evidence category based on the proximity of the gene (within 500 nucleotides) to either a CAGE peak (RNA), one of various chromatin marks (DNA), or both of these (both), or evidence based on supra-median expression levels (expr_level), or no supporting evidence found from any of the analyzed features (none), see Methods for details.
polyadenylation: polyadenylation status classification for RNA Atlas genes based on majority vote across samples (detailed in Methods). Levels: polyadenylated, non-polyadenylated, bimorphic or undetermined.
class: classification of the RNA Atlas genes in Annotated, PreRep, or RNA Atlas only based on comparisons against other up to date gen sets annotations (see Methods).
match_biotype: biotype annotation of the gene in the matching gene set (protein-coding or non-coding, or no-match if the gene is RNA Atlas only)
nsamps_0.1_TotalRNA: number of samples in which the gene is expressed in 0.1 TPM or higher in total RNA-seq
nsamps_0.5_TotalRNA: number of samples in which the gene is expressed in 0.5 TPM or higher in total RNA-seq
sortedsamps_0.1_TotalRNA: list of samples in which the gene is expressed in 0.1 TPM or higher in total RNA-seq (ordered by decreasing expression value)
nsamps_0.1_polyA: number of samples in which the gene is expressed in 0.1 TPM or higher in polyA-seq
nsamps_0_5_polyA: number of samples in which the gene is expressed in 0.5 TPM or higher in polyA-seq
sortedsamps_0.1_polyA: list of samples in which the gene is expressed in 0.1 TPM or higher in polyA-seq (ordered by decreasing expression value)
max_N_exons: maximum number of exons found across individual isoforms
max_ORF_length: ORF length of the largest isoform
fraction_aa_cons_chimp: fraction of the amino acidic sequence that is conserved in chimpanzee
fraction_nt_cons_chimp: fraction of the DNA sequence that is conserved in chimpanzee
mass_spec_evidence: whether peptides matching the candidate protein were found in mass spectrometry data from the Human Proteome Map with an estimated FDR below 1% (see Methods).
blastp_evalue: BLASTp expected value (E-value) for the best hit.

Table S14. Table of known polyadenylated and non-polyadenylated genes used as reference for the polyadenylation status classification. This list was generated based on Yang et al. (2011)[21].
gene_name: unique gene identifier
polyadenylation: polyadenylation status based on Yang et al. (2011)

Table S15. List and annotation of 160 genes with variable polyadenylation status across samples.
gene_name: unique gene identifier
biotype: gene biotype (lincRNA, asRNA, PCG)

n_polyA_samples: number of samples that have at least 100 counts in both total and polyA RNA-seq data and a positive polyadenylation score (log2 normalized ratio)
n_non_polyA_samples: number of samples that have at least 100 counts in total RNA-seq data and a negative polyadenylation score (log2 normalized ratio)
plus_sample: most extreme polyadenylated sample (i.e. with highest log2 normalized ratio) for this gene
minus_sample: most extreme non-polyadenylated sample (i.e. with lowest log2 normalized ratio) for this gene
normalized_logratio_plus_sample: normalized log2 ratio score for the most extreme polyadenylated sample
normalized_logratio_minus_sample: normalized log2 ratio score for the most extreme non-polyadenylated sample
polyA_gene_tpm_ratio: ratio of TPM gene level expression from polyA RNA-seq between plus and minus sample
TotalRNA_gene_tpm_ratio: ratio of TPM gene level expression from total RNA-seq between plus and minus sample
dominant_plus.polyA: dominant transcript (with highest expression among all gene's isoforms) for the most extreme polyadenylated sample in polyA RNA-seq data
dominant_minus.polyA: : dominant transcript (with highest expression among all gene's isoforms) for the most extreme non-polyadenylated sample in polyA RNA-seq data
dominant_plus.Total: dominant transcript (with highest expression among all gene's isoforms) for the most extreme polyadenylated sample in total RNA-seq data
dominant_minus.Total: dominant transcript (with highest expression among all gene's isoforms) for the most extreme non-polyadenylated sample in total RNA-seq data
fraction_plus.polyA: fraction of total gene expression represented by the dominant transcript for the most extreme polyadenylated sample in polyA RNA-seq data
fraction_minus.polyA: fraction of total gene expression represented by the dominant transcript for the most extreme non-polyadenylated sample in polyA RNA-seq data
fraction_plus.Total: fraction of total gene expression represented by the dominant transcript for the most extreme polyadenylated sample in total RNA-seq data
fraction_minus.Total: fraction of total gene expression represented by the dominant transcript for the most extreme non-polyadenylated sample in total RNA-seq data
tpm_plus.polyA: TPM expression of dominant transcript for the most extreme polyadenylated sample in polyA RNA-seq data
tpm_minus.polyA: TPM expression of dominant transcript for the most extreme non-polyadenylated sample in polyA RNA-seq data
tpm_plus.Total: TPM expression of dominant transcript for the most extreme polyadenylated sample in total RNA-seq data
tpm_minus.Total: TPM expression of dominant transcript for the most extreme non-polyadenylated sample in total RNA-seq data

Table S16. Significantly imprinted genes in 203 tissue and cell type samples.

seq_type: indicates if the gene was detected with total RNA-seq, polyA RNA-seq or both (joint). These genes are filtered on a difference between observed & expected heterozygotes > 30
chr: chromosome in which the gene is located
gene_name: unique gene identifier
polyadenylation: polyadenylation status classification based on majority vote across samples (detailed in Methods).
#SNPs total RNA: number of significantly imprinted SNPs per gene in total RNA-seq
#SNPs polyA: number of significantly imprinted SNPs per gene in polyA RNA-seq
obs/exp Hz total RNA: observed vs expected heterozygous samples per gene (most conservative estimate in case of multiple SNPs) in total RNA-seq
obs/exp Hz polyA: observed vs expected heterozygous samples per gene (most conservative estimate in case of multiple SNPs) in polyA RNA-seq
known: a comparison is made with GTEx, geneimprint.com and literature. V: known imprinted gene compared to GTEx, geneimprint.com and literature. $: previously detected as imprinted in blood by our methodology, unpublished results (https://bit.ly/2oCR6eD)

Table S17. Annotation of fusion genes labels retrieved by FusionCatcher[22] and criteria used for filtering false positives.
Label:label retrieved by FusionCatcher tool
Description: label description
Source: reference to the source, if any, where the fusion was previously found
Code: manual annotation indicating the probability of the fusion being a false positive, with 2 indicating low probability, 1 indicating high probability and 0 indicating very high probability. Only fusions with label 2 were retained.

Tables S18-20. Per sample mRNA, pre-mRNA, and m/p-ratio expression profiles of protein-coding transcripts (quantile-normalized and log2-transformed). Each column represents a sample profiled by RNA Atlas. Each row represents a RefSeq transcript and its corresponding gene symbol.
(Related to Figure 5)

Table S21. Experimentally-verified targets of canonical regulators including TFs and miRNAs. (Related to Figure 5)
regname: gene symbol of TFs/RBPs; miRBase ID of mature miRNAs
targname: gene symbol of targets
biotype(reg): biotype of regulators (miRNA or PCG)
biotype(targ): biotype of targets (PCG only)
class(reg): classification of regulators (Annotated for TFs/RBPs; miRBase for mature miRNAs)
class(targ): classification of targets (Annotated only)
MOA(TR:trans;PTR:posttrans): mode of action of regulators (either transcriptional or post-transcriptional)

Table S22. LongHorn-inferred targets of canonical regulators including TFs and miRNAs. (Related to Figure 5)

regname: gene symbol of TFs/RBPs; miRBase ID of mature miRNAs in miRBase; RNA Atlas ID (RNAATLASMIRXXXXX) of mature miRNAs predicted by miRDeep2
targname: gene symbol of targets
biotype(reg): biotype of regulators (miRNA or PCG)
biotype(targ): biotype of targets (PCG, asRNA, lincRNA, or circRNA)
class(reg): classification of regulators (Annotated for TFs/RBPs; miRBase and miRDeep2 predicted for mature miRNAs)
class(targ): classification of targets (Annotated for PCG; Annotated, PreRep, or RNA Atlas only for ncRNAs)
MOA(TR:trans;PTR:posttrans): mode of action of regulators (either transcriptional or post-transcriptional)

Table S23. High-confidence set of miRBase and miRDeep2 predicted miRNAs with strong evidence of post-transcriptional regulation in multiple tissues, and their LongHorn-inferred targets. (Related to Figure 5)
miRNA_id: miRBase IDs of mature miRNAs in miRBase; RNA Atlas ID (RNAATLASMIRXXXXX) of mature miRNAs predicted by miRDeep2.
class: classification of mature miRNAs (miRBase or miRDeep2 predicted)
numtarget: number of protein-coding target transcripts predicted by LongHorn
targetlist: list of protein-coding target transcripts predicted by LongHorn with their RefSeq IDs and corresponding gene symbols

Table S24. lncRNA network predicted by LongHorn using RNA Atlas expression profiles. (Related to Figure 6)
regname: gene symbol of regulators
targname: gene symbol of targets
biotype(reg): biotype of regulators (asRNA, lincRNA, or circRNA)
biotype(targ): biotype of targets (PCG only)
class(reg): classification of lncRNAs (Annotated, PreRep, or RNA Atlas only)
class(targ): classification of targets (Annotated only)
MOA(TR:trans;PTR:posttrans): mode of action of regulators (Co-factor, Guide, Decoy(TF), or Decoy(MiRNAIRBP))
mediatorcount: number of mediators modulated by the regulator
mediatorlist: list of mediators modulated by the regulator

Table S25. Significance of correlation deviations for LongHorn-inferred lncRNA targets. (Related to Figure 6)
RegName: gene symbol of regulators
TargName: gene symbol of targets
Biotype(reg): biotype of regulators (antisense, lincRNA, or circRNA)
Biotype(targ): biotype of targets (protein-coding only)
MOA(TR:trans;PTR:posttrans): mode of action of regulators (Co-factor, Guide, Decoy(TF), or Decoy(MiRNAIRBP))
Pval(PairedT-test;one-tailed): analytical p-values for the significances of correlation deviation using the paired Student's T test.
Pval(permutation;min=0.01): non-parametric p-values for the significances of correlation deviation estimated by permutation testing. The minimum attainable p-value is 0.01.

adjPval(PairedT-test;one-tailed;Benjamini-Hochberg): analytical p-values for the significances of correlation deviation using the paired Student's T test with the Benjamini-Hochberg adjustment for multiple comparisons
adjPval(permutation;Benjamini-Hochberg): non-parametric p-values for the significances of correlation deviation estimated by permutation testing with the Benjamini-Hochberg adjustment for multiple comparisons

Table S26. Large-scale verification of LongHorn-predicted lncRNA targets using FANTOM6 ASO-mediated lncRNA knockdown experiments followed by transcriptome profiling with RNA-Seq in human primary dermal fibroblast (HDF) cells. (Related to Figure S26).
LncRNA Name: lncRNA gene names
ASO ID: identifiers of antisense oligonucleotide (ASO) provided by the FANTOM6 consortium
# Gene (profiled): total number of genes profiled by RNA-Seq
# Gene (dysreg at P<0.05): number of profiled genes that were dysregulated at p<0.05 following targeting a specific lncRNA with ASO. The significance of gene dysregulation was estimated from DESeq2 by the FANTOM6 consortium.**# Gene (pred_target)**: number of LongHorn-predicted lncRNA targets
# Gene (pred_target & dysreg at P<0.05): number of LongHorn-predicted lncRNA targets that were also dysregulated at p<0.05 following targeting the same lncRNA with ASO. The significance of gene dysregulation was estimated from DESeq2 by the FANTOM6 consortium.
pFET: significance of overlap between LongHorn-predicted lncRNA targets and dysregulated genes using the Fisher's Exact Test. Significant one-sided p-values lower than 0.05 are in red.**Odds Ratio (OR):** the odds of being dysregulated gens in LongHorn-predicted targets relative to that in all genes profiled by RNA-Seq for a specific lncRNA. OR>1 and OR<1 indicate that LongHorn-predicted targets were more or less likely to be dysregulated, respectively. Odds ratios larger than 1 are in red.

Table S27. Transcriptome profiling by RNA-Seq upon the transfection of MALAT1-targeting sgRNAs and non-targeting controls (NC1 and NC2) based on CRISPR-Cas9 interference (CRISPRi) system in HEK293 cells. (Related to Figure S27) Each row and column correspond to the profiled protein-coding genes and the sgRNA identities, respectively. The gene abundances were measured in counts per million (CPM).

Table S28. Significant lncRNAs in Hallmark Gene Set enrichment analysis. (Related to Figure 7)
lncrna: gene symbol of lncRNAs
class: classification of lncRNAs (Annotated, PrRep, or RNA Atlas only)
biotype: biotype of lncRNAs (asRNA or lincRNA)
hallmark_gene_set: name of hallmark gene sets
overlap_size: number of overlapping genes between LongHorn-inferred lncRNA targets and hallmark gene sets
target_enrichment(pFET): significance of overlap estimated by Fisher's Exact Test

References

1. Frankish A, Diekhans M, Ferreira A-M, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res*. 2019;47:767. doi:10.1093/nar/gky955
2. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*. 2015. doi:10.1038/nbt.3122
3. Forrest ARR, Kawaji H, Rehli M, et al. A promoter-level mammalian expression atlas. *Nature*. 2014;507(7493):462-470. doi:10.1038/nature13182
4. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518(7539):317-329. doi:10.1038/nature14248
5. Liu SJ, Horlbeck MA, Cho SW, et al. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science (80-)*. 2017;355(6320):eaah7111. doi:10.1126/science.aah7111
6. Hon CC, Ramilowski JA, Harshbarger J, et al. An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature*. 2017;543(7644):199-204. doi:10.1038/nature21374
7. Iyer M, Niknafs Y, Malik R, et al. The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet*. 2015;47(3):199-208. doi:10.1038/ng.3192
8. You BH, Yoon SH, Nam JW. High-confidence coding and noncoding transcriptome maps. *Genome Res*. 2017;27(6):1050-1062. doi:10.1101/gr.214288.116
9. Pertea M, Shumate A, Pertea G, et al. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol*. 2018;19(1):208. doi:10.1186/s13059-018-1590-2
10. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016;44(D1):D733-D745. doi:10.1093/nar/gkv1189
11. Kozomara A, Griffiths-Jones S. MiRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*. 2014;42(D1):68-73. doi:10.1093/nar/gkt1181
12. Friedländer MR, MacKowiak SD, Li N, Chen W, Rajewsky N. MiRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res*. 2012;40(1):37-52. doi:10.1093/nar/gkr688
13. De Rie D, Abugessaisa I, Alam T, et al. An integrated expression atlas of miRNAs and their promoters in human and mouse. *Nat Biotechnol*. 2017;35(9):872-878. doi:10.1038/nbt.3947
14. Backes C, Fehlmann T, Kern F, et al. miRCarta: a central repository for collecting miRNA candidates. *Nucleic Acids Res*. 2018;46. doi:10.1093/nar/gkx851
15. Fromm B, Domanska D, Høye E, et al. MirGeneDB 2.0: The metazoan microRNA complement. *Nucleic Acids Res*. 2020;48(D1):D132-D141. doi:10.1093/nar/gkz885

16. Cunningham F, Amode MR, Barrell D, et al. Ensembl 2015. *Nucleic Acids Res*. 2015. doi:10.1093/nar/gku1010
17. Cole C, Byrne A, Adams M, Volden R, Vollmers C. Complete characterization of the human immune cell transcriptome using accurate full-length cDNA sequencing. *bioRxiv*. 2019:589-601. doi:10.1101/761437
18. Workman RE, Tang AD, Tang PS, et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat Methods*. 2019;16(12):1297-1305. doi:10.1038/s41592-019-0617-2
19. Mukherji S, Ebert MS, Zheng GXY, Tsang JS, Sharp PA, Van Oudenaarden A. MicroRNAs can generate thresholds in target gene expression. *Nat Genet*. 2011;43(9):854-859. doi:10.1038/ng.905
20. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst*. 2015. doi:10.1016/j.cels.2015.12.004
21. Yang L, Duff MO, Graveley BR, Carmichael GG, Chen LL. Genomewide characterization of non-polyadenylated RNAs. *Genome Biol*. 2011;12(2). doi:10.1186/gb-2011-12-2-r16
22. Nicorici D, Satalan M, Edgren H, et al. *FusionCatcher - a Tool for Finding Somatic Fusion Genes in Paired-End RNA-Sequencing Data*.; 2014. doi:10.1101/011650
23. Gleeson J, 2# TAL, Harrison PJ, Haerty W, Clark MB. Nanopore direct RNA sequencing detects differential expression between human cell populations. doi:10.1101/2020.08.02.232785
24. Leger A, Amaral PP, Pandolfini L, et al. RNA modifications detection by comparative Nanopore direct RNA sequencing. *bioRxiv*. 2019:1-29. doi:10.1101/843136