nature biotechnology

RESOURCE

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

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

Lucia Lorenzi[1,2,21], Hua-Sheng Chiu[3,21], Francisco Avila Cobos[1,2], Stephen Gross[4], Pieter-Jan Volders[1,2,5], Robrecht Cannoodt[1,2,6,7,8], Justine Nuytens[1,2], Katrien Vanderheyden[1,2], Jasper Anckaert[1,2], Steve Lefever[1,2], Aidan P. Tay[9,10], Eric J. de Bony[1,2], Wim Trypsteen[1,2], Fien Gysens[1,2], Marieke Vromman[1,2], Tine Goovaerts[11], Thomas Birkballe Hansen[12], Scott Kuersten[4], Nele Nijs[13], Tom Taghon[14], Karim Vermaelen[15], Ken R. Bracke[15], Yvan Saeys[5,6], Tim De Meyer[2,11], Nandan P. Deshpande[16], Govardhan Anande[17,18], Ting-Wen Chen[19], Marc R. Wilkins[16], Ashwin Unnikrishnan[17,18], Katleen De Preter[1,2], Jørgen Kjems[12], Jan Koster[20], Gary P. Schroth[4], Jo Vandesompele[1,2], Pavel Sumazin[3] ✉ and Pieter Mestdagh[1,2] ✉

Existing compendia of non-coding RNA (ncRNA) are incomplete, in part because they are derived almost exclusively from small and polyadenylated RNAs. Here we present a more comprehensive atlas of the human transcriptome, which includes small and polyA RNA as well as total RNA from 300 human tissues and cell lines. We report thousands of previously uncharacterized RNAs, increasing the number of documented ncRNAs by approximately 8%. To infer functional regulation by known and newly characterized ncRNAs, we exploited pre-mRNA abundance estimates from total RNA sequencing, revealing 316 microRNAs and 3,310 long non-coding RNAs with multiple lines of evidence for roles in regulating protein-coding genes and pathways. Our study both refines and expands the current catalog of human ncRNAs and their regulatory interactions. All data, analyses and results are available for download and interrogation in the R2 web portal, serving as a basis for future exploration of RNA biology and function.

RNA sequencing technologies have enabled the interrogation of the human transcriptome at nucleotide resolution, exposing distinct RNA biotypes beyond protein-coding genes (PCGs). A plethora of regulatory ncRNAs—including microRNAs (miRNAs), long intergenic non-coding RNAs (lincRNAs), antisense RNAs (asRNAs) and circular RNAs (circRNAs)—have been identified and are being explored as potential players in human development and disease, including cancer[1–3]. Several consortium-based efforts have contributed to the discovery and quantification of these RNA biotypes in heterogeneous sample collections[4–13]. The resulting transcriptome landscape is a community resource to study RNA biology and to characterize regulatory mechanisms and gene functions, as well as to identify predictive biomarkers for human diseases[14–17]. However, these studies have relied mostly on analyses of small and polyadenylated RNA transcriptomes. Consequently, we still lack a systematic survey of non-polyadenylated and circularized transcripts and their relationship to other RNA biotypes. To capture a more complete diversity of the human transcriptome, we profiled a heterogeneous collection of 300 human samples—including 45 tissues, 162 cell types and 93 cell lines (Fig. 1a and Supplementary Table 1)—using three complementary RNA sequencing technologies. From these samples, we generated strand-specific small RNA (298 samples), polyA (295 samples) and total RNA (296 samples) libraries that were sequenced at median depths of 13 million, 60 million and 125 million paired-end reads, respectively, for a total of 125 billion reads (Supplementary Fig. 1). The resulting datasets include profiles of PCGs, lincRNAs, asRNAs, circRNAs and miRNAs, and their analysis produced a carefully constructed transcriptome and a matching expression atlas. Moreover, the broad intron coverage from total RNA sequencing enabled data-driven predictions of

[1]Center for Medical Genetics, Ghent University, Ghent, Belgium. [2]Cancer Research Institute Ghent (CRIG), Ghent, Belgium. [3]Texas Children's Cancer Center, Baylor College of Medicine, Houston, TX, USA. [4]Illumina, Inc., San Diego, CA, USA. [5]VIB-UGent Center for Medical Biotechnology, VIB, Ghent, Belgium. [6]Data Mining and Modelling for Biomedicine Group, VIB Center for Inflammation Research, Ghent, Belgium. [7]Department of Applied Mathematics, Computer Science, and Statistics, Ghent University, Ghent, Belgium. [8]Data Intuitive, Lebbeke, Belgium. [9]Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation, New South Wales, Sydney NSW, Australia. [10]Department of Biomedical Sciences, Macquarie University, New South Wales, Sydney NSW, Australia. [11]Department of Data Analysis and Mathematical Modelling, Ghent University, Ghent, Belgium. [12]Interdisciplinary Nanoscience Centre (iNANO), Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark. [13]Biogazelle, Zwijnaarde, Belgium. [14]Department of Diagnostic Sciences, Ghent University, Ghent, Belgium. [15]Department of Respiratory Medicine, Ghent University, Ghent, Belgium. [16]Systems Biology Initiative, School of Biotechnology and Biomolecular Sciences, UNSW Sydney, Sydney NSW, Australia. [17]Adult Cancer Program, Lowy Cancer Research Centre, UNSW Sydney, Sydney NSW, Australia. [18]Prince of Wales Clinical School, UNSW Sydney, Sydney NSW, Australia. [19]Institute of Bioinformatics and Systems Biology, National Yang Ming Chiao Tung University, Hsinchu, Taiwan. [20]Department of Oncogenomics, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands. [21]These authors contributed equally: Lucia Lorenzi, Hua-Sheng Chiu. ✉e-mail: sumazin@bcm.edu; pieter.mestdagh@ugent.be

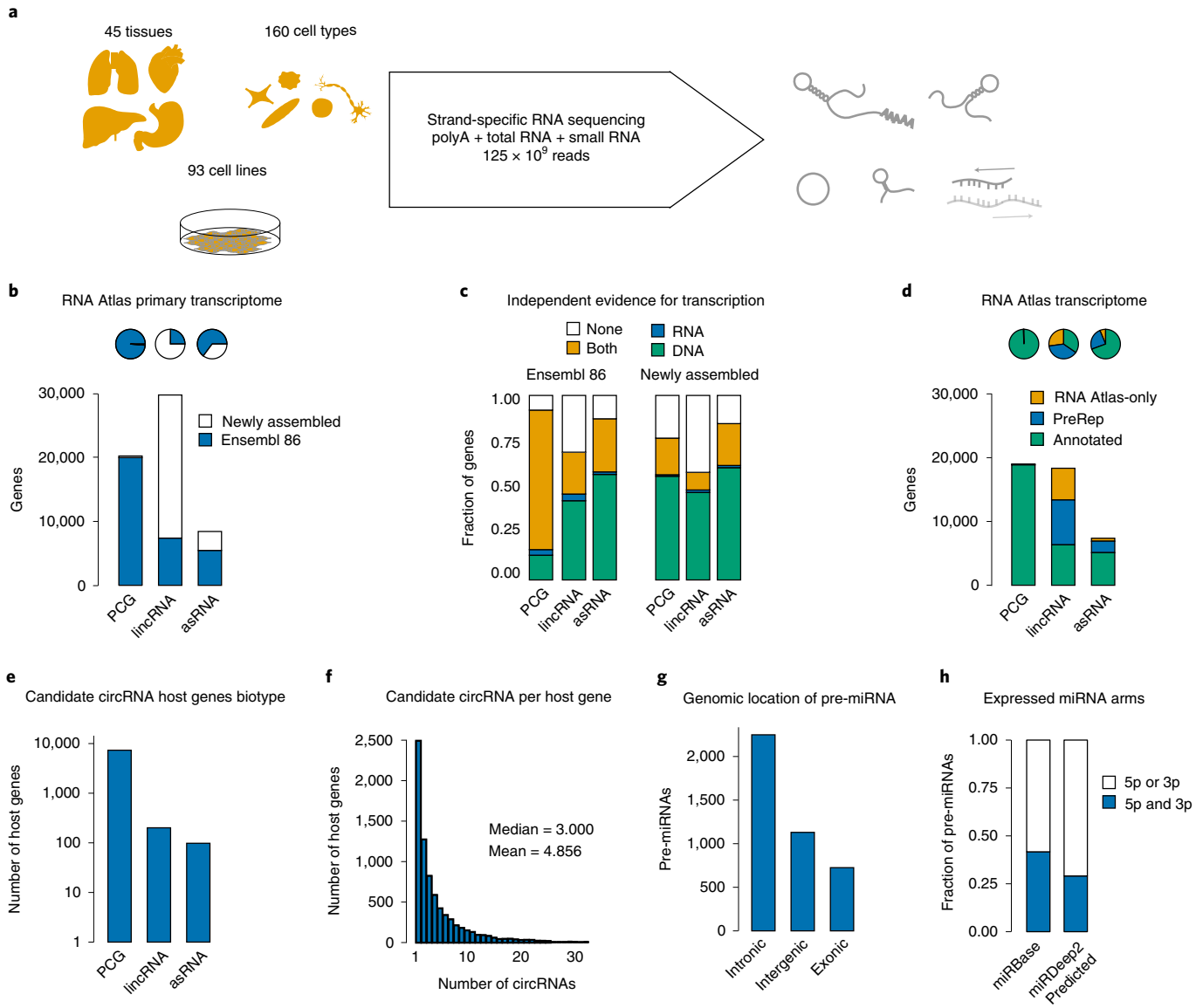


Fig. 1 | RNA Atlas transcriptome generation and annotation. a, We profiled RNA in a heterogeneous collection of tissues, cell types and cell lines using complementary strand-specific technologies. **b**, Number and fraction of known genes in Ensembl version 86 and newly assembled PCG, lincRNA and asRNA genes in the complete RNA Atlas transcriptome. **c**, Fraction of known and newly assembled genes with independent evidence for transcriptional activity according to chromatin state (DNA), CAGE peak (RNA), both or neither. **d**, The number and fraction of Annotated, PreRep and RNA Atlas-only PCG, lincRNA and asRNA genes in the RNA Atlas transcriptome. **e**, The frequencies of PCG, lincRNA and asRNA host types for RNA Atlas circRNAs. **f**, The distribution of circRNA counts per host. **g**, Annotation of RNA Atlas miRNA precursors relative to their nearest genes. **h**, The fractions of miRBase- and miRDeep2-predicted pre-miRNA candidates with expressed mature miRNAs from both pre-miRNA arms or one pre-miRNA arm.

transcriptional and post-transcriptional modulation of gene expression by ncRNAs. Our results are available for download and analysis through the R2 web portal (http://r2platform.com/rna_atlas).

Results

Assembling a comprehensive human transcriptome reveals many single-exon lincRNAs. The assembly of the RNA Atlas transcriptome was guided by Ensembl annotation (v86)[18] (Fig. 1b, Supplementary Fig. 2 and Supplementary Tables 2 and 3; see Methods for details) and included known and newly assembled PCGs, lincRNAs and asRNAs. To gather independent evidence supporting the transcriptional activity of these genes, we integrated public cap analysis of gene expression (CAGE) sequencing data generated by the FANTOM consoritum[6] and various chromatin state profiles that are associated with transcription or enhancer activity from comprehensive cell and tissue collections (Epigenomics Roadmap[19]). Most Ensembl and newly assembled genes—88% and 62%, respectively—were closely associated with either CAGE peaks or relevant chromatin states within 500 base pairs of transcription start sites (TSSs) (Fig. 1c and Supplementary Fig. 3). ncRNAs were primarily supported by chromatin marks, including enhancer marks that were previously shown to have a significant predictive value for lncRNA function[20]. PCGs, on the other hand, were more strongly associated with CAGE and active transcription chromatin states. Genes supported by CAGE peaks or chromatin states were retained in what we refer to as the RNA Atlas transcriptome, which is composed of RNA Atlas genes (Fig. 1d). An additional 642 genes without CAGE peak or chromatin state support, but with supramedian expression levels, were also included (Methods).

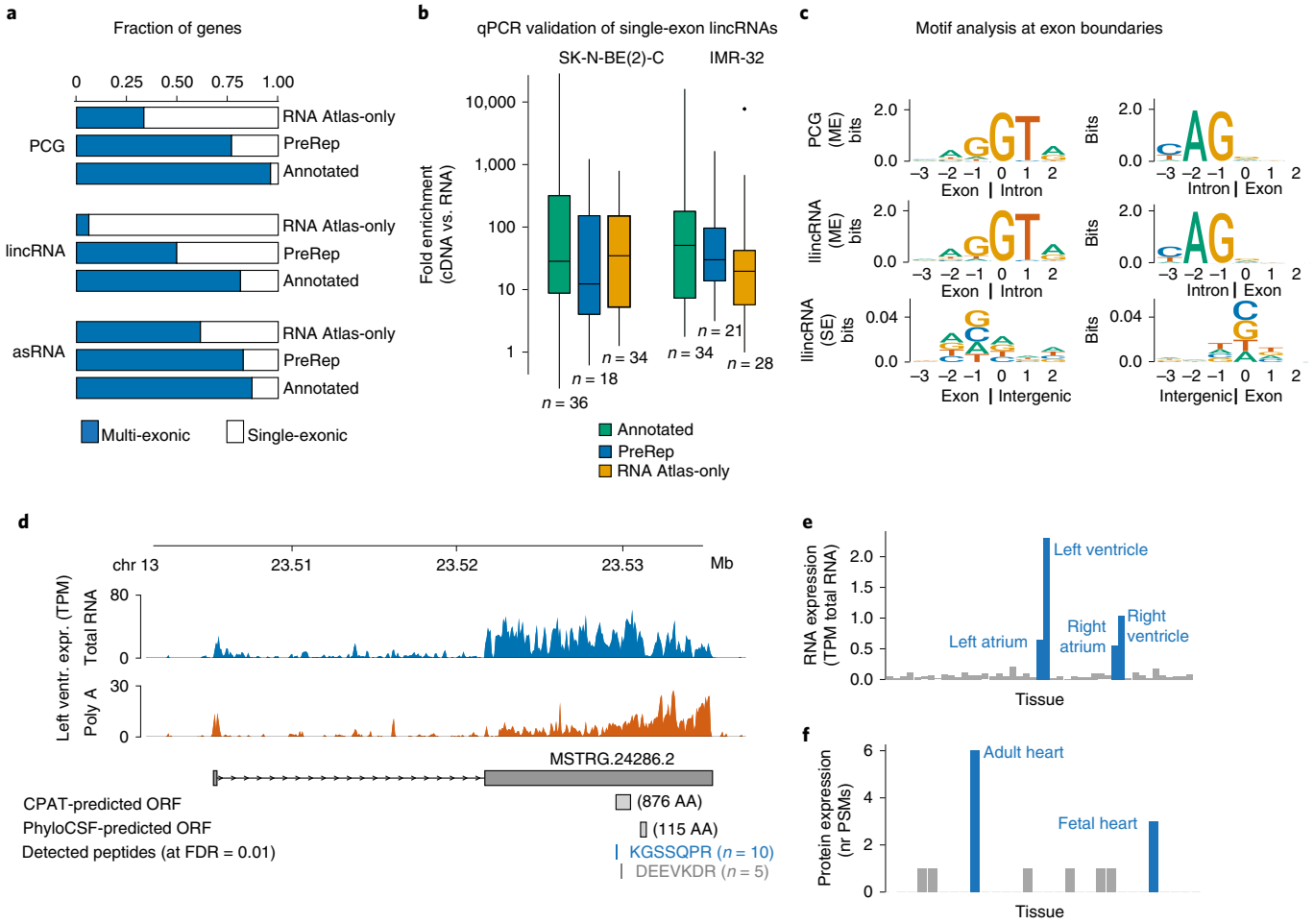


Fig. 2 | The RNA Atlas transcriptome catalogued many single-exon lncRNAs and revealed previously non-annotated PCGs. a, The fraction of single-exon and multi-exon genes in the PCG, lincRNA and asRNA biotypes. **b**, qPCR validation of RNA Atlas single-exon genes in two cell lines: SK-N-BE(2)-C (left) and IMR-32 (right). Box plots display the $\log_{10}$ fold increase of qPCR signal in the RT (cDNA) versus no-RT (RNA) samples for genes that were amplified (Cq < 35, total amplified genes = 105) in each cell line. 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 interquartile range from the upper and lower box hinges, respectively. 'Outlier' genes beyond the limits of whiskers are plotted as individual points. **c**, Sequence motif analysis of exon/intron boundaries or exon/intergenic boundaries for single-exon genes; the *y* axis shows the information content measured in bits. **d**, An example of a heart-specific PCG with matching peptides from public mass spectrometry data. Expression of this gene is specific for heart tissue samples (blue bars) at both RNA (**e**) and peptide (**f**) levels. PSM, peptide spectrum match.

This transcriptome consisted of 18,962 PCGs, 18,364 lincRNAs and 7,374 asRNAs. After transcriptome assembly, we annotated the RNA Atlas transcriptome against GENCODE version 33 (ref. [21]), the curated set of RefSeq[22] version 200 (the combination of which is further referred to as Annotated RNAs) and genes reported by four large-scale RNA annotation efforts, including the FANTOM5 (ref. [7]) stringent set, CHESS[9], MiTranscriptome[10] and BIGTranscriptome[13], and with the model RefSeq version 200 set (the combination of which is further referred to as Previously Reported ('PreRep' for short) RNAs); see details in Methods and Supplementary Fig. 4c. Based on these comparisons, we classified our set of RNA Atlas genes as Annotated, PreRep or RNA Atlas-only (see Fig. 1d, and note that these gene sets are mutually exclusive). Annotated genes included most PCGs (99%) and asRNAs (69%). However, 65% of RNA Atlas lincRNAs were RNA Atlas-only (27%) or PreRep (38%). Overall, RNA Atlas genes included 5,471 RNA Atlas-only lncRNAs, 11,453 Annotated lncRNAs and 8,814 PreRep lncRNAs (Supplementary Fig. 4a). We note that 990 of the RNA Atlas-only genes were also identified in the less stringent collection of FANTOM5 robust genes. Pairwise comparisons of RNA Atlas lncRNAs and lncRNAs found in other catalogs consistently showed higher association of RNA Atlas lncRNAs to DNase and enhancer chromatin states compared to lncRNAs not present in RNA Atlas (Supplementary Fig. 4d). The analysis of ENCODE transcription factor (TF) occupancy in lncRNA promoters suggested that the promoters of RNA Atlas lncRNAs were more likely to be occupied by TFs than previously predicted or annotated lncRNAs that are not in RNA Atlas. In total, 69%, 71% and 77% of RNA Atlas-only, Annotated and PreRep lncRNA proximal promoters, respectively, were identified with at least two ENCODE TF binding sites, compared to 46% and 49% of previously predicted and annotated lncRNA promoters, respectively, that were not included in RNA Atlas.

The analysis of total RNA sequencing assays identified 38,023 candidate circRNAs with more than four back-splice reads (Supplementary Table 4). Of these, 37,128 were derived from RNA Atlas genes and were retained in the RNA Atlas transcriptome. We identified 13 pre-existing circRNA databases[23], and most (99%) of the RNA Atlas circRNAs were annotated in at least one circRNA

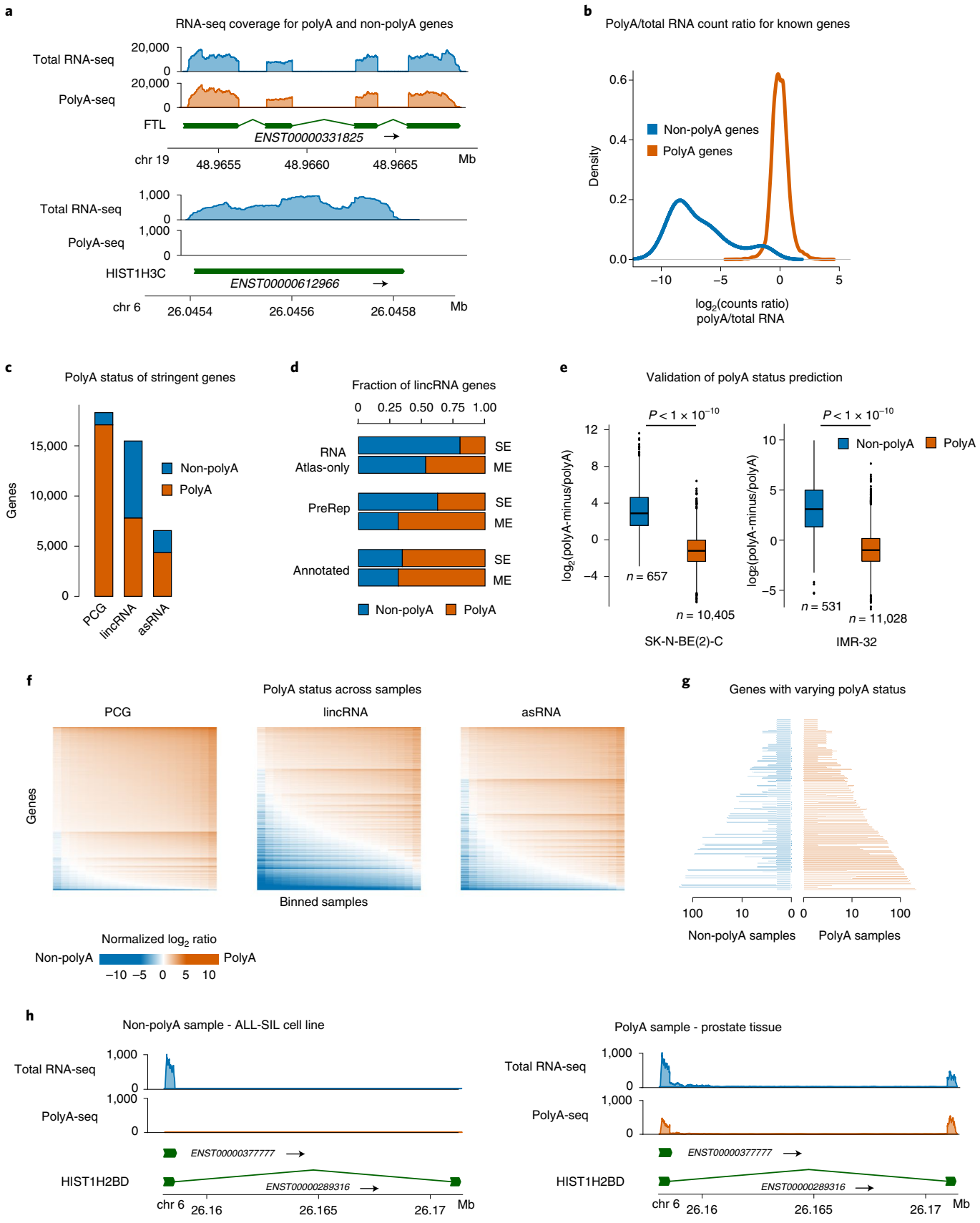
database (only 446 circRNAs were exclusively found in RNA Atlas), with 70% present in at least five of the 13 databases (Methods and Supplementary Fig. 5c,d). Almost all circRNAs (98%) were processed from PCG hosts, with a median of three circRNAs per host (Fig. 1e,f). Most circRNAs spanned at least four exons, and the circRNAs were flanked by introns that were significantly longer than the introns not flanking circRNAs ($P < 1 \times 10^{-10}$ by Wilcoxon rank-sum test; Supplementary Fig. 5a,b). These circRNA characteristics are in line with previous reports[24,25]. Finally, we identified 5,213 candidate mature miRNAs that passed a minimal abundance criterion of ten reads in our small RNA sequencing assays. These included 1,646 miRBase[26] (v22) miRNAs and 3,567 (68%) miRDeep2-predicted[27] miRNAs, of which 2,600 (73%) showed no overlap to genomic coordinates of miRNAs included in any of three other miRNA resources (FANTOM5 (ref. [8]), miRCarta[28] and MirGeneDB[29]; Supplementary Table 5 and Supplementary Fig. 4b). Predicted precursor transcripts for these candidate miRNAs were predominantly located in intronic and intergenic regions, and the fraction of precursors giving rise to both a 3p and a 5p mature form was lower for miRDeep2-predicted miRNAs (29%) than for miRBase (42%) miRNAs (Supplementary Table 6 and Fig. 1g,h). Precursor miRNAs are processed from larger primary miRNA transcripts whose TSSs are not easily characterized, with 1,130 precursors located in intergenic regions. Consequently, we did not attempt to integrate CAGE sequencing or chromatin state data to filter miRNA candidates. Instead, we used ncRNA target inference algorithms to identify miRNAs with evidence for post-transcriptional target regulation, selecting 316 miRNAs and their targets—details to follow.

Most PreRep and RNA Atlas-only lncRNAs were lincRNAs, and most of these (68%) were single-exon genes (Fig. 2a). Notably, 4,877 of the 5,471 (89%) RNA Atlas-only lncRNAs were single-exon gene models. Single-exon lincRNAs are often removed from transcriptome assemblies to guard against contaminating DNA. However, our stranded RNA sequencing workflow allowed for testing DNA fragments based on strand identity, with random DNA fragments expected to map to both strands in a nearly equal ratio. In contrast, the mean exonic strandedness for the single-exon lincRNAs was 96%, which was similar to that of multi-exon lincRNAs (94%) and PCGs (95%) (Supplementary Fig. 6). This test suggested that the single-exon lincRNAs do not originate from contaminating DNA fragments. Our conclusion was further validated experimentally by qPCR in RNA samples that were not reverse transcribed into cDNA for 110 single-exon genes, including 42 RNA Atlas-only, 27 PreRep and 41 Annotated genes. For those single-exon genes that were successfully amplified ($Cq < 35$, $n = 105$), we observed a significant fold increase of qPCR signal in the RT versus no-RT samples by Wilcoxon signed-rank test ($P < 1 \times 10^{-10}$; Fig. 2b, Supplementary Fig. 7 and Supplementary Table 7). Moreover, all single-exon RNA Atlas lincRNAs had CAGE peaks (13%) or chromatin states that indicated active transcription (98%) near their predicted TSSs. Unlike multi-exon PCGs or lincRNAs, and similarly to single-exon PCGs, single-exon lincRNAs were not flanked by canonical splice sites (Fig. 2c and Supplementary Fig. 8a). Moreover, the distances between single-exon lincRNAs and their nearest genes were significantly larger (median, 2.3 kb) than distances between consecutive exons in multi-exon PCGs (median, 0.511 kb; $P < 1 \times 10^{-10}$) or lincRNAs (median, 1.17 kb; $P < 1 \times 10^{-10}$); P values were computed by Wilcoxon rank-sum tests (Supplementary Fig. 8b). Together, these observations suggest that RNA Atlas single-exon lincRNAs are not fragments of multi-exon genes that lacked junction reads in our assembly. The fact that single-exon and multi-exon RNA Atlas-only genes show almost identical expression distributions (Supplementary Fig. 9) further strengthens that notion. Finally, an analysis of publicly available single-molecule RNA sequencing datasets (Oxford Nanopore Technologies (ONT)) identified non-spliced and uniquely mapping ONT reads overlapping 372 (7.6%) of the RNA Atlas-only single-exon genes (Supplementary Tables 8–12 and Supplementary Fig. 10). Of note, we could not find any spliced ONT reads overlapping RNA Atlas-only single-exon genes across the four analyzed datasets, further supporting the single-exon status of these genes.

Although most RNA Atlas-only genes were ncRNAs, our workflow also revealed a handful of new candidate PCGs. Based on in silico predictions and cross-species protein conservation, we identified 104 candidate PCGs that were not previously annotated as such but were likely protein coding based on evaluations by CPAT[30], PhyloCSF[31] and BLASTp[32]; note that PCG identification required significant positive assessment by all three methods. The BLASTp E value of all 104 selected candidates was below 1×10^{-3}, and 90 of these had an E value less than 1×10^{-5} (see Methods and Supplementary Table 13 for details). These 104 genes included nine RNA Atlas-only genes, 80 genes that matched non-coding genes in either the Annotated or PreRep sets and 15 PreRep PCGs. The coding potential of these genes was further substantiated through a re-analysis of mass spectrometry data from the Human Proteome Map[33], revealing peptides matching 20 (19%) of these candidate PCGs (false discovery rate (FDR) < 0.01). In addition, we found that, whereas the median percentage of conserved amino acids in chimpanzee was lower for RNA Atlas candidates (96%) than Annotated PCGs (99%), 84% of our predicted open reading frames (ORFs) were still more conserved than 10% of Annotated PCGs, indicating that most candidate PCGs fall within an acceptable range of conservation observed in Annotated PCGs (Supplementary Fig. 11). In addition, our 104 candidate PCGs were more tissue specific and had fewer exons and shorter ORF lengths than known PCGs (Supplementary Fig. 11). Finally, we identified peptides whose expression profiles across tissues correlated with that of their

Fig. 3 | Analyses of RNA polyadenylation status. a, Read coverage profiles from polyA sequencing and total RNA sequencing libraries for a known polyadenylated gene (upper panel) and a known non-polyadenylated gene (lower panel). **b**, Distributions of the normalized $\log_2$ ratio of counts from polyA sequencing versus total RNA sequencing in an individual sample (human umbilical vein endothelial cell) for known polyadenylated and non-polyadenylated genes. **c**, The number of polyadenylated and non-polyadenylated genes for each RNA biotype based on majority votes across samples; RNA Atlas genes with ten or more counts in at least one sample and with an uneven majority vote across samples are shown. **d**, The relative frequencies of polyadenylated and non-polyadenylated lincRNAs. **e**, Validation of non-polyadenylated genes through polyA-minus sequencing in two RNA Atlas cell lines; box plots show the $\log_2$ counts ratio between polyA-minus and polyA sequencing for genes classified as non-polyadenylated (blue) and polyadenylated (orange). 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 interquartile range from the upper and lower box hinges, respectively. 'Outlier' genes beyond the limits of whiskers are plotted as individual points. P values were estimated by two-sided Wilcoxon rank-sum tests. **f**, Heat maps showing polyadenylation status of genes across samples. We plotted the mean normalized $\log_2$ ratio of the expression of binned genes with 100 or more counts in at least 20 samples; genes were sorted into 20 non-empty bins based on normalized expression. **g**, Selection of genes with varying polyadenylation across samples. The number of samples in each category is shown with values in $\log_{10}$ scale. Genes are sorted by evenness (that is, difference in number of samples in each category) and by number of polyadenylated samples. **h**, An example of a gene (HIST1H2BD) whose varying polyadenylation across samples can be explained by differential expression of alternatively polyadenylated isoforms. Coverage profiles from total RNA sequencing and polyA sequencing are shown for a non-polyadenylated sample (ALL-SIL cell line, left panel) and a polyadenylated sample (prostate tissue, right panel).

template mRNAs, underscoring the validity of the results. These included a heart-specific peptide whose template mRNA was specifically expressed in heart tissues (Fig. 2d–f) and a peptide with high abundance in T cells and B cells whose template mRNA showed the highest expression in spleen and lymph node (Supplementary Fig. 12). Protein domain analysis of the corresponding heart-specific



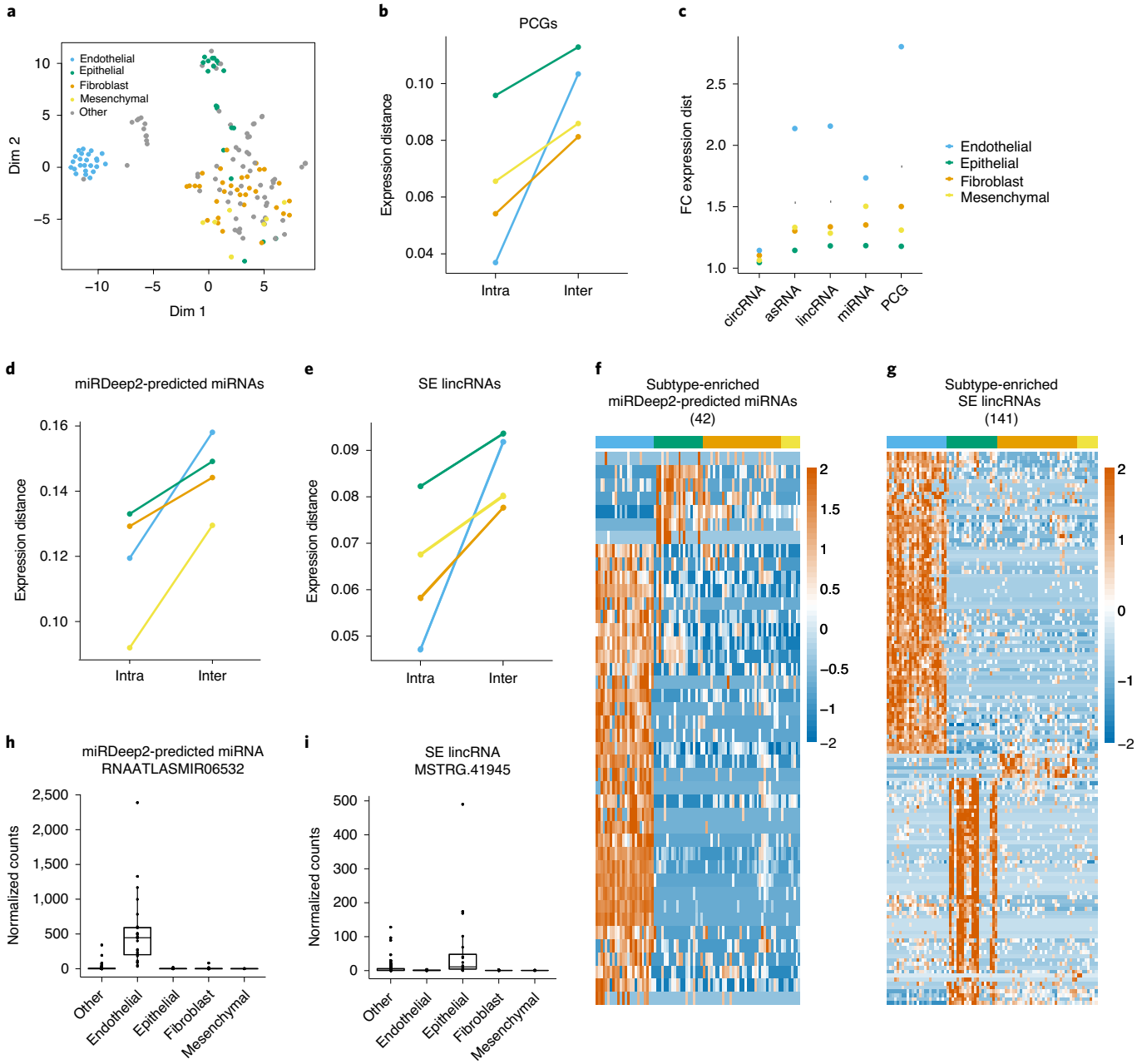


Fig. 4 | The association between sample ontology and expression distance. a, t-SNE plot of the RNA Atlas cell types based on PCG expression. Samples were colored according to four biological cell subtypes of interest (epithelial, mesenchymal, fibroblast and endothelial). **b**, For each biological cell subtype, we showed the median PCG expression-based distances between pairs of samples from the same subtype (intra-distances) and across subtypes (inter-distances). **c**, The distribution of fold changes between median inter- and intra-distances of RNA biotypes in each of the four cell subtypes. **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 heat map for miRDeep2-predicted candidate miRNAs significantly upregulated in cell subtypes. **g**, Expression heat map for single-exon lincRNAs significantly upregulated in cell subtypes. **h**, **i**, Examples of an endothelial-specific miRDeep2-predicted candidate miRNA (**h**) and an epithelial-specific single-exon lincRNA (**i**). Expression distributions are shown as box plots of normalized counts as computed by the DESeq function for each sample group: endothelial ($n=24$), epithelial ($n=20$), fibroblast ($n=32$), mesenchymal ($n=8$) and other ($n=209$). 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 interquartile range from the upper and lower box hinges, respectively. 'Outlier' samples beyond the limits of whiskers are plotted as individual points.

sequence of this peptide revealed a TLV_coat domain, which is also present in human syncytin genes. Homology to syncytin was further validated through BLAST analysis. Syncytin proteins mediate cell fusion during placental development, a process also occurring in muscle and cardiac tissues[34].

Integrating polyA and total RNA sequencing data reveals thousands of non-polyadenylated genes. We used our matching polyA and total RNA sequencing data to test the polyadenylation status of RNA Atlas genes. We reasoned that normalized expression of polyadenylated genes should be nearly equal in polyA and total RNA

libraries from each sample, whereas, for non-polyadenylated genes, normalized expression profiles should be markedly lower in polyA libraries (Fig. 3a). Indeed, the distribution of normalized $\log_2$ count ratios (polyA/total RNA) for known polyadenylated genes was centered around zero and was significantly higher than that of known non-polyadenylated genes (Fig. 3b, Supplementary Fig. 13 and Supplementary Table 14). For each of 291 samples for which matching polyA and total RNA data were available, we calculated the $\log_2$ ratio cutoff that most accurately classified known polyadenylated ($n=5{,}987$) and non-polyadenylated ($n=117$) genes[35] and subsequently applied it to establish the polyadenylation status of RNA Atlas genes (see Methods and Supplementary Table 3 for details). Only genes with at least ten counts were included in the analysis. As expected, most PCGs (93%) were classified as polyadenylated based on majority votes across all samples (Fig. 3c). For lincRNAs and asRNAs, the fraction of polyadenylated genes was 48% and 63%, respectively. Notably, more than 75% of the RNA Atlas-only and 60% of the PreRep single-exon lincRNAs were classified as non-polyadenylated (Fig. 3d), demonstrating the added value of total RNA sequencing to detect this specific RNA biotype. Polyadenylated and non-polyadenylated lncRNAs showed similar cross-species conservation and promoter TF occupancy (Supplementary Fig. 14). To empirically validate our polyA status assessment methodology, we established a polyA-minus RNA sequencing protocol by depleting polyadenylated transcripts from total RNA libraries and applied it to two RNA Atlas cell lines. Non-polyadenylated genes showed significantly higher polyA-minus/polyA count ratios than polyadenylated genes ($P<1\times10^{-10}$ by Wilcoxon rank-sum test; Fig. 3e).

To study potential changes in gene polyadenylation status across samples, we limited our analysis to genes with at least 100 counts in 20 or more samples. $\log_2$ count ratio distributions across samples suggested shifts from polyadenylated to non-polyadenylated states for a subset of genes in each biotype category (Fig. 3f). A fraction of these genes, including 41% of the lincRNAs, 43% of the asRNAs and 30% of the PCGs, were classified as both polyadenylated and non-polyadenylated—each in at least five samples per classification. To evaluate the cause of these shifts, we selected the most extreme cases based on $\log_2$ count ratios in each class, identifying 160 genes, including 83 PCGs, 36 lincRNAs and 41 asRNAs (Fig. 3g and Supplementary Table 15). We found that variable polyadenylation status might be driven by differential expression of alternatively polyadenylated isoforms in 57 (36%) of these genes. One example is the histone gene HIST1H2BD that is transcribed into a single-exon non-polyadenylated PCG in the ALL-SIL cell line and a two-exon polyadenylated PCG in prostate tissues (Fig. 3h). The remaining 103 genes did not show changes in isoform usage, suggesting that other mechanisms might alter polyadenylation (Supplementary Fig. 15).

RNA biotype expression reflects sample ontology. We verified that the RNA Atlas expression data reflect several well-established features of the transcriptome, such as non-coding RNA expression specificity[36], imprinting[37] and cancer fusion gene expression[38]. As expected, we observed a strong enrichment of mRNA fusion genes in cancer cell lines compared to non-malignant cell types and tissues, and we detected 20 known imprinted genes that featured consistent mono-allelic expression over the large majority of tissues and cell types (Supplementary Figs. 16 and 17 and Supplementary Tables 16 and 17). We confirmed that ncRNAs were expressed in a more tissue-specific manner than PCGs even after normalization for RNA biotype abundance (Supplementary Fig. 18). For circRNAs, tissue specificity resembled that of PCGs (Supplementary Fig. 18). In total, 96% of 1,320 previously catalogued tissue-specific RNA Atlas RNAs[39] were confirmed as tissue specific in our dataset (Supplementary Fig. 19). In conclusion, the analysis of RNA Atlas data confirmed these previously proposed transcriptome characteristics.

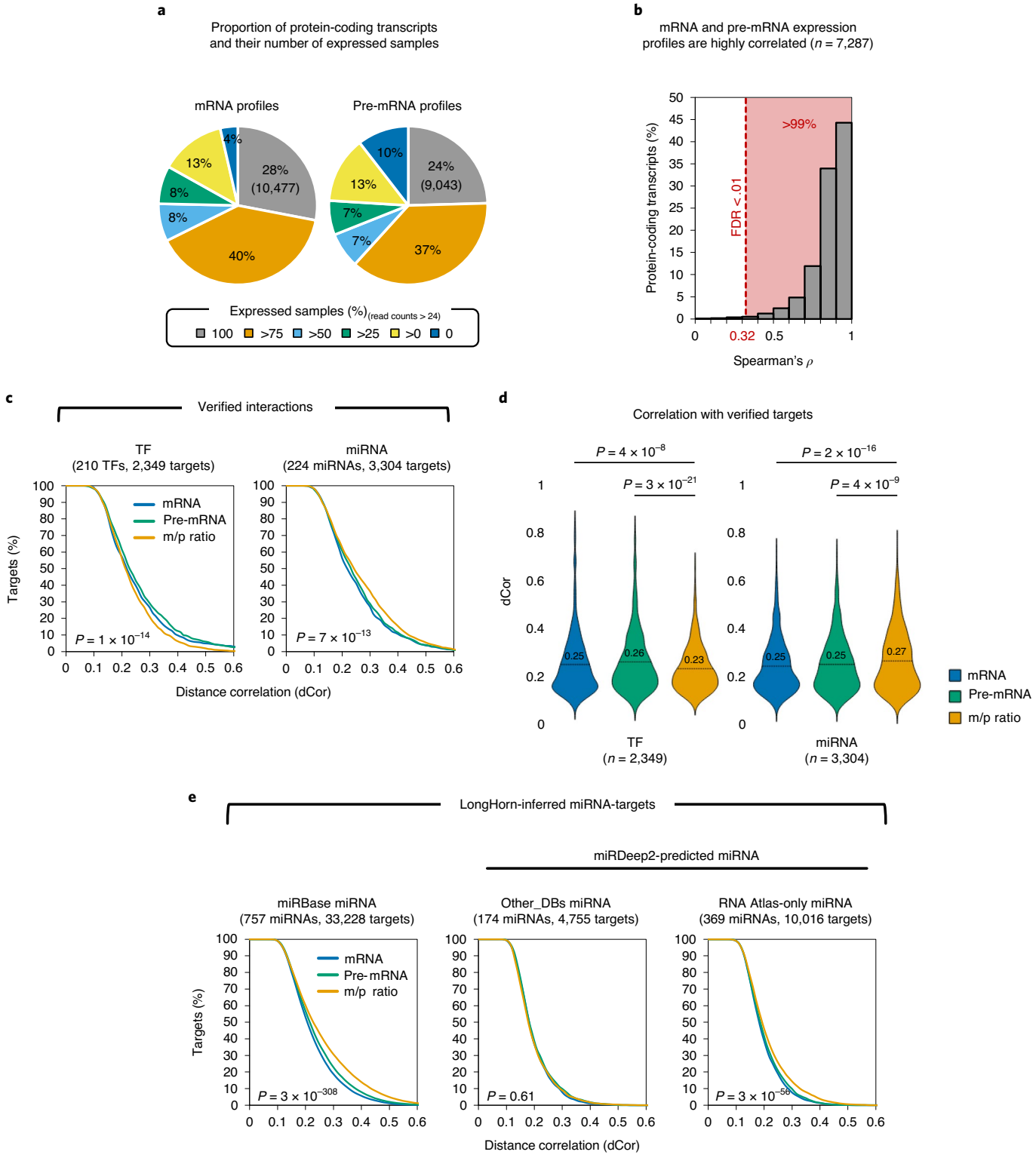
We used two-dimensional unsupervised clustering to evaluate the relationship among gene expression profiles, RNA biotypes and cellular subtypes. Clustering based on PCG expression profiles suggested that closely associated cell types had similar transcriptomes (Fig. 4a). Most strikingly, epithelial cells ($n=20$), endothelial cells ($n=24$), fibroblasts ($n=32$) and mesenchymal cells ($n=8$) clustered together and were distinct from transcriptional profiles of the other cell types in the RNA Atlas dataset. Namely, PCG expression-based distances between samples within a subtype (intra-distance) were consistently smaller than expression-based distances between these samples and all other samples in the dataset (inter-distance) ($P<1\times10^{-5}$, Wilcoxon rank-sum test; Fig. 4b). This was observed for all four subtypes and was most pronounced for endothelial cells for which the median inter-distance was 2.8-fold higher than the median intra-distance.

When calculating intra- and inter-distances for RNA biotypes, all biotypes—including miRDeep2-predicted miRNAs and single-exon lincRNAs—showed inter-distances that were significantly higher than intra-distances (Fig. 4c–e), suggesting that these biotypes are expressed in a coordinated manner. This is supported by the fact that a subset of mirDeep2-predicted miRNAs ($n=42$) and single-exon lincRNAs ($n=141$) were significantly upregulated ($\log_2$ fold change >3 and a Benjamini–Hochberg FDR <0.01) in individual cell subtypes and demonstrated subtype-specific expression profiles (Fig. 4f–i). To further validate these findings, we repeated our analysis, focusing on cancer cell lines within each cancer type. Inter-cancer-type distances were significantly higher than intra-cancer-type distances for all biotypes (Supplementary Fig. 20a–c). Significant differences were also observed for the miRDeep2-predicted miRNAs and single-exon lincRNAs (Supplementary Fig. 20d,e), confirming our observations about biological subtypes. Moreover, we identified several miRDeep2-predicted miRNAs ($n=119$) and single-exon lincRNAs ($n=567$) that were significantly overexpressed in individual cancer types (Supplementary Fig. 20f–i). These results indicated that RNA Atlas genes, including miRDeep2-predicted candidate

Fig. 5 | Total RNA transcriptomes facilitated the use of intron expression profiles to study regulatory modalities. a, Distribution of protein-coding transcripts whose exonic (mRNA, left) and intronic (pre-mRNA, right) expression estimates were supported by at least 24 unique reads in 0%, >0%, >25%, >50%, >75% or 100% of RNA Atlas-profiled samples. **b**, Distribution of Spearman's correlations between pre-mRNA and mRNA expression profiles; $n=7{,}287$ expression values were $\log_2$ transformed. **c**, Cumulative distributions of dCor between TF or miRNA expression profiles and their verified target's pre-mRNA and mRNA expression profiles and the ratios between these profiles (m/p ratio). *P* values were computed as the geometric means of the one-sided *P* values, estimated by the paired Student's *t*-test, for delta correlations between regulators and their target's expression profiles, where delta correlations compared correlations of regulators and their target's pre-mRNA or mRNA versus their m/p ratio profiles. **d**, On average, correlations between the profiles of regulators and their verified target's mRNA and pre-mRNA showed no significant difference; however, correlations between regulator and target m/p ratio profiles were significantly lower and higher for TFs and miRNAs, respectively. The averaged dCor is given as a dashed line on the plot along with a number next to it. One-sided *P* values were estimated by the paired Student's *t*-test. **e**, LongHorn-inferred direct targets of miRBase and RNA Atlas-only miRNAs had improved regulator-target m/p ratio correlations, but this difference was not significant for Other_DB miRNAs, which were identified in previous studies. Both RNA Atlas-only and Other_DB miRNAs were predicted by miRDeep2 using RNA Atlas-profiled samples; similar to **c**, *P* values were computed as the geometric means of the one-sided *P* values estimated by the paired Student's *t*-test. The number of tested regulators and interactions is in parentheses.

miRNAs and single-exon lincRNAs, showed expression patterns that closely reflect sample ontology relationships. Furthermore, these non-random expression patterns support our assertion that RNA Atlas single-exon lincRNAs do not derive from DNA fragments contaminating the RNA sequencing libraries.

We note that, in our analysis, circRNA expression profiles were the least correlated with cell types (Fig. 4c and Supplementary Fig. 20c). Because of technical limitations—circRNAs can only be quantified using reads spanning the back-splice junction—the number of available reads that quantify circRNA expression was typically lower and less quantifiable. When limiting the analysis to more abundant circRNAs, we observed an increase in the difference between inter- and intra-cell-type cluster distances(Supplementary Fig. 21). That increase was proportional to the abundance of selected circRNAs, and only the 1% most abundant circRNAs produced results that were similar to those observed for other RNAs.



Overall, 44 and 77 circRNAs were significantly upregulated with $\log_2$ fold changes >3 and a Benjamini–Hochberg-based FDR < 0.01 in at least one of the four major cell types or ten cancer types (Supplementary Fig. 21).

Total RNA transcriptomes facilitate the use of intron expression profiles to study regulatory modalities. Both intronic and exonic expression profiles could be accurately estimated for thousands of RNA Atlas transcripts in each of our total RNA profiles (Fig. 5a). Consequently, we were able to use these estimates as surrogates for pre-mRNA and mRNA expression profiles, respectively (Supplementary Tables 18–20). Analysis of pre-mRNA and mRNA expression profile estimates suggested a universally significant, but less than perfect, concordance for most transcripts (Fig. 5b). Notably, pre-mRNA and mRNA expression deviated significantly more for genes with longer 3′ untranslated regions (UTRs) ($P < 8 \times 10^{-37}$), which might be due to tighter 3′ UTR-mediated post-transcriptional regulation[40]. Namely, transcriptional regulation is expected to affect both pre-mRNA and mRNA expression profiles, whereas post-transcriptional regulation is expected to cause deviations between these profiles. Thus, the ratio between a gene's estimated mRNA and pre-mRNA expression profiles (m/p ratio) is expected to be correlated with changes in the gene's post-transcriptional but not its transcriptional regulation.

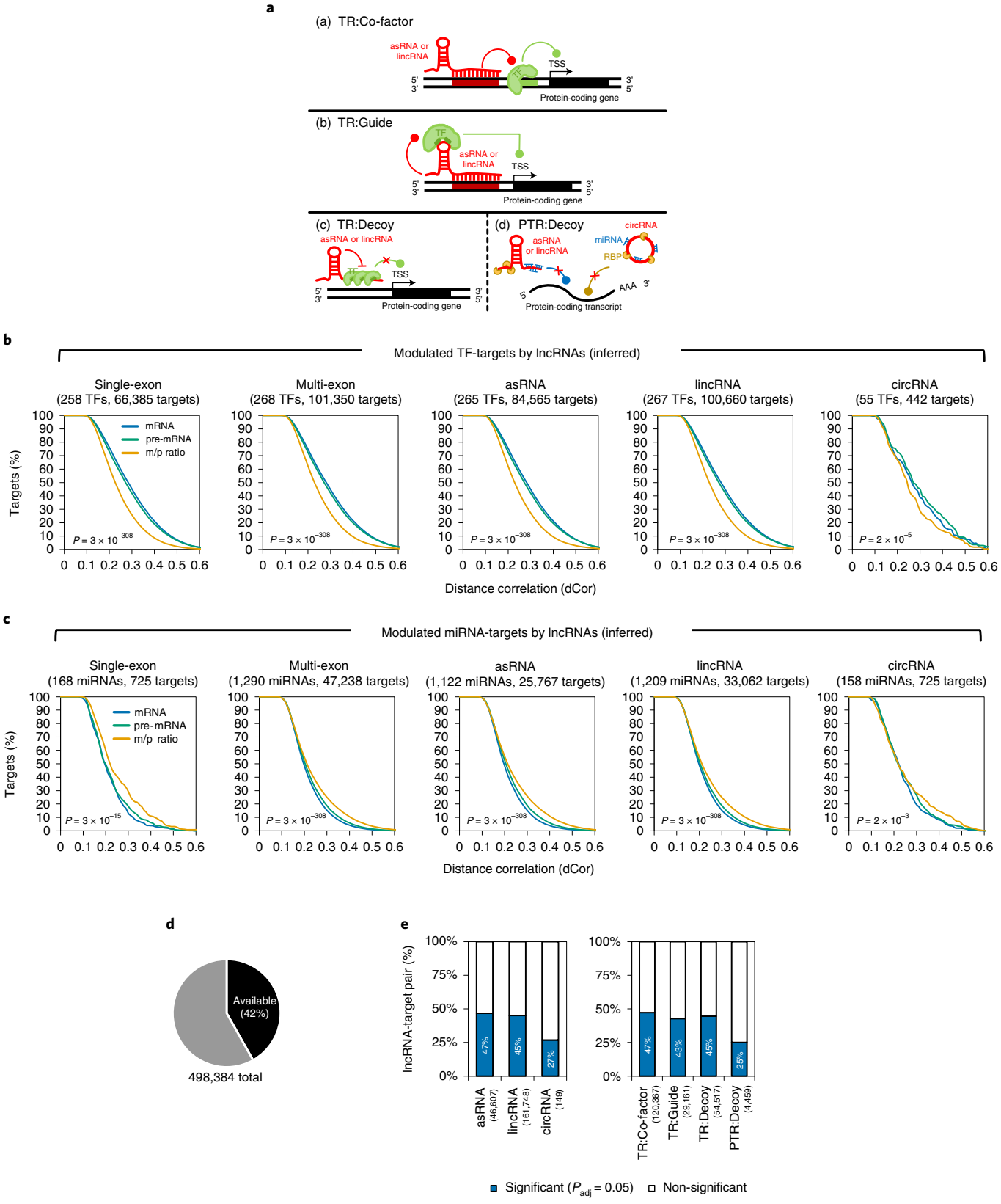
To further study the effects of transcriptional and post-transcriptional regulation in RNA Atlas, we collected verified TF and miRNA targets for 210 TFs and 224 miRBase miRNAs (Supplementary Table 21). TF expression profiles had significantly higher correlations with both the TF's target pre-mRNA and mRNA expression profiles than with the target's mRNA/pre-mRNA ratio (m/p ratio). In contrast, and as expected, miRNA expression profiles had significantly higher correlations with the target m/p ratio than with pre-mRNA and mRNA expression profiles of the target (Fig. 5c,d). These observations are in concordance with the hypothesis that mRNA and pre-mRNA estimates can indicate regulation modalities. Moreover, they provided a means to functionally evaluate miRNAs that were identified in this study with respect to the post-transcriptional regulation of their predicted targets. However, extending this observation to predicted TF and miRNA targets required accurate and dataset-specific regulator-target predictions (Supplementary Table 22), and sequence-based target predictions alone showed little evidence of differences between pre-mRNA and m/p ratio correlations (Supplementary Fig. 22c). In contrast, analyses focusing on miRNAs with predicted targets by Cupid[41] and LongHorn[42] suggested that correlations between miRBase miRNAs and their target m/p ratio profiles were significantly higher than with their target pre-mRNA profiles. Note that LongHorn uses mRNA expression estimates to predict interactions, and, consequently, LongHorn-inferred regulator and target mRNA profiles—but not m/p ratios—are expected to be correlated. We also classified miRDeep2-predicted miRNAs into two sets: one containing miRNAs previously identified in other DBs (Other_DB miRNAs) and the other with miRNAs exclusively found in RNA Atlas (RNA Atlas-only miRNAs). As in the case of miRBase miRNAs, RNA Atlas-only miRNAs had significantly higher correlations with their target m/p ratio than pre-mRNA profiles according to the paired Student's *t*-test ($P < 3 \times 10^{-56}$; Fig. 5e), but this trend was not observed for Other_DB miRNAs.

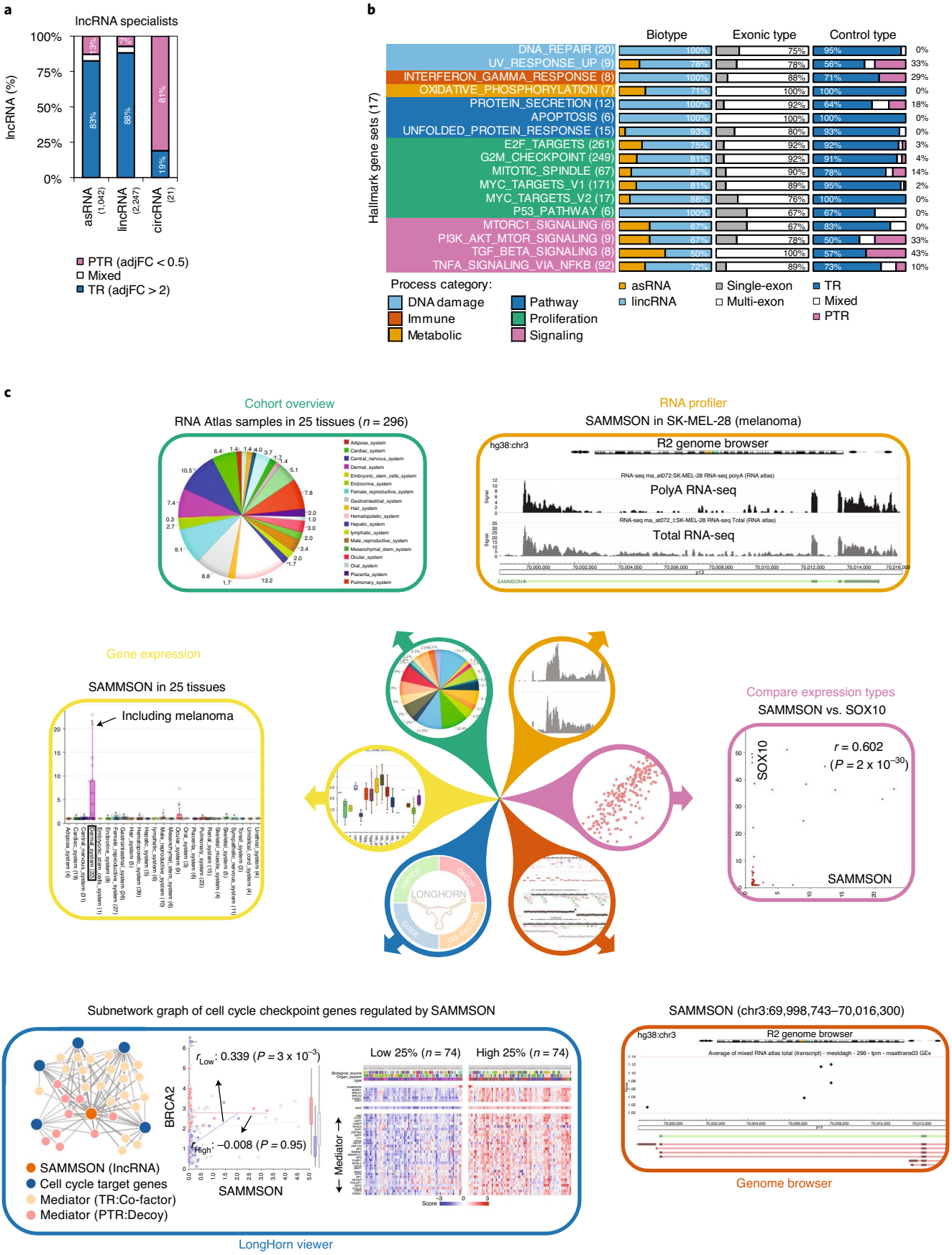
We set restrictive criteria when searching for miRNAs whose expression profiles had significantly higher correlations with target m/p ratio than mRNA and pre-mRNA profiles. Namely, only miRNAs with at least one predicted target with adequate pre-mRNA expression profiles in all RNA Atlas samples were included in the analysis. In total, 757 miRBase miRNAs and 543 miRDeep2-predicted candidate miRNAs satisfied this requirement. Of these, 709 miRBase miRNAs (94%) and 469 miRDeep2-predicted miRNAs (86%) had at least one target for which we observed a higher correlation between the miRNA expression profiles and its target's m/p ratio profiles compared to the target's mRNA and pre-mRNA profiles. However, to test that the expression profiles of miRNAs are significantly more likely to have a stronger correlation with their target's m/p ratios, we required that a significantly greater number of predicted targets have higher miRNA to m/p ratio correlations ($P < 0.05$ by paired Student's *t*-test). In total, of the 735 miRBase miRNAs and 525 miRDeep2-predicted miRNAs that had multiple interactions with target pre-mRNA expression available, 211 miRBase miRNAs (29%) and 105 miRDeep2-predicted miRNAs (20%) satisfied this requirement (Supplementary Table 5). Our analysis suggests that these 211 miRBase miRNAs and 105 miRDeep2-predicted miRNAs are functionally regulating multiple post-transcriptional targets across multiple samples in the RNA Atlas dataset. We note that 18/105 miRDeep2-predicted and 207/211 miRBase miRNAs with functional evidence were also annotated in FANTOM5, miRCarta or MirGeneDB (Supplementary Table 23).

Evidence for transcriptional and post-transcriptional regulation by lncRNAs. Our ability to distinguish transcriptional from post-transcriptional regulation could also provide insight into lncRNA function. We first applied LongHorn[42] to infer regulatory networks downstream of RNA Atlas lncRNAs, including single-exon and multi-exon lincRNAs, asRNAs and circRNAs. LongHorn (Fig. 6a) predicted lncRNA targets by evaluating them as modulators of transcriptional regulation—where lncRNAs are modeled to alter TF regulation as co-factors, guides, or decoys—and of post-transcriptional regulation as miRNA- and RNA-binding protein (RBP) decoys[43–46]. We evaluated whether LongHorn predictions were supported by the expected correlations between regulator

Fig. 6 | Evidence for regulation by lncRNAs. a, LongHorn inferred interactions by evaluating four distinct models for lncRNA regulation, including transcriptional regulation (TR) by lncRNAs as co-factors, guides or TF decoys and post-transcriptional regulation (PTR) as decoys for miRNAs and RBPs. **b**, Predicted TF-target interactions with no supporting evidence from expression did not show correlation differences between TF-target pre-mRNA and m/p ratio (Supplementary Fig. 22c). However, evidence for lncRNA regulation, including regulation by single-exon and multi-exon lncRNAs, antisense, intergenic and circular lncRNAs, resulted in TF-target pre-mRNA profiles that were significantly better correlated than TF-target m/p ratio profiles; the number of tested TFs and interactions for each lncRNA biotype is given in parentheses. Note that $P < 3 \times 10^{-308}$ are *P* values beyond machine precision. Similarly, (**c**) miRNA-target interactions modulated by single-exon and multi-exon lncRNAs and antisense, intergenic and circular lncRNAs showed higher miRNA-target m/p ratio correlations. *P* values represented in **b** and **c** were computed as the geometric means of the one-sided *P* values from the paired Student's *t*-test for delta correlations between regulators and their target's expression profiles, where delta correlations compared correlations of regulators and their target's pre-mRNA or mRNA versus their m/p ratio profiles. **d**, LongHorn predicted 498,384 lncRNA-target interactions, and, among these, 42% had sufficient intronic and exonic read counts to evaluate m/p ratio in all profiled samples. Note that lncRNA-target pairs that were predicted to interact according to multiple models were counted multiple times. **e**, Proportion of interactions for which regulator-target m/p ratio correlations were significantly different from regulator-target pre-mRNA correlations. Results are shown by the biotype of modulating lncRNAs (left) and the type of regulation for these interactions (right); one-sided *P* values were estimated by permutation testing after adjusted for multiple comparisons using the Benjamini–Hochberg procedure, and the number of lncRNAs is given in parentheses (Methods). The regulators were either TFs or miRNAs mediating the lncRNA-target interactions.

expression (that is, TF or miRNA) and target mRNA, pre-mRNA and m/p ratio estimates. For each LongHorn prediction, we selected TF-target interactions based on ENCODE ChIP-seq[47] and TF binding motifs and miRNA-target interactions based on sequence analysis[42]. Analogously to verified TF-target interactions (Fig. 5c), and for all types of lncRNA modulators considered, pre-mRNA and mRNA correlations were consistently higher than m/p ratio correlations for predicted TF-target interactions (Fig. 6b and Supplementary Fig. 22a). Note that predicted TF-target interactions with no evidence of lncRNA modulation did not have substantial differences





between pre-mRNA and m/p ratio correlations (Supplementary Fig. 22c). Similarly, all types of lncRNA-mediated miRNA-target interactions showed significantly higher m/p ratio correlations than mRNA or pre-mRNA correlations (Fig. 6c and Supplementary Fig. 22b). As additional controls, we verified that randomized regulator-target pair sets that were generated by permutation testing

Fig. 7 | Interpretation of lncRNA function. a, LongHorn identified lncRNAs that are enriched for predicted transcriptional or post-transcriptional interactions, or for both, relative to other lncRNA species. asRNAs and lincRNAs were more likely to be identified as transcriptional regulators, whereas circRNAs were more likely to be identified as post-transcriptional regulators; each of these lncRNAs was required to have at least ten LongHorn-inferred targets; the number of lncRNAs included in each category is given in parentheses. **b**, Seventeen MSigDB's hallmark gene sets were predicted to be significantly regulated by at least five lncRNAs at $P < 0.01$. *P* values were one sided and estimated by FDR-adjusted Fisher's exact test. lncRNAs were predicted to preferentially target proliferation and signaling pathways; the total number of regulating lncRNAs in each pathway is provided in parentheses. **c**, The full lncRNA-target prediction data are available for download and analysis on the R2 platform. R2 also allows the analysis and visualization of lncRNA abundance and regulatory network modules predicted by LongHorn. We show an example of an R2 analysis for the lncRNA SAMMSON.

did not show differences in correlations even for 100,000 simulated interactions (Supplementary Figs. 22d,e and 25). Lastly, we note that target, miRNA and TF expression variabilities, and the number of samples in which they were expressed, did not qualitatively influence our observations and conclusions that TFs and miRNAs, whose regulation was evidenced by LongHorn-inferred lncRNA modulators, are significantly less and more correlated with their target's m/p expression ratios, respectively (Supplementary Figs. 23 and 24). Collectively, this result suggests that differences observed for predicted interactions were not expected by random chance, independently of the number of regulator-target pairs considered or expression features of targets and regulators (Fig. 6, Supplementary Figs. 22–25 and Supplementary Tables 24 and 25).

These observations suggested that all lncRNA biotypes, including single-exon lncRNAs and non-polyadenylated lncRNAs, are effectively altering TF and miRNA regulation. In addition, for many lncRNA species, the analysis pointed to lncRNA specialization as either transcriptional or post-transcriptional regulators. In total, we predicted that 498,384 lncRNA-mediated interactions and targets included in 208,543 of these (Fig. 6d) had pre-mRNA, mRNA and m/p ratio expression estimates in all samples. Differences between the correlations of regulator expression and the pre-mRNA and m/p ratio expression profiles of their targets, for each regulatory modality, were significant for most of LongHorn's predicted interactions at Benjamini–Hochberg-adjusted $P < 0.05$ by permutation testing (Fig. 6e, Methods and Supplementary Table 25). However, relatively fewer predicted circRNA-mediated and post-transcriptional decoy-mediated interactions had significant correlation differences. An analysis of lncRNA regulatory models inferred by LongHorn suggested that circRNAs are predominantly post-transcriptional decoys, whereas other lncRNAs predominantly modulate transcription (Fig. 7a). Although these observations are based on just 21 circRNAs and might, therefore, not generalize to all circRNAs or all decoys, they are in line with earlier observations demonstrating the enrichment of circRNAs in the cytoplasm and lncRNAs in the nucleus[24,48,49]. In total, of the 1,221 single-exon lncRNAs investigated, 960 (79%) were predicted to modulate at least one interaction that had significant differences between the correlations of a regulator and the target's pre-mRNA and m/p ratio. Similarly, 4,092 (88%) of our multi-exon lncRNAs showed this relationship. To validate these predicted interactions, we analyzed RNA sequencing data from lncRNA perturbation experiments (publicly available and newly generated). Our result suggested that, for 20/24 (83%) and 15/24 (63%) of the tested lncRNAs, LongHorn-inferred lncRNA targets were more likely to be dysregulated (odds ratio > 1) and significantly dysregulated ($P < 0.05$, by Fisher's exact test) after the downregulation of the corresponding lncRNA, respectively. Details of this analysis are described in Methods (Supplementary Figs. 26 and 27 and Supplementary Tables 26 and 27).

In total, 3,310 lncRNAs, including 2,012 Annotated, 1,174 PreRep and 124 RNA Atlas-only, were associated with ten or more LongHorn-inferred targets. Our analysis suggested that many of these lncRNAs preferentially target key pathways in disease and development (FDR-adjusted Fisher's exact test < 0.01). To study this further, we catalogued lncRNAs according to their predicted targets' enrichment in hallmark pathways[50] and their specialization as transcriptional or post-transcriptional modulators. In total, 17 pathways were enriched in targets from at least five lncRNAs (Fig. 7b). Our analysis suggested that, overall, lncRNAs preferentially target proliferation and signaling pathways. The full analysis is given in Supplementary Table 28, and pathway enrichments for 15 lncRNAs from each class (Annotated, PreRep and RNA Atlas-only) are depicted in Supplementary Fig. 28. The full lncRNA-target prediction data are available for download and analysis on the R2 platform (http://r2platform.com/rna_atlas). R2 allows for comparing profiles across technologies, visualizing lncRNA expression across samples and studying regulatory network modules predicted by LongHorn, including analyzing correlations among lncRNAs, TFs, miRNAs and predicted targets. Figure 7c depicts an example of the R2 analysis module using the lncRNA SAMMSON[16].

Discussion

By applying three complementary RNA sequencing technologies on a heterogeneous collection of tissues, cell types and cell lines, we assembled a comprehensive human transcriptome covering all major RNA biotypes. Our effort both complements other consortium-based efforts aimed at generating human expression atlases[4–13] and extends the scopes of RNA catalogs by integrating multiple RNA sequencing technologies, which allowed for a variety of follow-up analyses. Namely, the total RNA sequencing component of the RNA Atlas dataset revealed novel non-polyadenylated lincRNAs, enabled us to determine transcript polyadenylation status and allowed pre-mRNA expression estimates by quantifying intronic RNA abundance. The latter was crucial to distinguishing transcriptional from post-transcriptional regulation, which helped define a set of regulatory RNAs with evidence of post-transcriptional regulation in multiple tissues. We note that, although we imposed minimal abundance criteria for the candidate miRNAs, we did not implement other criteria defined in the miRBase high-confidence checklist[26] and previously used by the FANTOM consortium[8]. Notably, several miRBase miRNAs have validated target genes but do not adhere to the high-confidence criteria, which are mainly related to sequence characteristics. We, therefore, reasoned that miRNA-like behavior and evidence for regulatory function outweigh sequence characteristics when cataloguing miRNAs.

The integration of miRNA and whole-transcriptome profiling across a variety of tissues enabled analyses that allowed us to evaluate RNA Atlas-predicted ncRNA species for functional evidence. Consequently, in addition to supporting the inclusion of previously predicted miRNAs and lncRNAs and identifying new species, we were also able to collect multiple lines of evidence for their functional relevance in human cells and to interpret their function through both transcriptional and post-transcriptional regulatory interactions. The resulting RNA Atlas dataset and analysis products serve as a resource to mine the expression and regulatory landscapes of multiple RNA biotypes and contain a unique collection of ncRNAs together with their functional interpretation. Dedicated experimental validation studies based on genetic perturbations coupled to phenotypic or molecular readouts should follow to evaluate ncRNA function[51] in each studied context. Moreover, we envision

that the ncRNA regulatory interactions that are presented will serve as a starting point for follow-up studies to gain insights into the mode of action of hundreds of ncRNAs.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-021-00936-1.

Received: 10 December 2019; Accepted: 26 April 2021;
Published online: 17 June 2021

References

1. Esteller, M. Non-coding RNAs in human disease. *Nat. Rev. Genet.* **12**, 861–874 (2011).
2. Chen, L.-L. The biogenesis and emerging roles of circular RNAs. *Nat. Rev. Mol. Cell Biol.* **17**, 205–211 (2016).
3. Lorenzi, L. Long noncoding RNA expression profiling in cancer: challenges and opportunities. *GenesÿChromosomes Cancer* **58**, 191–199 (2019).
4. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
5. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
6. Forrest, A. R. R. et al. A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
7. Hon, C. C. et al. An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature* **543**, 199–204 (2017).
8. De Rie, D. et al. An integrated expression atlas of miRNAs and their promoters in human and mouse. *Nat. Biotechnol.* **35**, 872–878 (2017).
9. Pertea, M. et al. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.* **19**, 208 (2018).
10. Iyer, M. et al. The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47**, 199–208 (2015).
11. Vo, J. N. et al. The landscape of circular RNA in cancer. *Cell* **176**, 869–881 (2019).
12. Regev, A. et al. The Human Cell Atlas. *eLife* **6**, e27041 (2017).
13. You, B. H., Yoon, S. H. & Nam, J. W. High-confidence coding and noncoding transcriptome maps. *Genome Res.* **27**, 1050–1062 (2017).
14. Melé, M. et al. Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).
15. Arun, G., Diermeier, S. D. & Spector, D. L. Therapeutic targeting of long non-coding RNAs in cancer. *Trends Mol. Med.* **24**, 257–277 (2018).
16. Leucci, E. et al. Melanoma addiction to the long non-coding RNA SAMMSON. *Nature* **531**, 518–522 (2016).
17. Hosono, Y. et al. Oncogenic role of THOR, a conserved cancer/testis long non-coding RNA. *Cell* **171**, 1559–1572 (2017).
18. Cunningham, F. et al. Ensembl 2015. *Nucleic Acids Res.* **43**, D662–D669 (2015).
19. Roadmap Epigenomics Consortium, K. A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–329 (2015).
20. Liu, S. J. et al. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355**, eaah7111 (2017).
21. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
22. O'Leary, N. A. et al. Reference Sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
23. Vromman, M., Vandesompele, J. & Volders, P.-J. Closing the circle: current state and perspectives of circular RNA databases. *Brief Bioinform.* **22**, 288–297 (2021).
24. Jeck, W. R. et al. Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* **19**, 141–157 (2013).
25. Memczak, S. et al. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495**, 333–338 (2013).
26. Kozomara, A. & Griffiths-Jones, S. MiRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42**, 68–73 (2014).
27. Friedländer, M. R., MacKowiak, S. D., Li, N., Chen, W. & Rajewsky, N. MiRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.* **40**, 37–52 (2012).
28. Backes, C. et al. miRCarta: a central repository for collecting miRNA candidates. *Nucleic Acids Res.* **46**, D160–D167 (2018).
29. Fromm, B. et al. MirGeneDB 2.0: the metazoan microRNA complement. *Nucleic Acids Res.* **48**, D132–D141 (2020).
30. Wang, L. et al. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41**, e74 (2013).
31. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275–i282 (2011).
32. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
33. Kim, M. S. et al. A draft map of the human proteome. *Nature* **509**, 575–581 (2014).
34. Frese, S. et al. Long-term endurance exercise in humans stimulates cell fusion of myoblasts along with fusogenic endogenous retroviral genes in vivo. *PLoS ONE* **10**, e1032099 (2015).
35. Yang, L., Duff, M. O., Graveley, B. R., Carmichael, G. G. & Chen, L. L. Genomewide characterization of non-polyadenylated RNAs. *Genome Biol.* **12**, R16 (2011).
36. Cabili, M. et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).
37. Baran, Y. et al. The landscape of genomic imprinting across diverse adult human tissues. *Genome Res.* **27**, 927–936 (2015).
38. Yoshihara, K. et al. The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene* **34**, 4845–4854 (2015).
39. Uhlén, M. et al. Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
40. Gaidatzis, D., Burger, L., Florescu, M. & Stadler, M. B. Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. *Nat. Biotechnol.* **33**, 722–729 (2015).
41. Chiu, H. et al. Cupid: simultaneous reconstruction of microRNA-target and ceRNA networks. *Genome Res.* **25**, 257–267 (2015).
42. Chiu, H. S. et al. Pan-cancer analysis of lncRNA regulation supports their targeting of cancer genes in each tumor context. *Cell Rep.* **23**, 297–312 (2018).
43. Karreth, F. A. & Pandolfi, P. P. CeRNA cross-talk in cancer: when ce-bling rivalries go awry. *Cancer Discov.* **3**, 1113–1121 (2013).
44. Poliseno, L. et al. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**, 1033–1038 (2010).
45. Salmena, L., Poliseno, L., Tay, Y., Kats, L. & Pandolfi, P. P. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* **146**, 353–358 (2011).
46. Tay, Y., Rinn, J. & Pandolfi, P. P. The multilayered complexity of ceRNA crosstalk and competition. *Nature* **505**, 344–352 (2014).
47. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
48. Salzman, J., Gawad, C., Wang, P. L., Lacayo, N. & Brown, P. O. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS ONE* **7**, e30733 (2012).
49. Djebali, S. Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
50. Liberzon, A. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
51. Ramilowski, J. A. Functional annotation of human long noncoding RNAs via molecular phenotyping. *Genome Res.* **30**, (2020).

Methods

Sample cohort. A total of 300 human samples were used in this study, including 45 tissues, 162 cell types and 93 cell lines, of which 89 are cancer cell lines derived from 13 different types of cancer (Supplementary Table 1). RNA of individual cell types was obtained from ScienCell Research Laboratories or isolated from cell types collected at Ghent University Hospital. RNA from collected cell types and (cancer) cell lines was isolated using the miRNeasy kit (Qiagen) according to the manufacturer's instructions. RNA samples from normal human tissues were obtained from Ambion and BioChain.

Library prep and sequencing. For each RNA sample, three different strand-specific RNA libraries were prepared. Small RNA libraries were generated using the TruSeq Small RNA Library Prep Kit (Illumina) according to the manufacturer's instructions, using 750-ng input RNA. Library size selection was performed using a Pippin Prep device (Sage Science). Total RNA libraries were generated using the TruSeq Stranded Total RNA Library Prep Kit with Ribo-Zero Gold (Illumina) according to the manufacturer's instructions using 1 µg of input RNA. PolyA RNA libraries were generated using the TruSeq Stranded mRNA Library Prep Kit (Illumina) according to the manufacturer's instructions using 1 µg of input RNA. Small RNA libraries were pooled (volume-based pooling, 48 libraries per pool), and pools were quantified using the High Sensitivity dsDNA assay on a Bioanalyzer device (Agilent). PolyA and total RNA library pools were quantified using the Standard Sensitivity NGS Fragment Kit (cat. no. DNF-473) on a Fragment Analyzer (Advanced Analytical). Small RNA library pools were sequenced on a NextSeq 500 instrument (Illumina) using a high-output flow cell, 76 cycles. Pooled polyA and total RNA libraries were sequenced on a HiSeq 4000 instrument (Illumina) with paired-end 76 cycle reads.

Transcriptome assembly from polyA and total RNA sequencing libraries. PolyA and total RNA reads were mapped to the hg38 reference genome (primary assembly, canonical chromosomes, repeats from RepeatMasker and Tandem Repeats Finder soft masked, http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/) with TopHat[52] version 2.1.0 (Bowtie2 (ref. [53]) v2.2.6) using the --no-coverage-search option and --library-type=fr-firststrand. The Ensembl[18] transcriptome (v86, which was the latest available at the start of the project—download link: ftp://ftp.ensembl.org/pub/release-86/gtf/homo_sapiens) was provided to guide the mapping of reads to known transcripts first. All other parameters were set to default values. Transcriptomes were assembled in each sample and each library type separately using StringTie (ref. [54]) (v1.3.3). Default parameters were used, and the Ensembl[18] reference annotation (v86) was supplied to guide each assembly (Supplementary Table 2 and Supplementary Fig. 2b).

All individual transcriptomes were then merged together with the reference Ensembl transcriptome (v86) using StringTie merge applying a cutoff of 1 transcript per million (TPM) (Supplementary Fig. 1c–e) and minimum transcript length of 200 nucleotides, with all other parameters set to default values.

CuffCompare[55] (v2.2.1) was used to compare the newly assembled transcripts with the reference annotation. Non-annotated transcripts with classification codes other than 'x', 'u', 'j' or '=' were removed (this included 20,539 transcripts with codes 'c', 'e', 'I', 'o', 'p' or 's') as well as transcripts spanning two or more known, non-overlapping, adjacent loci.

We then calculated the overlap of all known and non-annotated exons with repetitive elements in the genome using BEDTools[56] (intersect). The repeats regions were retrieved from the UCSC Table Browser[57] (Group: Repeats; Track: RepeatMasker). The fraction of exonic sequence overlapping repeats was computed for each gene. Non-annotated, non-coding, single-exon genes with fewer than 200 consecutive non-repeat nucleotides were filtered out. Regions overlapping repeats within the remaining non-coding, single-exon genes were hard-masked (bases were converted to Ns) using BEDTools[56] (v2.27.1, maskfasta). After these filtering steps, the polyA/total RNA sequencing-derived transcriptome contained 422,083 transcripts, including all transcripts in Ensembl version 86 annotation. This transcriptome was quantified using Kallisto quant[58] (flag --rf-stranded and all other parameters set to default values) across all polyA and total RNA sequencing libraries.

After quantification, newly assembled and Ensembl genes belonging to the biotypes 'protein_coding', 'antisense' and 'lincRNA' were retained. Because of previous filtering steps at transcript level, non-annotated genes are either intergenic (lincRNA) or antisense (asRNA) to reference genes. Ensembl and newly assembled genes with expression levels below 0.1 TPM in all samples were removed.

Selection of the RNA Atlas transcriptome. Independent evidence for transcription of RNA Atlas PCG, lincRNA and asRNA genes was obtained by integrating results from CAGE sequencing from the FANTOM consortium[6] and various chromatin states from the Roadmap Epigenomics Project[19]. The following chromatin states, indicative for active transcription, were considered: active transcription start site (1_TssA), transcription (5_Tx5, 6_Tx and 7_Tx3), transcribed and regulatory (9_TxReg), transcribed and enhancer (10_TxEnh5 and 11_TxEnh3), active_enhancer (13_EnhA1 and 14_EnhA2) and bivalent_promoter (23_PromBiv)[19]. For each TSS of genes with expression values greater than or equal to 0.1 TPM in at least one tissue from the RNA Atlas and not being part of chromosome Y (chromatin states did not include information for that chromosome), we used the Zipper plot approach[59] to retrieve the closest CAGE sequencing[6] and chromatin state[19] peak across all samples from FANTOM5 and the Roadmap Epigenomics Project, respectively. We refined our gene set based on presence of the aforementioned peaks within 500 nucleotides upstream or downstream the TSS and further classified the genes across the different RNA biotypes into four categories: (1) evidence at both DNA (chromatin state) and RNA (CAGE sequencing) levels; (2) evidence at RNA level only; (3) evidence at DNA level only; and (4) no evidence. Genes belonging to one of the first three categories were retained in the RNA Atlas transcriptome. We also included 642 genes that do not present close association to any CAGE peak or chromatin state peaks but whose expression levels were higher than the median expression levels for genes that present both levels of evidence. Specifically, we retained genes with no independent evidence if their expression in at least four samples from either polyA or total RNA libraries was higher than the median value of the mean expression across samples for genes with both levels of evidence (4.67 TPM and 2.20 TPM for polyA and total RNA data, respectively). This set included 473 PCGs, 21 asRNAs and 148 lincRNAs.

Comparison with reference databases and gene sets from other large-scale efforts. Ensembl (v86) was used to guide the assembly of the primary transcriptome and for biotype definition at project inception. We later annotated our RNA Atlas set using GENCODE (v33) and RefSeq (v200) annotations (as of 1 May 2020), in addition to other recent transcriptome resources derived from large-scale studies, including FANTOM5 (ref. [7]), BigTranscriptome[13], MiTranscriptome[10] and CHESS[9]. We note that the Ensembl/Havana merged gene set is equivalent to the GENCODE gene set[21]. To perform these comparisons, we first combined the set of genes in GENCODE and 'curated' RefSeq records (that is, including only 'NM_' and 'NR_' identifiers—see https://www.ncbi.nlm.nih.gov/refseq/about) into a joint transcriptome that we named Annotated RNAs. In the same way, we combined all RNAs predicted by FANTOM5 stringent set, BigTranscriptome, MiTranscriptome, CHESS and the 'model' set of RefSeq (that is, including only 'XM_' and 'XR_' identifiers—see https://www.ncbi.nlm.nih.gov/refseq/about/) into a joint transcriptome that we named Previously Reported (PreRep) RNAs. See 'Generation of combined transcriptomes' below for details on how all combinations were performed.

We then computed the overlap of the set of RNA Atlas RNAs with each of these combined transcriptomes separately, using the CuffCompare[55] tool. In each case, each RNA Atlas gene was considered to overlap the reference transcriptome if at least one of its transcripts was assigned one of the class codes '=', 'j', 'c', 'o' or 'e'. We used this information to classify the set of RNA Atlas genes into three categories: those that overlap genes in the Annotated set, named 'Annotated' RNAs; those that overlap genes in the PreRep set but not in the Annotated set, named 'PreRep' RNAs; and those that do not overlap genes in either transcriptome, named 'RNA Atlas-only' RNAs.

Note that the PreRep transcriptome includes the set of FANTOM5 stringent genes. We next checked the overlap between RNA Atlas genes and FANTOM5 robust genes with the CuffCompare[55] tool. Using the criteria described in the previous paragraph, 990 (18%) RNA Atlas-only genes were found to overlap FANTOM5 robust genes. We added the overlap with this dataset as an additional column in Supplementary Table 3. Additionally, we performed the same comparison using the GffCompare[60] tool, an updated version of CuffCompare, to evaluate possible differences between both tools. In this case, we also considered the class codes 'k', 'm' and 'n', not present in CuffCompare, as indicative of overlap. The differences in the results obtained by both tools are marginal (with six extra genes reported as matches by CuffCompare but not by GffCompare) and related to differences in the assigned class codes for a small fraction of transcripts. Overall, this suggests that using GffCompare instead of CuffCompare would not affect the conclusions put forward in this study.

Construction of Venn diagram for lncRNAs. The sections that compose the Venn diagram in Supplementary Fig. 4a were calculated as follows. For RNA Atlas lncRNAs (sections in white numbers), the classification approach described above, based on the overlap with transcripts contained in each of the other two datasets, was used. To retrieve information for those lncRNAs not included in RNA Atlas (sections in black numbers), we first generated a joint transcriptome by combining the RNA Atlas, Annotated and PreRep sets (see 'Generation of combined transcriptomes'). From this combined annotation, we first filtered out all genes that contained RNA Atlas transcripts. From the remaining non-RNA Atlas combined loci, we selected those classified as lncRNAs based on a consensus call across transcripts (see 'Transcript and gene biotype definition for combined annotations') and then classified them as PreRep only (if the lncRNA contained only transcripts coming from PreRep set), Annotated only (if the lncRNA contained only transcripts coming from the Annotated set) or common to PreRep and Annotated (those lncRNAs that had transcripts from both datasets). We finally summarized this information into the Venn diagram shown in the figure.

Generation of combined transcriptomes. Combined transcriptomes were obtained with the CuffCompare[55] tool (output files with '.combined.gtf' suffix). This tool performs a joint union at transcript level by keeping all individual transcripts from all databases including tracking information of where they come

from, and it only collapses transcripts across references into unique entries if they have exactly matching intron chain (that is, only admits differences in the start and end exons). As for locus level, it merges and assigns unique locus identifiers to all transcripts that overlap in a certain genomic region. Note that this is a rather conservative approach because it retains small differences between annotations and often results in relatively large transcript sets. Also note that the resulting gene definition in certain loci might differ from the individual gene definitions across databases being combined, and, therefore, it is not always possible to define a one-to-one relationship between a gene in an individual database and a gene in the resulting combined annotation.

Transcript and gene biotype definition for combined annotations. Different biotype definitions across datasets were grouped in five main categories: 'protein-coding genes', 'lncRNAs' (this includes antisense, lincRNAs and divergent and sense-overlapping ncRNAs), 'pseudogenes', 'other genes' (this category includes rRNA, tRNA, scaRNA, snoRNA, sRNA, miRNA and other small RNAs) and 'uncertain' (the biotype is not defined, ambiguous or uncertain).

For transcripts in the PreRep dataset that matched transcripts in the Annotated dataset (class codes '=', 'j', 'c' or 'e' from CuffCompare), the biotype from Annotated was taken (and, within this, GENCODE over RefSeq annotation was taken if they differed). For transcripts in the PreRep dataset that did not have a match in the Annotated dataset, a majority vote across the individual databases was taken among the non-uncertain annotations; in case of ties, 'protein-coding' biotype was prioritized first, then 'pseudogene', then 'lncRNA' and, finally, 'other'. Genes in the combined annotations were annotated using these rules: if any transcript in the locus is 'protein-coding', the entire locus is annotated as protein-coding; otherwise, a majority vote among the non-uncertain biotypes is taken where ties are solved by taking the priority order of 'pseudogene' first, then 'lncRNA', then 'other'.

TF binding analyses. We sought to identify an independent objective method to evaluate prediction accuracy for the different sets of lncRNAs. We argued that accurate lncRNA gene models are more likely to harbor regulatory elements in their predicted promoters, which would support their status as independently regulated transcriptional units. To assess this, we collected all TSSs for Annotated, PreRep and RNA Atlas-only lncRNAs and compared TF occupancy (number of TFs with significant -qval <0.05- ENCODE ChIP-seq peaks) at 2-kb promoters centered at lncRNA TSSs. Our evaluation, of nearly 140,000 lncRNA TSSs, was based on ENCODE-identified binding sites for 161 TFs in 90 cell types and tissue samples. Note that only 36 (40%) of these sample types were profiled by RNA Atlas, whereas all of these sample types were included in analyses of lncRNAs from other curation efforts. When computing overlaps, we matched cell lines and tissue types, and, in total, 11 cell lines and 25 tissue types (from non-matching donors) were profiled by both RNA Atlas and ENCODE.

Evaluation of coding potential. To assess the protein-coding potential of the newly assembled transcripts, two algorithms were used: the Coding-Potential Assessment Tool[30] (CPAT, v2.0.0) and PhyloCSF[31] (obtained from https://github.com/mlin/PhyloCSF, 18 January 2015). The CPAT code was slightly modified so that the predicted ORF sequence is returned in the output. CPAT was run using the provided hexamer table and logit model. The recommended probability cutoff of 0.364 was used to identify putative coding ORFs.

Because the PhyloCSF pipeline is based on the GRCh37/hg19 reference genome, all genomic coordinates were first converted using the liftOver tool on the UCSC Genome Browser website[57]. To run PhyloCSF, whole-genome alignments of 46 species are obtained from the UCSC Genome Browser website[57] and processed using the Phylogenetic Analysis with Space/Time Models package[61] (PHAST, v1.4) into the required input format. PhyloCSF was run in three reading frames using the ATGStop setting; all ORFs of at least ten codons were considered. A cutoff score of 60.7876, based on benchmarking with Ensembl (v90)[62] transcripts, was used to identify putative coding ORFs. In total, 188 newly assembled genes had at least one isoform scored as protein-coding by both tools.

Alignment of the protein sequence to other animal proteins via BLASTp. The existence of similar proteins across different species (cross-species protein conservation) was evaluated through alignment of our predicted ORFs to UniProt protein reference clusters (UniRef90) using BLASTp (BLAST+[32] package v2.9.0). Off 188 evaluated genes, 109 had at least one hit with an E value < 0.001. When compared to up-to-date annotations (see sections above), we found that 21 of these candidates were annotated as non-coding genes in the Annotated dataset; 59 and 15 were annotated as non-coding and protein-coding, respectively, in the PreRep dataset; and nine were unique for RNA Atlas (that is, RNA Atlas-only class). Finally, five of these genes were annotated as protein-coding genes in the Annotated dataset and were excluded from our list of new candidate PCGs. Note that our final list of 104 new candidate protein-coding genes includes only genes in the RNA Atlas transcriptome that were predicted to have high coding potential by both evaluated algorithms (CPAT and PhyloCSF—see above) and that have at least one BLASTp hit with an E value ≤ 0.001. Detailed information for these 104 candidate protein-coding genes is provided in Supplementary Table 7.

Conservation with chimpanzee. Whole-genome alignments of 99 species against human (multiz100way, hg38) were obtained from the UCSC website[57] and processed using the PHAST[61] package (v1.3). From these whole-genome alignment files, local alignments of the human (hg38) and chimpanzee (panTro4) genomes were extracted for the predicted ORFs. From these, the fraction of conserved bases and amino acids (through in silico translation) was calculated for each predicted ORF. The approach was repeated for Ensembl protein-coding ORFs for comparison.

Mass spectrometry validation. Mass spectrometry validation of the predicted proteins was conducted on the draft map of the human proteome dataset[33]. Briefly, this dataset consists of deep proteomic profiling of 17 adult tissues, seven fetal tissues and six purified primary hematopoietic cells. Raw files were obtained from the PRoteomics IDEntifications (PRIDE) database[63] (project PXD000561) and converted to Mascot generic format (MGF) using the msconvert tool in the ProteoWizard package[64].

Analysis of the tandem mass spectrometry data was performed using Ionbot (unpublished, based on the work of Silva et al.[65], https://ionbot.cloud), a sequence database search tool based on machine learning capable of performing rapid open modification and open mutation searches. Ionbot was used under a beta-tester version supplied by Sven Degroeve and Lennart Martens (Ghent University, VIB). Briefly, peptide databases were created as a full in silico trypsin digestion (allowing up to one missed cleavage) of the protein sequence dataset consisting of all human proteins in the UniProt in the UniProtKB database[66] (Swiss-Prot subset, 21,008 proteins) and the CPAT- and PhyloCSF-predicted proteins. Decoy peptides were obtained by applying the same digestion on the reversed target proteins. Ionbot was run in the open modification and open mutation mode. In addition, carbamidomethylation of cysteine was set as fixed and oxidation of methionine as variable modification. The FDR was estimated with the target decoy approach. Only peptide spectrum matches with an estimated FDR below 1% were retained.

Identification and quantification of circRNAs. circRNAs were identified from total RNA sequencing data using two independent workflows—find_circ2 (ref. [25]) (n = 85,470) and CIRCexplorer2 (ref. [67]) (n = 204,857)—using genome build hg19. For downstream analysis, the mean circRNA count across methods was used. Only circRNAs identified by both tools (n = 62,832) and with mean counts between tools higher than 4 in at least one sample were retained (n = 38,030). Genomic positions of 38,023 circRNAs were successfully converted to hg38 coordinates using the liftOver tool (UCSC)[57]. The back-splice acceptor and donor sites from each circRNA were annotated relative to other linear splice sites and gene coordinates from PCGs, asRNAs and lincRNAs. circRNAs with a back-splice acceptor and donor site overlapping genes in the RNA Atlas transcriptome were retained as RNA Atlas circRNAs (n = 37,128).

Flanking intron length analysis. We compared the length of the introns (both upstream and downstream) that flanked the circRNAs to introns from genes that do not produce a circRNA isoform. The flanking introns were found to be unusually long when compared to the non-circRNA introns, as shown in Supplementary Fig. 5. The median length for the flanking introns was 6,304 bp compared to the median value for non-circRNA introns, which was observed to be 1,041 bp. Statistical significance of the difference was assessed with the Wilcoxon rank-sum test. Box plots were drawn in R to display the results.

Overlap of RNA Atlas circRNAs with multiple circRNA databases. We then compared the 37,128 circRNAs derived from loci in the RNA Atlas set to 13 public circRNA resources, including circBase (n = 91,793), CircAtlas (n = 609,839), circbank (n = 140,135), circRNADb (n = 32,863), CSCD (n = 1,220,058), CIRCpediav2 (n = 177,43), MiOncoCirc (n = 227,055), TSCD (n = 284,293), circad (n = 592), CircR2disease (n = 450), Circ2Disease (n = 225), exoRBase (n = 57,412) and CircRiC (n = 92,564). This dataset, previously described by Vromman et al.[23], was collected by downloading all available circRNA database exports, processing the circRNA entries using RStudio (v1.2.1335) and converting all back-splice junction positions to the hg38 genome build. As some databases do not provide strand information, when circRNA entries were found with non-assigned strands, this information was borrowed, if available, from other circRNAs (including RNA Atlas circRNAs) with exact matching chr, start and end positions. After this strand information imputation, we computed the number and fractions of exact matches between RNA Atlas circRNAs and circRNAs in each individual database (Supplementary Fig. 5c,d).

Candidate miRNA identification and quantification. Reads from small RNA sequencing libraries were processed with the FASTX-Toolkit[68] (v0.0.14) to remove adapters, filter reads by quality (a quality score of at least 20 in 80% of the sequence was required) and collapse non-unique reads. Processed reads were then mapped against the hg38 genome with Bowtie[69] (v1.1.2) allowing no mismatches in the first 25 bases of the read (-n = 0 and l = 25) and using the '--best' option to force reporting of up to ten (k = 10) best alignments (all other parameters were set to default values). Candidate miRNAs were predicted with miRDeep2 (ref. [27]) (v2.0.0.8), using mapped reads per sample and all human miRNAs in miRBase[26]

version 22 as input. Novel miRNA predictions with non-zero estimated probability were aggregated across samples, retaining only the prediction with maximum counts from both mature forms in a given sample in cases of predictions with partially overlapping coordinates. Reads mapped to the aggregated newly predicted miRNAs and human miRNAs from miRbase version 22 were then quantified across all samples. For each miRNA, counts from the canonical mature form and non-canonical mature forms (isoMiRs) were aggregated. Only candidate miRNAs with ten or more counts in at least one sample were retained.

Overlap of RNA Atlas candidate miRNAs with other miRNA databases. We compared our miRNA candidates against mature miRNAs annotated in other well-known miRNA resources, including the set of candidate miRNAs identified by FANTOM5 (ref. [8]) ($n=564$) and miRNAs annotated in miRCarta[28] ($n=24,742$) and MirGeneDB[29] ($n=1,017$) databases. Similarly to what we did for lncRNAs, we performed a three-set comparison among RNA Atlas candidates (these included 1,646 miRBase (v22) miRNAs that are expressed in RNA Atlas samples and 3,567 miRDeep2-predicted miRNAs from our RNA sequencing data), all mature miRNAs annotated in miRBase ($n=2,712$) and the combination of FANTOM5 candidate miRNAs, miRCarta and MirGeneDB ($n=25,420$). Any overlap between genomic coordinates of mature miRNAs in the same strand was considered as a match. This information was summarized in a Venn diagram (Supplementary Fig. 4b).

Genomic analyses of single-exon lincRNAs. The distance from each unique RNA Atlas exon to its closest upstream or downstream exon was retrieved with BEDTools[56] (v2.27.1, bedtools closest -io -D a -s). A two-sample Wilcoxon rank-sum test was used to compare the distances between single-exon lincRNAs and multi-exon lincRNA exons. Sequence motifs at the exon–intron boundary of multi-exon genes and exon–intergenic boundary of novel single-exon genes were determined by calculating the frequency of each nucleotide at each position of the region starting 3 bases upstream and ending 3 bases downstream of the boundary. This was done for multi-exon PCG exons, multi-exon lincRNA exons and single-exon lincRNA exons. The information content was computed for each position, and the relative frequencies of each base in each position were represented as a sequence logo with the R package ggseqlogo[70]. For strandedness analyses, we selected unique exons with no overlap with any feature on the opposite strand. Only exons with ten or more counts on the correct strand in at least one sample were considered. The strandedness for each selected exon was defined using the sample with maximum normalized expression on the correct strand, as the percentage of reads mapping to the exon on the correct strand relative to all reads mapping to the exon regardless of the strand.

RT–qPCR validation of single-exon genes. We performed qPCR validation of single-exon genes using RNA from two RNA Atlas cell lines: SK-N-BE(2)-C and IMR-32. We designed specific forward and reverse primers for the amplification of a total of 110 single-exon genes, including 42 RNA Atlas-only genes, of which 24 were ubiquitously expressed, 11 were specifically expressed in SK-N-BE(2)-C and seven were specifically expressed in IMR-32; 27 PreRep genes, of which 19 were ubiquitously expressed, two were specifically expressed in SK-N-BE(2)-C and six were specifically expressed in IMR-32; and 41 Annotated genes, of which 35 were ubiquitously expressed, four were specifically expressed in SK-N-BE(2)-C and two were specifically expressed in IMR-32. Primers were designed using primerXL[71] (Supplementary Table 2). For each gene in each sample, two qPCR reactions were performed, one on cDNA and one on RNA (to assess amplification of contaminating DNA). cDNA was produced using the iScript Advanced Kit (Bio-Rad) with a mix of random primers and oligo-dT primers on 2 µg of input RNA and a reaction volume of 20 µl. All qPCR reactions were performed in a total volume of 5 µl containing 2.5 µl of SsoAdvanced mastermix (Bio-Rad), 2 µl of forward and reverse primers (5 µM) and 0.5 µl of cDNA (10 ng µl^{-1}) or the equivalent mass of RNA. Reactions were run on a LightCycler 480 (Roche) in 384-well plates using the following thermal cycling protocol: 2 min of enzyme activation at 95.0 °C (temperature ramp rate of 4.8 °C s^{-1}), followed by 45 cycles of 5 s at 95 °C (temperature ramp rate of 4.8 °C s^{-1}) and 30 s at 60 °C (temperature ramp rate of 2.5 °C s^{-1}). For melting curve analysis, denature was performed at 95 °C for 5 s (temperature ramp rate of 4.8 °C s^{-1}), followed by cooling to 60 °C for 1 min (temperature ramp rate of 2.5 °C s^{-1}) and then heating to 95 °C at a ramp rate of 0.11 °C s^{-1} with 5 acquisitions per °C. Final cooling was performed during 3 min at 37.0 °C (temperature ramp rate of 2.5 °C s^{-1}).

Analysis of overlap between ONT reads in public datasets and RNA Atlas-only single-exon genes. *RNA sequencing libraries.* Three direct RNA sequencing libraries and one R2C2 sequencing library were used in this study. Workman et al.[72], Gleeson et al.[73] and Leger et al.[74] sequenced the transcriptome of human GM12878 cells, human SH-SY5Y neuroblastoma cells and MOLM13 cells, respectively. After basecalling, Workman et al. generated approximately 14.9 million direct RNA reads, whereas Gleeson et al. and Leger et al. generated approximately 2.7 million and 2.3 million direct reads, respectively. Furthermore, Cole et al.[75] sequenced the transcriptome of whole blood samples from humans. After basecalling and processing of the raw R2C2 reads with their C3POa pipeline, Cole et al. generated approximately 5.1 million R2C2 consensus reads.

Direct RNA sequencing libraries from Workman et al. were obtained via the GitHub repository https://github.com/nanopore-wgs-consortium/NA12878, and direct RNA sequencing libraries from Gleeson et al. and Leger et al. were obtained from the European Nucleotide Archive with dataset identifiers PRJEB39347 and PRJEB35148, respectively. The R2C2 consensus reads from Cole et al. were obtained via https://users.soe.ucsc.edu/~vollmers/PBMC_data/R2C2_reads.fa.

Data processing. To ensure that the direct reads from Workman et al., Gleeson et al. and Leger et al. were high quality, FASTQ files containing direct reads were analysed with FastQC (v0.11.8). After the assessment of read quality, the start of each direct read was trimmed by 10 bp using NanoFilt (v2.6.0)[76]. NanoFilt (v2.6.0) was also used to remove reads that were assigned a score of less than 7, because these reads are considered low quality according to the ONT guidelines, and to remove reads less than 60 bases in length, because almost all lncRNAs annotated in Ensembl are more than 60 bases. By contrast, for R2C2 consensus reads, no additional filtering was required because raw R2C2 reads were processed, and low-quality reads were removed, with the C3P0a pipeline. Direct RNA reads and R2C2 consensus reads were then mapped to the reference human genome from Ensembl (release 100, GRCh38) using minimap2 (v2.16)[77] according to the software guidelines (-ax splice -uf). Reads that could not be mapped, and reads that mapped to more than one location in the human genome, were removed using SAMtools[78].

ONT validation of RNA Atlas-only single-exon genes. Custom scripts were used to map the reads in each dataset onto the RNA Atlas-only single-exon genes. Reads in each dataset that mapped to more than one unique single-exon gene were removed. Moreover, we required that the strand of the read and of the single-exon RNA Atlas-only gene to be consistent. To ensure high-confidence validation of these RNA Atlas-only genes, we also mapped the reads in each dataset onto transcripts in Ensembl (release 100, GRCh38), and reads that mapped to both RNA Atlas-only and Ensembl transcripts were removed. Custom scripts were then used to compare the number of RNA Atlas-only single-exon genes that were validated by the reads in each dataset.

Analysis of polyadenylation status. For these analyses, we used read count expression data obtained with htseq-count[79] (v0.11.0) from TopHat BAM files. We used 291 samples for which we have expression data from both polyA and total RNA sequencing libraries. Samples 'RNA_AT294' and 'RNA_AT296' were not included in these analyses because they had a very high fraction of mitochondrial reads in polyA RNA sequencing libraries. First, we generated a list of known polyadenylated and non-polyadenylated genes based on Yang et al.[35] by selecting those genes that were annotated as either polyadenylated or non-polyadenylated in both cell lines used in that study. To normalize counts between matching polyA and total RNA sequencing libraries for differences in library size and library complexity, we calculated the mean count of the 900 most abundant known polyadenylated PCGs in both libraries and used the mean count ratio between libraries (polyA/total RNA) as a scaling factor. For most samples, this ratio was below 1 and was, thus, used to scale counts in the total RNA library. In cases where this ratio was higher than 1, we used the inverse of this ratio to scale the polyA library. Scaling was done by subsampling counts from the relevant library to obtain similar counts for polyadenylated genes in both libraries.

Polyadenylation status was determined as follows: only genes with at least ten counts in total RNA libraries were classified; otherwise, their polyadenylation status was considered as 'undetermined' ($n=3,065$). Genes with zero counts in polyA and at least ten counts in total RNA were classified as non-polyadenylated. For genes with non-zero counts in polyA and at least ten counts in total RNA libraries, a classification approach was taken, based on the $\log_2$ ratio of counts between polyA and total RNA libraries. First, a sample-specific $\log_2$ ratio cutoff was determined based on the distributions of known polyadenylated and known non-polyadenylated genes[35]. From these, only polyadenylated genes with at least ten counts in the polyA library and only non-polyadenylated genes with at least ten counts in the total RNA library were retained.

A sample-specific $\log_2$ ratio cutoff was determined by taking the value that maximizes the accuracy (number of true polyadenylated genes + number of true non-polyadenylated genes)/(total number of genes) of the classification of known genes into non-polyadenylated ($\log_2$ ratio below the cutoff) and polyadenylated ($\log_2$ ratio above the cutoff). Because the set of known polyadenylated genes is much larger than the set of known non-polyadenylated genes, we subsampled the polyadenylated genes to match the number of non-polyadenylated genes to obtain a balanced dataset. We repeated this approach 100 times and took the mean selected cutoff across iterations. We then derived a general classification for each RNA Atlas gene by taking the majority vote across samples and defining the following categories: (1) polyadenylated (number of polyadenylated samples > number of non-polyadenylated samples); (2) non-polyadenylated (number of non-polyadenylated samples > number of polyadenylated samples); and (3) bimorphic (number of polyadenylated samples = number of non-polyadenylated samples). Heat maps of gene polyadenylation across samples were plotted per biotype (PCG, lincRNA and antisense). For this, the samples were first sorted based on a normalized $\log_2$ ratio obtained by subtracting the sample specific cutoff

from the $\log_2$ ratios (to make them comparable across samples). Sorted samples were then binned in a total of 20 bins, and the mean corrected ratio for each bin was calculated and plotted in the heat map.

To select genes with changing polyadenylation status across samples, we considered genes that are expressed in at least two samples with a corrected $\log_2$ ratio below −4 and a read count of at least 100 in total RNA and expressed in at least two samples with a corrected $\log_2$ ratio above zero and read counts of at least 100 in both total RNA and polyA. This resulted in 160 genes. To get insights into the factors driving the observed changes in polyadenylation status at gene level, we analyzed changes in expression levels of individual transcripts from these genes. Specifically, we retrieved the dominant transcripts in each library from the most extreme polyadenylated sample (that is, with highest $\log_2$ normalized ratio) and the most extreme non-polyadenylated sample (that is, with lowest $\log_2$ normalized ratio). We computed the fraction of total gene expression represented by the dominant transcripts and evaluated differences in dominance and fraction of expression between the polyadenylated and non-polyadenylated samples (Supplementary Table 9). By analyzing these parameters, we defined two cases in which the variability in gene-level polyadenylation across samples can be explained by differential expression of alternatively polyadenylated isoforms: (1) those genes that present a different dominant transcript in total RNA datasets from the polyadenylated and the non-polyadenylated samples while presenting the same dominant transcript in the polyA and the total RNA datasets from the polyadenylated sample ($n=48$); or (2) those genes that present the same dominant transcript in total RNA datasets from the polyadenylated and the non-polyadenylated samples but whose fraction of total gene expression is lower in the polyadenylated sample compared to the non-polyadenylated sample. Besides, the dominant transcript for the polyadenylated sample in its polyA dataset is not the same as the dominant transcript in its total RNA dataset ($n=9$).

PolyA-minus sequencing. PolyA-minus libraries were generated for two RNA Atlas cell lines, SK-N-BE(2)-C and IMR-32, in duplicates. In brief, 500 ng of RNA was first depleted for rRNA using the Ribo-Zero Gold approach (Illumina) followed by the polyA selection procedure as implemented in the TruSeq mRNA Library Prep protocol (Illumina). Rather than discarding the polyA-minus fraction, two additional rounds of polyA selection were performed on that fraction, each time maintaining the polyA-minus fraction as input for the polyA selection step. The final polyA-minus fraction was concentrated using RNA Clean XP beads (Agencourt) before proceeding with library prep (according to the TruSeq mRNA Library Prep manual). In parallel, the polyA-plus fraction, obtained after the first polyA selection step, was also processed for library prep and sequencing. Libraries were quantified using qPCR (Kapa) and equimolarly pooled for sequencing on a NextSeq 500, high-output flow cell, paired-end sequencing, 75 cycles per read (Illumina). Sequencing reads were mapped to the hg38 reference genome (primary assembly, canonical chromosomes, repeats from RepeatMasker and Tandem Repeats Finder soft masked—http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/) with TopHat[52] v2.1.0 (Bowtie2 (ref. [53]) v2.2.6) using the --no-coverage-search option and --library-type=fr-firststrand. The Ensembl transcriptome (v86)[18] was provided to guide the mapping of reads to known transcripts first. All other parameters were set to default values. Read count expression data were then derived from the mapped reads using htseq-count[79] (v0.11.0). Genes with at least ten mean counts between replicates in either polyA-minus or polyA-plus libraries were selected, and the ratio of polyA-minus versus polyA-plus counts scaled by library size was calculated. This ratio was then compared between genes that were previously classified as polyadenylated or non-polyadenylated in the corresponding sample based on the ratio of counts from polyA and total RNA libraries.

Expression specificity. Expression specificity was computed for each RNA biotype and each sample type (cell types, cell lines and tissues) separately, using a specificity score based on the Jensen–Shannon divergence metric[36]. For PCGs, asRNAs and lincRNAs, specificity was calculated using TPM values; for miRNAs and circRNAs, we used reads per million (RPM) values. These expression metrics are not directly comparable, because they come from library preparations that capture different sets of RNAs or, in the case of circRNAs, from reads mapping to one particular location in the transcript (that is, the back-splice junction) rather than the entire transcript (as is the case for PCGs, lincRNAs and asRNAs). To directly compare expression levels between these biotypes, we quantified PCG, lincRNA and asRNA expression based on junction reads. We compared the specificity score distributions of back-splice junctions (circRNAs) and linear junctions (PCGs, lincRNAs and asRNAs) before and after correcting for differences in expression abundance between these biotypes. For this, we subsampled subsets of splice junctions from each biotype so that the subsampled distributions of maximum expression values matched a common normal-like distribution with mean equal to the median of the mean values for the different biotype distributions and standard deviation equal to the mean minus one-third of the smallest absolute extreme value (that is, minimum and maximum values of the distributions) among all biotype distributions.

Validation of tissue-specific RNAs from external datasets. Expression data from 23 tissues with matched RNA Atlas tissues were retrieved from the Tissue Atlas of the Human Protein Atlas (HPA)[39]. We selected 1,320 tissue-specific genes within the HPA dataset with an expression value of at least 5 TPM and a fold change of at least 10 between the first and the second tissue with highest expression values. The selected HPA markers were considered as cross-validated in RNA Atlas if they presented the highest expression in the same tissue. For all selected biomarkers, the $\log_2$ fold change between the expression in the matching tissue and the highest expression among the remaining 22 tissues was calculated.

Fusion genes. Fusion genes were identified with FusionCatcher[80] across all polyA sequencing samples. In each sample, fusions labeled as probable false positives and fusions known to occur in healthy samples (Supplementary Table 21, codes 0 and 1) were filtered out. Also, the fusion transcripts were required to have zero 'counts of common mapping reads'—that is, reads that map on both partners—and a minimum of four unique reads mapping on the fusion junction. Finally, within each sample, transcript fusions were collapsed at gene level—that is, if multiple junctions occurred at different joint points or reciprocally between the same pair of genes, they were counted only once—and the distribution of number of junctions per sample was compared among cell lines, cell types and tissues using two-sample Wilcoxon rank-sum tests.

Imprinting. To detect imprinting, data were first further processed according to Goovaerts et al.[81], which relies on SAMtools (v0.1.19) for initial variant calling and genotyping (and sequencing error estimation) by SeqEM (v1.0). Only variants present in dbSNP (v150) were retained, and insertions, deletions and loci corresponding with mutations from the Human Gene Mutation Database were removed. Detection of imprinting and other statistical analyses were performed in R (v3.3.2). Used filters were: coverage > 4, number of samples ≥ 75, minor allele frequency > 0.15 and estimated sequencing error rate ≤ 0.035. As outlined earlier[81], for the detection of imprinting across tissues per single-nucleotide polymorphism (SNP), a mixture model of homozygous and heterozygous samples was fit to the RNA sequencing data, with weights derived from Hardy–Weinberg equilibrium. The mixture model takes into account sequencing errors and partial imprinting. Unpublished before, also the degree of inbreeding in the underlying population is taken into account when estimating the fractions of heterozygous and homozygous loci—that is, the weights of the mixture model. The degree of inbreeding is estimated as a hyperparameter—that is, the median degree of inbreeding over all SNPs passing the quality filters (further described in Goovaerts et al.[81]), leading to an estimate of 0.102. A likelihood ratio test is used to assess whether the model supports the absence of apparently heterozygous loci (which feature, on average, a 1:1 ratio of both alleles).

This methodology was applied on the total RNA sequencing samples from 203 tissue and cell type samples, excluding cancer/cell line samples given their frequent loss of imprinting[81]. Next to using total RNA sequencing, also the polyA RNA sequencing dataset was queried (177 samples), whereas the coverage of the small RNA sequencing was too low to apply the methodology (data not shown). Other used filters were: goodness of fit > 1.2, symmetry > 0.05, median imprinting ≥ 0.8 and estimated $\hat{\imath} \geq 0.6$; for more details, see Goovaerts et al.[81]. Additionally, as we aimed to identify consistently imprinted loci, we focused on SNPs featuring a minimal difference between expected and observed heterozygous samples (based on SeqEM RNA sequencing genotyping) of 30. An FDR of 0.05 was used to call imprinting significant. We relied on RNA Atlas annotation, complemented by Ensembl annotation where relevant. In case of overlapping genes, the gene in which the SNP was located in an exonic region or UTR was selected. For additional validation, genotyping data from Illumina Human1M-Duo BeadChip for ten cell lines (HEK293T, SK-N-SH, A549, HL-60, K562, MCF-7, OVCAR-3, T-47D, JURKAT and H1 hESC) were downloaded from ENCODE for validation of imprinting in these cell lines. Note that these data had not been used during the screening phase. For virtually all sufficiently covered sample/gene combinations featuring heterozygous SNPs, at least one sample–SNP combination showed allelic expression patterns compatible with imprinting/mono-allelic expression.

Expression-based distances and differential expression analysis. t-distributed stochastic neighbor embedding (t-SNE)[82] was applied on the PCG expression data for all cell types or cancer cell lines, and the two first dimensions were used to plot a visual representation of the clustering. Weighted expression correlations (w_cor) for all pairs of samples were calculated for all RNA biotypes (using the cov.wt function in R[83]), using counts normalized by variance stabilizing transformation (VST, DESeq2)[84] as input and the average of sigmoid transformations of VST normalized counts for both samples as weights. Expression distances (expr_dist) were derived from these values as: $\text{expr\_dist} = 1 - (\text{w\_cor} + 1) / 2$.

Expression distances were compared among cell types from four biological subtypes (epithelial cells ($n=20$), endothelial cells ($n=24$), fibroblasts ($n=32$) and mesenchymal cells ($n=8$)) or cancer cell lines from 12 cancer types (B-ALL ($n=8$), breast cancer ($n=6$), central nervous system cancer ($n=6$), colon cancer ($n=7$), leukemia (n=6), melanoma ($n=9$), neuroblastoma ($n=11$), non-small cell lung cancer ($n=9$), ovarian cancer ($n=7$), prostate cancer ($n=2$), renal cancer ($n=8$) and T-ALL ($n=8$)) to measure inter- and intra-group distances. Wilcoxon rank-sum tests were performed between intra and inter distances for each group.

Differential gene expression analyses (DESeq2)[84] were performed to identify candidate miRNAs, single-exon lincRNAs and circRNAs with a significant differential expression between cell subtypes or cancer types (prostate cancer and leukemia were excluded from this analysis, for having only two cell lines belonging to the cancer type and for including a rather heterogeneous collection of cell lines, respectively). Those genes with a $\log_2$ fold change of at least 3 and a Benjamini–Hochberg-based FDR lower than 0.01 were selected, and expression data were visualized in heat maps. For circRNAs, we repeated the previous analyses on cell types and cancer cell lines by using sequential subsets of top expressed circRNAs. For this, we sorted the circRNAs based on their mean back-splice counts across samples within the sample sets used in each case and took the top 20%, 10%, 5%, 3%, 2% or 1% expressed circRNAs to calculate sample–sample distances and compared the results between subsets.

Expression estimation of exons (mRNA), introns (pre-mRNA) and their ratios from total RNA sequencing data. We sought to estimate the exonic (mRNA surrogate) and intronic (pre-mRNA surrogate) expressions of protein-coding transcripts. For each total RNA sequencing BAM file profiled in RNA Atlas, the featureCounts v1.6.0 program[85] was applied to enumerate read counts in each exon and intron regions.

Exon annotations were downloaded from the UCSC Genome Browser[57] in December 2017 (track: NCBI RefSeq; table: refGene; assembly: GRCh38). We further extended exon boundaries by 10 base pairs to prevent exonic read boundaries near the exon junctions from being considered as intronic reads[40]. After extension, regions that were within two consecutive exons of a protein-coding transcript but did not overlap with any exons of other coding and non-coding transcripts were defined as intronic. We recorded exon and intron boundaries of each protein-coding transcript.

featureCounts was run on exons and introns separately, with reads summarized at feature level—that is, single-exon or intron (argument: -f)—and only primary alignments were counted (argument: --primary). Duplicate reads were excluded from the counting process (argument: --ignoreDup). Reads mapped to multiple genes (discordant reads) or locations (multi-mapping reads) were discarded.

Counts of reads matching entire exons and introns of the same transcript were used to represent its exonic (mRNA) and intronic (pre-mRNA) abundance, respectively. We required transcripts whose $\log_2$-transformed exonic and intronic expressions, after adding a pseudo-count of 8 to their raw read counts, are at least 5—that is, 24 counts—in every RNA Atlas sample[40]. In total, 7,287 RefSeq transcripts—corresponding to 3,555 genes—were kept for analysis. To calculate exon/intron ratios (mRNA/pre-mRNA ratios, or m/p ratios for short), we used the following formula for each protein-coding transcript:

$$\text{Exon/intron ratio (or m/p ratio)} = \log_2(\text{exonic read counts} + 8) - \log_2(\text{intronic read counts} + 8)$$

Each type of expression matrix—namely, mRNA, pre-mRNA and m/p ratio—was then separately normalized using quantile normalization over multiple RNA Atlas samples using the quantilenorm routine in MATLAB; note that we used the median of the ranked values rather than the mean to perform normalization.

Experimentally verified TF and miRNA targets. We compiled 6,476 experimentally verified TF-target pairs from three sources, including HTRIdb[86] (version: 03/20/2014), TRANSFAC Professional[87] (version: February 2013) and Supplementary Table 3 from Whitfield et al.[88] For pairs deposited at HTRIdb, we included only those verified by small- and mid-scale techniques. To err on the conservative side and reduce false-positive predictions, we removed protein–DNA candidate interactions whose proteins are co-factors rather than TFs in the TRANSFAC database. The list of 4,616 verified miRNA targets with strong experimental evidence, such as western blot or reporter assay, were selected from miRecords[89] (4/27/2013), TarBase[90] version 7, TRANSFAC Professional[87] (version: February 2013), miRTarBase[91] version 4.5 and Grosswendt et al. (Supplementary Table 2 (ref. [92])). We can keep only a proportion of these interactions whose targets had sufficient exonic and intronic expression across all profiled RNA Atlas samples. In total, 2,349 and 3,304 TF and miRNA target transcripts were included for analysis. Note that each miRNA was required to express in at least 20 RNA Atlas-profiled samples.

Regulatory regions and regulator sequences. We predicted TF and lncRNA binding sites on 21,550 proximal promoters of 17,044 protein-coding genes. Each 2-kbps promoter is ranging from −1 kb to +1 kbp relative to the TSS. About one-fifth of protein-coding genes had multiple proximal promoters. The 3′ UTRs were used to predict miRNA and RBP binding sites. In total, we compiled 37,515 3′ UTRs corresponding to 17,044 protein-coding genes. The median length of all 3′ UTRs is 1,016 bps. More than half of protein-coding genes had multiple 3′ UTRs.

While identifying lncRNAs that act as miRNA, RBP and TF decoys, we searched for binding sites of these regulators throughout the whole lncRNA transcript sequence. Similarly, we identified lncRNA binding sites in promoters that match any potential binding domains of lncRNAs without consideration to their potential structures. We applied triplexator[93] version 1.3.2 and TargetScan[94] version 6.0 to predict sequence-based triple-helix (or triplex) structures and miRNA binding sites, respectively. In total, we considered sequences of 5,213 mature miRNAs, 54,502 lncRNA transcripts (or 25,468 genes) and 33,554 circRNAs. We note that, in subsequent analyses, only RNA species that were expressed in at least 50% of samples were considered. Their sequences, including asRNAs, lincRNAs and circRNAs, were extracted from the human genome assembly GRCh38 stored at the UCSC Genome Browser using twoBitToFa[57].

Prediction of TF targets. We predicted targets for 636 human TFs based on both sequence and expression evidence. First, each predicted TF target was required to have significant binding evidence from either 751 ENCODE ChIP-seq[47,95] profiles or 1,618 human TF position weight matrices (PWMs) for 108 and 636 TFs, respectively. Second, we required each TF-target pair to exhibit significant co-expression pattern across RNA Atlas-profiled samples.

ENCODE ChIP-seq datasets were profiled in 37 immortal cell lines, and more than 60% of them are in K562 ($n=121$ for 61 TFs), GM12878 ($n=113$ for 64 TFs), HepG2 ($n=97$ for 51 TFs), A549 ($n=67$ for 27 TFs) and H1-hESC ($n=62$ for 36 TFs). More than one-third of TFs had at least two replicates in the same cell line. Human TF PWMs were collected from five sources, including motifs annotated in Factorbook[96] (see Supplementary Table 2 in their paper; $n=86$ for 76 TFs), motifs of quality A-D in HOCOMOCO version 9 (ref. [97]) ($n=427$ for 395 TFs), high-confidence motifs in human TF[98] (see Supplementary Table 3 in their paper; $n=651$ for 357 TFs), JASPAR[99] v5_alpha ($n=103$ for 99 TFs) and SwissRegulon[100] downloaded on 18 March 2014 ($n=351$ for 331 TFs). To avoid matrix entries of value 0, a pseudo-count of 1 was added to each entry before calculating the relative occurrence frequencies of nucleotides at each position.

We interrogated each of 21,550 proximal promoters to see if there was a significant ChIP-seq peak ($q < 1\times10^{-10}$) or PWM-based binding site ($P<1\times10^{-5}$). The significance of motif scores on either the forward or reverse strand of the proximal promoters were compared to 5′ flanking regions of length 2 kbps of their cognate proximal promoters using the CREAD[101,102] package. Binding site evidence across multiple promoters associated with the same gene were aggregated to produce gene-level binding evidence. For any PCG that satisfied this sequence-based constraint, we further required significant distance correlation (dCor)[103] at $P<1\times10^{-9}$, as calculated using expression profiles of their regulating TFs and cognate protein-coding targets profiled in RNA Atlas. Note that only TFs and target genes of non-zero median absolute deviation (MAD) score were included for analysis. We applied permutation testing to estimate the significance of dCor by shuffling TFs' expression 100,000 times and then calculated the randomized dCor values. These values were used to fit parameters for a generalized extreme value (GEV) distribution using the MATLAB gevfit routine to obtain a non-parametric P value lower than 1×10^{-5} from the cumulative density of the resulting GEV distribution. For TF targets passed, both sequence and expression constraints were investigated for transcriptional lncRNA modulation. We predicted 105,029 interactions between TFs and their protein-coding targets significantly modulated by lncRNAs. Moreover, 102,338 TF-target interactions had target transcripts of adequate exonic and intronic coverage to compute m/p ratio profiles.

Prediction of miRNA and RBP targets. We predicted targets of both types of post-transcriptional regulators through a two-step approach by requiring both sequence- and expression-based evidence. Specifically, 3′ UTRs of protein-coding transcripts and whole lncRNA transcripts were scanned for miRNA binding sites conserved across species (context score <-0.2) by TargetScan[94] version 6.0 and significant RBP binding peaks at $P<1\times10^{-10}$. ENCODE eCLIP[104] datasets for 115 RBPs profiled in two human cancer cell lines—that is, K562 and HepG2—were downloaded from the UCSC Genome Browser. Among them, 66 and 49 RBPs were available in either one or two cell lines, respectively. Each RBP-cell line pair was performed in duplicate. Binding site evidence across multiple 3′ UTRs associated with the same gene were aggregated to produce gene-level binding evidence. We then asked if any pair of genes, either coding or non-coding, shared a significantly large common regulator program at adjusted $P_{FET} < 0.01$. For each qualified gene pair and their common regulators, we measured if correlation changes between a common miRNA/RBP and any of these two genes had evidence for being modulated by lncRNA expressions using delta dCor; see the 'lncRNA target predictions using LongHorn' section below. A pair of regulator-target significantly modulated by at least one lncRNA at $P<0.05$ was finally selected. miRNA/RBP targets that passed both sequence and expression constraints were investigated for post-transcriptional lncRNA modulation. In total, 66,623 predicted interactions between miRNAs and their protein-coding targets were significantly modulated by lncRNAs, and, among them, 46,126 miRNA-target transcripts had adequate exonic, intronic and m/p ratio reads and could be included in further analyses to compare correlations of regulator and target mRNA and pre-mRNA expression profiles. Note that, similarly to experimentally verified miRNA targets, each miRNA, including both miRBase-annotated and miRDeep2-predicted miRNAs, was required to be expressed in at least 20 RNA Atlas-profiled samples. RBPs were required to have a non-zero MAD score.

dCor. For each experimentally verified and LongHorn-inferred TF and miRNA target, we applied dCor[103] to measure co-expression patterns between a regulator—namely, a TF, RBP or miRNA—and its target using the target's mRNA (exonic),

pre-mRNA (intronic) or m/p ratio profiles. dCor is able to capture non-linear relationships between two variables, which is a common scenario in the biological world. The dCor value is always non-negative, and dCor = 0 means that two variables are completely independent[103].

lncRNA target predictions using LongHorn. LongHorn[43] predicts lncRNA interactions through respectively integrating statistical evidence from modulation of transcriptional and post-transcriptional regulation by TFs, miRNAs and RBPs. Transcriptional lncRNAs can physically interact with either proximal promoters (TR:Guide and TR:Co-factor) or TFs (TR:Guide and TR:Decoy) to alter their target pre-mRNA abundance. Guide and Co-factor lncRNAs form RNA–DNA triple-helix structures (triplex, for short) with proximal promoters[93]. The former recruit TFs to bind their transcript's promoter regions and synergistically enhance the transcription rates of their targets. The latter can either activate or inhibit the regulatory activities of TFs, which have their own binding sites on the same promoter. Decoy lncRNA, in turn, influences target pre-mRNA transcription and expression by altering the amount of TF and miRNA/RBP molecules available to target proximal promoters and 3′ UTRs. In our model, lncRNAs can regulate target pre-mRNA (TR:Decoy) and mRNA (PTR:Decoy) abundances in nucleus and cytoplasm. Note that, while predicting lncRNAs acting as guides or decoys of TFs, we used PWMs to scan TF binding sites from on lncRNA transcript sequences.

LongHorn reverse engineers transcriptional and post-transcriptional interactions on a genome-wide basis at first; see the sections 'Prediction of TF targets' and 'Prediction of miRNA and RBP targets' above. To estimate the significance of modulation, we calculated delta dCor for each triplet, consisting of a lncRNA, a regulator (TF/miRNA/RBP) and a protein-coding target. According to the lncRNA expression in each triplet, we partitioned RNA Atlas samples into four quartiles, from the lowest to the highest, and required this lncRNA to satisfy two constraints, including (1) it was not correlated with the regulator ($P > 0.1$) (independence constraint) and (2) its expression fold change was more than 2× between the fourth and the first quartiles of the samples (range constraint). Then, comparing the first and the fourth sample quartile, we required a non-parametric $P < 0.05$ for the delta dCor between the regulator and the target against a bootstrapping-based null hypothesis. For significant triplets that are associated with the same lncRNA-target pair, we integrated their P values across their common regulators using either the Fisher method[105] (transcriptional) or the weighted Brown method[106] (post-transcriptional). While combining significance from multiple tests, the Brown method takes into account miRNAs and RBPs in the same genomic cluster, which are often co-expressed, to avoid inflating the integrated P values. TargetScan context scores were used to sort predicted miRNA binding sites from lowest to highest; we then used their percentile ranks as weights for integrating P values of significant triplets. After integration, we set a cutoff of adjusted $P < 0.01$ for significant lncRNA-target pairs. Note that both lncRNAs and their protein-coding targets were required to have a non-zero MAD score across profiled RNA Atlas samples. We also note that correlations between pre-mRNA and mRNA expression estimates (Fig. 5b) were maintained for both lncRNA targets and the randomized interactions used in permutation testing.

Validation of LongHorn-inferred target prediction. To test the accuracy of LongHorn target inference on a macro scale, we evaluated LongHorn predictions using data generated by the FANTOM6 consortium (Ramilowski et al.)[51]. FANTOM6 used antisense oligonucleotides (ASOs) to knock down 285 lncRNAs in HDF—a human primary dermal fibroblast cell line—and selected 154 of these lncRNAs for gene expression profiling by RNA sequencing based on their high knockdown efficiencies. We systematically tested if downregulation of these lncRNAs by ASOs lead to the dysregulation of their LongHorn-inferred RNA Atlas targets, albeit in HDF cells only. We chose to focus our analysis on 24 of the tested lncRNAs that had at least 20 LongHorn-inferred targets in RNA Atlas cells and tissues. Each of these 24 lncRNAs was associated with a list of dysregulated genes that were expressed in RNA sequencing (TPM > 0) and significantly upregulated or downregulated after transfection with ASOs at $P < 0.05$, as estimated by Ramilowski et al.[51]. For each of these 24 lncRNAs, we collected genes that were expressed (TPM > 0) and significantly upregulated or downregulated after lncRNA targeting at $P < 0.05$ in the FANTOM6 data as estimated by Ramilowski et al.[51]. We then tested the significance of the overlap—namely, genes that were predicted to be lncRNA targets in RNA Atlas and were also dysregulated after ASO-targeting lncRNA in FANTOM6—between our predictions and this set using Fisher's exact test. The results suggested that, for 20/24 (83%) and 15/24 (63%) of the lncRNAs, LongHorn-inferred RNA Atlas targets were more likely to be dysregulated (odds ratio > 1) and significantly dysregulated at $P < 0.05$ by Fisher's exact test after the downregulation of the corresponding lncRNA, respectively. Odds ratios were computed using the following formula:

$$\text{odds ratio} = \frac{\text{no. of predicted targets that were dysregulated/no. of predicted targets}}{\text{no. of profiled genes that were dysregulated/no. of profiled genes}}$$

Complete accounting is provided in Supplementary Table 26. An example gene set enrichment analysis (GSEA) plot for the enrichment of LongHorn-inferred EMX2OS targets is given in Supplementary Fig. 26.

To further test LongHorn-predicted RNA Atlas targets using unpublished data, we chose to target the lncRNA MALAT1 by CRISPR interference (CRISPRi) (see experimental details below) and profiled gene expression by RNA sequencing in HEK293 cells. MALAT1 is a highly expressed lncRNA that had been successfully targeted by CRISPRi in past works and was predicted to target 723 genes in RNA Atlas cells and tissues. We compared the effects of MALAT1-targeting single guide RNAs (sgRNAs) to two controls (NC1 and NC2) in Supplementary Fig. 27. In total, >2,000 out of >17,200 profiled genes in these assays were identified as differentially expressed at $P < 0.05$ relative to each control independently. Of the 723 LongHorn-predicted MALAT1 targets, 140 and 190 were significantly dysregulated at $P < 0.05$ relative to NC1 and NC2, respectively, corresponding to $P < 5 \times 10^{-10}$ and $P < 4 \times 10^{-16}$ by Fisher's exact test, respectively. Significantly upregulated or downregulated genes were selected by comparing their expressions, measured in counts per million (CPM), in samples transfected with MALAT1-targeting sgRNAs (eight replicates) over each non-targeting control (NC1 or NC2 with 12 replicates each) independently based on two-tailed Student's t-tests. Normalized profiles are given in Supplementary Table 27.

Of course, dysregulated genes after lncRNA targeting might not be direct targets of lncRNAs, and true lncRNA targets might not have been identified as dysregulated because of biological or technical reasons. However, when a sufficient number of true targets is tested, we expect to observe enrichments in their dysregulation. Here, the analysis of FANTOM6 data suggested that predicted target sets of most tested lncRNAs were significantly enriched in dysregulated genes, and the analysis of our assays targeting MALAT1 confirmed this observation.

CRISPRi-mediated transcriptional silencing of lncRNA MALAT1. MALAT1 was silenced in HEK293T cells using the CRISPRi method. First, nuclease-deficient dCas9-KRAB-MeCP21 (Addgene plasmid no. 110821) was stably introduced in HEK293T cells using the piggy-transposase system (System Biosciences, cat. no. PB210PA-1) according to the manufacturer's recommendations. dCas9-KRAB-MeCP2-positive HEK293T cells were selected using 10 μg ml^{-1} of blasticidin. Next, sgRNAs targeting a window of 300 bp upstream and downstream of the MALAT1 TSS were designed using the DeskGen tool[107]. Two control sgRNAs (GAACGACTAGTTAGGCGTGTA and GTGCGATGGGGGGGTGGGTAGC) were selected from Horlbeck et al.[108] and Gilbert et al.[109] sgRNA sequences were then amended with 5′ and 3′ appendixes as specified by the Guide-it sgRNA In Vitro Transcription Kit (Takara Bio, cat. nos. 632638, 632639, 632635, 632636 and 632637), and single-stranded DNA oligos were purchased from IDT. dsDNA in vitro transcription template was generated with the Guide-it kit according to the manufacturer's instructions. In vitro transcription was performed at 37 °C for 4 h. Finally, 12,000 cells per well were seeded in 96-well plates (Corning, cat. no. 3596) in 180 μl of RPMI cell culture medium. Twenty-four hours after seeding, sgRNAs were transfected with lipofectamine reagent CRISPRMAX (Invitrogen, cat. no. CMAX00003) to a final concentration of 0.5 ng μl^{-1} in 200 μl. Seventy-two hours after the transfection, cells were lysed with SingleShot lysis buffer (Bio-Rad, cat. no. 172-5081). Quantseq RNAseq library preparation (Lexogen) was performed according to the manufacturer's protocol using 5 μl of cell lysate as input. Libraries were quantified by qPCR, pooled and sequenced on a NextSeq 500 (Illumina). FASTQ files were processed using an in-house RNA sequencing pipeline. First, FastQC (v0.11.8) was used for data quality control, after which adapter sequences, polyA readthrough and low-quality reads were removed via bbduk[110] (BBMap v38.26). Next, reads were mapped with STAR[111] (v2.6.0c) against the hg38 reference genome, and gene counts were determined via HTSeq[79] (v.0.11.0).

The RNA Atlas lncRNA-target set. We defined the RNA Atlas lncRNA-target set by evaluating the deviations of the dCor between each mediating regulator and the target pre-mRNA or m/p ratio expression profiles. Specifically, LongHorn predicted lncRNA-target pairs by providing (1) target identities, (2) regulation model including transcriptional or post-transcriptional interactions and (3) the list of regulators that are predicted to mediate the interactions. Each lncRNA-target interaction was associated with two distributions of the dCor: one was using pre-mRNA, and the other was using m/p ratio expression profiles of the target. We anticipated that post-transcriptional targets of lncRNAs will have higher dCor values with their mediating regulators while using m/p ratio than pre-mRNA expression profiles of targets. However, for transcriptional targets of lncRNAs, the relationships were expected to be in the opposite direction—with m/p ratio less correlated than the pre-mRNA expression profiles. We evaluated differences between these two distributions by either the paired Student's t-test (parametric and one tailed) or permutation testing (non-parametric) and required RNA Atlas lncRNA-target interactions to be significant at $P < 0.05$ in non-parametric tests after adjusted for multiple comparisons using the Benjamini–Hochberg procedure.

To estimate the non-parametric significance of lncRNA-target interactions, we first partitioned the data into four quantiles with low to high regulator expression variability based on MAD. We randomly selected 10,000 out of all possible regulator-target pairs in each quantile to form a randomized set of 40,000 pairs. All transcript combinations per pair were compiled. Then, for

each predicted lncRNA-target interaction that was mediated by *N* regulators, we selected *N* regulator-transcript pairs at random to match quantiles and calculated a randomized Student's *t*-score from the paired Student's *t*-test by comparing the two distributions of the dCor values between regulators and target pre-mRNA or m/p ratio expression profiles. This selection criteria ensured that the only variable perturbed is the lncRNA-target association: regulators have similar expression and are predicted to regulate the target. The process was repeated 100 times to compute the null distribution of Student's *t*-scores with a minimum attainable *P* value of 0.01. To adjust for multiple comparisons, we applied the Benjamini–Hochberg procedure to control the FDR. Each pair in the RNA Atlas lncRNA-target set was a predicted interaction that had a computed Student's *t*-score at Benjamini–Hochberg-adjusted $P < 0.05$ after comparing with the null distribution. Note that the regulators were assigned based on the regulation model—that is, TFs and miRNAs were for transcriptional and post-transcriptional lncRNA-target interactions, respectively.

Transcriptional and post-transcriptional specialists. To identify lncRNA specialists with unusual number of transcriptional or post-transcriptional interactions, we first normalize the size of LongHorn-inferred transcriptional and post-transcriptional interactomes to obtain a scaling ratio (σ). In RNA Atlas, LongHorn predicted 480,333 and 18,051 targets whose regulators' activities were transcriptionally and post-transcriptionally modulated by lncRNA, respectively. Namely, the scaling ratio is 26.610 for transcriptional interactions. For each RNA Atlas-profiled lncRNA, including asRNAs, lincRNAs and circRNAs, with at least ten predicted targets, we calculated the adjusted fold change (adjFC) using the following formula to determine if it is a specialist:

$$\text{adjFC} = \frac{\text{no. of transcriptional interactions}/\sigma}{\text{no. of post-transcriptional interactions}}$$

The lncRNA is a transcriptional or post-transcriptional specialist if the adjFC is larger than 2 or smaller than 0.5, respectively. By calculating this statistic, we revealed lncRNAs extensively involved in pathways of transcriptional or post-transcriptional gene regulation.

Hallmark GSEA. We sought to study if lncRNAs regulate key biological pathways through searching for significant overlaps between their LongHorn-inferred targets and 50 MSigDB hallmark gene sets[50], which can be broken into eight basic categories, including Cellular Component, Development, DNA Damage, Immune, Metabolic, Proliferation and Signaling Pathways. We calculated *P* values for the significance of overlap using Fisher's exact test and adjusted for multiple comparisons based on Bonferroni correction. For each lncRNA-gene set pair, an adjusted *P* value lower than 0.01 was considered to be a significant association.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All types of RNA entities can be readily explored via the online R2: Genomics Analysis and Visualization Platform (http://r2.amc.nl) and via a dedicated accessible portal (http://r2platform.com/rna_atlas). This portal includes genome browser profiles for the total RNA as well as polyA tracks for all samples. All samples can also be used for correlations, differential signals and many more analyses. In addition, the LongHorn results, described in this manuscript, can be explored.

The raw data (FASTQ files) and processed expression measurement tables from all RNA biotypes across samples have been deposited in the National Center for Biotechnology Information's Gene Expression Omnibus (GEO) and are accessible through GEO series accession number GSE138734.

Code availability

Computer code used to generate the results presented in this manuscript is available at https://github.com/llorenzi90/RNA_Atlas.

References

52. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* **25**, 1105–1111 (2009).
53. Langmead Ben, StevenS. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2013).
54. Pertea, M. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
55. Trapnell, C. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
56. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
57. Haeussler, M. et al. The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.* **47**, D853–D858 (2019).
58. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
59. Cobos, F. A. et al. Zipper plot: visualizing transcriptional activity of genomic regions. *BMC Bioinformatics* **18**, 231 (2017).
60. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Res.* **9**, ISCB Comm J-304 (2020).
61. Hubisz, M. J., Pollard, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief Bioinform.* **12**, 41–51 (2011).
62. Zerbino, D. R. et al. Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
63. Vizcaíno, J. A. et al. The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* **41**, D1063–D1069 (2012).
64. Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24**, 2534–2536 (2008).
65. Silva, C. A. S. et al. Data-driven rescoring of metabolite annotations significantly improves sensitivity. *Anal. Chem.* **90**, 11636–11642 (2018).
66. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **46**, 2699 (2018).
67. Zhang, X. O. Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Res.* **26**, 1277–1287 (2016).
68. Gordon, A., Hannon, G. J. & Gordon. FASTX-Toolkit. http://hannonlab.cshl.edu/fastx_toolkit/
69. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
70. Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33**, 3645–3647 (2017).
71. Lefever, S. et al. High-throughput PCR assay design for targeted resequencing using primerXL. *BMC Bioinformatics* **18**, 400 (2017).
72. Workman, R. E. et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* **16**, 1297–1305 (2019).
73. Gleeson, J., Lane, T. A., Harrison, P. J., Haerty, W. & Clark, M. B. Nanopore direct RNA sequencing detects differential expression between human cell populations. Preprint at *bioRxiv* https://doi.org/10.1101/2020.08.02.232785 (2020).
74. Leger, A. et al. RNA modifications detection by comparative nanopore direct RNA sequencing. Preprint at *bioRxiv* https://doi.org/10.1101/843136 (2019).
75. Cole, C., Byrne, A., Adams, M., Volden, R. & Vollmers, C. Complete characterization of the human immune cell transcriptome using accurate full-length cDNA sequencing. *Genome Res.* **30**, 589–601 (2020).
76. De Coster, W., D'hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).
77. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3091–3100 (2018).
78. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* **25**, 2078–2079 (2009).
79. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics.* **31**, 166–169 (2015).
80. Nicorici, D. et al. FusionCatcher—a tool for finding somatic fusion genes in paired-end RNA-sequencing data. Preprint at https://doi.org/10.1101/011650 (2014).
81. Goovaerts, T. et al. A comprehensive overview of genomic imprinting in breast and its deregulation in cancer. *Nat. Commun.* **9**, 4120 (2018).
82. Van Der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
83. R Development Core Team. *R: A Language and Environment for Statistical Computing*. http://www.R-project.org (R Foundation for Statistical Computing, 2011).
84. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
85. Liao Y, Smyth GK, Shi W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btt656 (2014).
86. Bovolenta, L. A., Acencio, M. L. & Lemke, N. HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics* **13**, 405 (2012).
87. Matys, V. et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110 (2006).
88. Whitfield, T. W. et al. Functional analysis of transcription factor binding sites in human promoters. *Genome Biol.* **13**, R50 (2012).
89. Xiao, F. et al. miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.* **37**, D105–D110 (2009).
90. Vlachos, I. S. et al. DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Res.* **43**, D153–D159 (2015).

91. Da, H. S. et al. MiRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res*. **42**, D78–D85 (2014).
92. Grosswendt, S. et al. Unambiguous identification of miRNA: target site interactions by different types of ligation reactions. *Mol. Cell* **54**, 1042–1054 (2014).
93. Buske, F. A., Bauer, D. C., Mattick, J. S. & Bailey, T. L. Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic data. *Genome Res*. **22**, 1372–1381 (2012).
94. Garcia, D. M. et al. Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nat. Struct. Mol. Biol*. **18**, 1139–1146 (2010).
95. Landt, S. G. et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res*. **22**, 1813–1831 (2012).
96. Wang, J. et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res*. **22**, 1798–1812 (2012).
97. Kulakovskiy, I. V. et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-seq analysis. *Nucleic Acids Res*. **46**, D252–D259 (2018).
98. Lambert, S. A. et al. The human transcription factors. *Cell* **172**, 650–665 (2018).
99. Khan, A. et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res*. **46**, D260–D266 (2018).
100. Pachkov, M., Balwierz, P. J., Arnold, P., Ozonov, E. & Van Nimwegen, E. SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Res*. **41**, D214–D220 (2013).
101. Smith, A. D., Sumazin, P., Xuan, Z. & Zhang, M. Q. DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc. Natl Acad. Sci. USA* **103**, 6275–6280 (2006).
102. Smith, A. D., Sumazin, P., Das, D. & Zhang, M. Q. Mining ChIP-chip data for transcription factor and cofactor binding sites. Bioinformatics 21 (Suppl. 1), i403–i412 (2005).
103. Sz,kely, G. J., Rizzo, M. L. & Bakirov, N. K. Measuring and testing dependence by correlation of distances. *Ann. Stat*. **35**, 2769–2794 (2007).
104. Van Nostrand, E. L. et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods* **13**, 504–514 (2016).
105. Lury, D. A. & Fisher, R. A. Statistical methods for research workers. J. R. Stat. Soc. Ser. D Statistician https://doi.org/10.2307/2986695 (1972).
106. Brown, M. B. 400: a method for combining non-independent, one-sided tests of significance. *Biometrics* **31**, 987–992 (1975).
107. Hough, S. H., Ajetunmobi, A., Brody, L., Humphryes-Kirilov, N. & Perello, E. Desktop Genetics. *Per. Med*. **13**, 517–521 (2016).
108. Horlbeck, M. A. et al. Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *eLife* **5**, e19760 (2016).
109. Gilbert, L. A. et al. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442–451 (2013).
110. Bushnell, B. BBMap. https://sourceforge.net/projects/bbmap/
111. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

Acknowledgements

F.A.C. is supported by a Special Research Fund (BOF) scholarship of Ghent University (BOF.DOC.2017.0026.01). R.C. is supported by the Fonds Wetenschappelijk Onderzoek (11Y6218N). T.-W.C. is supported by grants from the Ministry of Science and Technology, Taiwan (MOST-109-2311-B-009 –002). A.U. is supported by research funding from the National Health and Medical Research Council (Australia) and the Leukemia & Lymphoma Society, the Leukemia Foundation and the Snowdome Foundation. G.A. is supported by a postgraduate scholarship from the Translational Cancer Research Network. M.R.W. and N.P.D. acknowledge support from the National Collaborative Research Infrastructure Strategy program, administered by Bioplatforms Australia. We thank N. Yigit, A. Barr, S. Pathak, L. Way and A. Mai for their contributions in library preparation and A. Yunghans, E. Jaeger and A. Moshrefi for their assistance in library organization and sequencing/tracking/data management. This project was funded by the European Union's Horizon 2020 Research and Innovation Programme under grant agreements 668858 and 826121 to P.M., P.S. and J. Koster and the Concerted Research Action of Ghent University (BOF/GOA 01G00819) to P.M. and K.B.

Author contributions

P.M., J.V. and P.S. conceived the idea and designed and supervised the project. L.L. and H.-S.C. contributed to the implementation and design of most bioinformatic analyses. L.L performed most of the raw sequencing data processing, transcriptome assembly and filtering, polyadenylation classification and most of the presented analyses for quality assessment and characterization of the generated transcriptome. H.-S.C., T.-W.C. and P.S. performed the analyses related to prediction and validation of regulatory interactions mediated by ncRNAs. F.A.C. and K.D.P. performed the analyses to select the RNA Atlas genes and contributed to quality validation of the transcriptome. S.G., S.K. and G.P.S. generated and sequenced the polyA and total RNA libraries. P.-J.V. performed the evaluation of coding potential, analyses of mass spectrometry data, alignment of candidate protein sequences to other animal proteins via BLASTp and analysis of conservation with chimpanzee. R.C. and Y. S. contributed to the analyses of RNA biotype expression and sample ontology associations. J.N. performed the polyA-minus sequencing and the qPCR experiments. K. Vanderheyden and J.N. generated and sequenced the small RNA libraries. J.A. implemented the identification of miRNAs and sequence motif analysis. S.L. designed the primers for the qPCR experiments and contributed to the graphic design of schematic figures. A.P.T. performed the analysis of overlap between ONT reads in public datasets and RNA Atlas-only single-exon genes. E.J.B., W.T. and F.G. performed the experiments of CRISPRi-mediated transcriptional silencing of lncRNA MALAT1. M.V. generated the integrated circRNA reference dataset used for comparisons with RNA Atlas circRNAs. T.G. and T.D.M. performed the imprinting analyses. T.B.H. and J. Kjems implemented the circRNA identification workflow. N.N. developed the polyA-minus sequencing protocol. T.T., K. Vermaelen and K.R.B. provided immune system-related cell lines and cell types. N.P.D., G.A., M.R.W. and A.U. performed analyses and annotation of circRNAs and contributed to the analysis of ONT reads in public datasets. J. Koster developed dedicated tools to analyze RNA Atlas data and results and implemented them in a dedicated RNA Atlas datascope in the online portal R2. P.M. led the writing of the manuscript in collaboration with L.L., H.-S.C. and P.S. L.L., H.-S.C., G.P.S., J.V., P.S. and P.M. contributed to the conceptualization, interpretation and discussion of results. All authors commented on the manuscript and contributed to the presentation of the data and results. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-021-00936-1.

Correspondence and requests for materials should be addressed to P.S. or P.M.

Peer review information *Nature Biotechnology* thanks **Steven Salzberg** and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Pieter Mestdagh

Last updated by author(s): Apr 22, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Fastq files from the sequencing data generated for this study were first deposited and downloaded from Illumina BaseSpace. Public datasets used in this study were downloaded from the corresponding public repositories which are described and cited in the Main or Methods sections.
Data analysis	All code used to analyse the data is described in the Methods section. Custom code is available upon request and will be made available through https://github.com/llorenzi90/RNA_Atlas

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All types of RNA entities can be readily explored via the online R2: Genomics analysis and visualization platform (http://r2.amc.nl), and via a dedicated accessible portal (http://r2platform.com/rna_atlas). This portal includes Genome browser profiles for the total RNA as well as polyA tracks for all samples. All samples can also be used for correlations, differential signals and many more analyses. In addition, the LongHorn results, described in this manuscript can be explored.
The raw data (fastq files) and processed expression measurement tables from all RNA biotypes across samples have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE138734 (ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE138734).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Because of the extension of the sample set and sequencing methods applied in this study, it was not feasible to include replicates for the profiled samples. However, when possible, we exploit similarity of samples based on ontology relationships to assess reproducibility.
Data exclusions	Data exclusion from particular analyses and the rationale behind it are described in Methods.
Replication	Because of the extension of the sample set and sequencing methods applied in this study, it was not feasible to include replicates for the profiled samples. However, when possible, we exploit similarity of samples based on ontology relationships to assess reproducibility.
Randomization	This is not relevant to our study
Blinding	This is not relevant to our study

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	cell lines were obtained from NCI60 or the Center for Medical Genetics Ghent (CMGG), Ghent Univeristy
Authentication	cell lines were authenticated prior to RNA isolation
Mycoplasma contamination	cell lines were tested and found to be negative for Mycoplasma contamination
Commonly misidentified lines (See ICLAC register)	MDA-MB-435, SNB-19