A conserved abundant cytoplasmic long noncoding RNA modulates repression of mRNAs by PUM2 in human cells

Ailone Tichon[1], Noa Gil[1], Tal Havkin Solomon [2], Doron Lemze[3], Shalev Itzkovitz[3], Noam Stern-Ginossar[2], Igor Ulitsky[1,*]

[1] Department of Biological Regulation, [2] Department of Molecular Genetics,

[3] Department of Molecular Cell Biology,

Weizmann Institute of Science, Rehovot 76100, Israel

* To whom correspondence should be addressed: igor.ulitsky@weizmann.ac.il

Abstract

Thousands of long noncoding RNA (lncRNA) genes are encoded in the human genome, and hundreds of them are evolutionary conserved, but their functions and modes of action remain largely obscure. Particularly enigmatic lncRNAs are those that are exported to the cytoplasm. We identified and characterized an abundant and highly conserved cytoplasmic lncRNA, which we denote Pumilio2-binding long intervening noncoding RNA (PUBLINC). Most of the sequence of this lncRNA is comprised of repetitive units that together contain at least 17 functional binding sites for PUM2, one of the two Pumilio homologs in mammals. Through binding to PUM2, PUBLINC modulates the mRNA levels and translation of PUM2 targets, enriched for genes involved in chromosome segregation during cell division. Our results suggest that some cytoplasmic lncRNAs function by modulating the activities of RNA binding proteins, an activity which positions them at key junctions of cellular signaling pathways.

Introduction

Genomic studies conducted over the past 15 years have uncovered the intriguing complexity of the transcriptome and the existence of tens of thousands of long noncoding RNA (lncRNA) genes in the human genome, which are processed similarly to mRNAs but appear not to give rise to functional proteins[1]. While some lncRNA genes overlap other genes and may be related to their biology, many do not, and these are referred to as long intervening noncoding RNAs, or lincRNAs. An increasing number of lncRNAs are implicated in a variety of cellular functions, and many are differentially expressed or otherwise altered in various instances of human

disease[2]; therefore, there is an increasing need to decipher their modes of action. Mechanistically, most lncRNAs remain poorly characterized, and the few well-studied examples consist of lncRNAs that act in the nucleus to regulate the activity of loci found *in cis* to their sites of transcription[3]. These include the XIST lncRNA, a key component of the X-inactivation pathway, and lncRNAs that are instrumental for imprinting processes, such as AIRN[4]. However, a major portion of lncRNAs are exported to the cytoplasm: indeed, some estimates based on sequencing of RNA from various cellular compartments suggest that most well-expressed lncRNAs are in fact predominantly cytoplasmic[1].

The functional importance and modes of action of cytoplasmic lncRNAs remain particularly poorly understood. Some lncRNAs that are transcribed from regions overlapping the start codons of protein-coding genes in the antisense orientation can bind to and modulate the translation of those overlapping mRNAs[5], and others have been proposed to pair with target genes through shared transposable elements found in opposing orientations[6]. Two lncRNAs that are spliced into circular forms were shown to act in the cytoplasm by binding Argonaute proteins (in one case, through ~70 binding sites for a miR-7 microRNA[7]) and act as sponges that modulate microRNA-mediated repression[7,8]. Such examples are probably rare, as few circRNAs and few lncRNAs contain multiple canonical microRNA binding sites (ref[9] and IU, unpublished results). It is not clear whether other cytoplasmic lncRNAs can act as decoys for additional RNA-binding proteins through a similar mechanism of offering abundant binding sites for the factors.

The Pumilio family consists of highly conserved proteins that serve as regulators of expression and translation of mRNAs that contain the Pumilio recognition element (PRE) in their 3' UTRs[101]. Pumilio proteins are members of the PUF family of proteins that is conserved from yeast to animals and plants, whose members repress gene expression either by recruiting 3' deadenylation factors and antagonizing translation induction by the poly(A) binding protein[11], or by destabilizing the 5' cap-binding complex. The drosophila Pumillio protein is essential for proper embryogenesis, establishment of the posterior anterior gradient in the early embryo, and stem cell maintenance. Related roles were observed in other invertebrates[10], and additional potential functions were reported in neuronal cells[12]. There are two Pumilio

proteins in humans, PUM1 and PUM2[10], which exhibit 91% similarity in their RNA binding domains, and which were reported to regulate a highly overlapping but not identical set of targets in HeLa cells[13]. Mammalian Pumilio proteins have been suggested to be functionally important in neuronal activity[14], ERK signaling[15], germ cell development[16], and stress response[14]. Therefore, modulation of PUM2 regulation is expected to have a significant impact on a variety of crucial biological processes.

Here, we characterize a previously annotated yet obscure lncRNA, which we denote Pumilio2-binding long intervening noncoding RNA (PUBLINC). PUBLINC is an abundant lncRNA with highly expressed sequence homologs found throughout plancental mammals. PUBLINC contains at least 17 functional binding sites for PUM2. By perturbing PUBLINC levels in osteosarcoma U2OS cells, we show that PUBLINC modulates the mRNA abundance of PUM2 targets, in particular those involved in mitotic progression. Further, using a luciferase reporter system we show that this modulation depends on the canonical PUM2 binding sites.

Results

PUBLINC is a highly abundant cytoplasmic lncRNAs conserved in mammals

In our studies of mammalian lncRNA conservation we identified a conserved and abundant lincRNA currently annotated as LINC00657 in human and 2900097C17Rik in mouse, which we denote here as Pumilio2-binding long intervening noncoding RNA (PUBLINC). PUBLINC produces a 5.3 Kb transcript that does not overlap other genes (**Figure 1A**), starts from a single strong promoter overlapping a CpG island, terminates with a single major canonical poly(A) site (**Figure 1B**), but unlike most long RNAs is unspliced. Similar transcripts with substantial sequence homology can be seen in EST and RNA-seq data from mouse, rat, rabbit, dog, cow, and elephant. PUBLINC does not appear to be present in opossum, where a syntenic region can be unambiguously identified based on both flanking genes but no evidence exists for a transcribed gene, and no homologs could be found in more basal vertebrates. PUBLINC is ubiquitously expressed across tissues and cell lines in human, mouse, and dog, with comparable levels across most embryonic and adult tissues (**Supplementary Figure 1)** with the exception of neuronal tissues, where PUBLINC is more highly expressed. In the most comprehensive dataset of gene expression in

normal human tissues, compiled by the GTEX project (http://www.gtexportal.org/), the ten tissues with the highest PUBLINC expression all correspond to different regions of the brain (highest level in the frontal cortex with a reads per kilobase per million reads (RPKM) score of 142), with levels in other tissues varying between an RPKM of 78 (pituitary) to 27 (pancreas). Comparable levels were also observed across ENCODE cell lines, with the highest expression in the neuroblastoma SK-N-SH cells (**Figure 1D**). The high expression levels of this lncRNA in the germ cells have probably contributed to the large number of closely related pseudogenes of PUBLINC found throughout mammalian genomes. There are four pseudogenes in human that share >90% homology with PUBLINC over >4 Kb, but they do not appear to be expressed, with the notable exception of a lincRNA transcript HCG11, which is expressed in a variety of tissues but at levels ~20-times lower than PUBLINC (based on GTEX and ENCODE data, **Figure 1D**). Due to this difference in expression levels we assume that while most of the experimental methods we used are not able to distinguish between PUBLINC and HCG11, the described effects likely stem from the PUBLINC locus and not from HCG11. Using single-molecule *in situ* hybridization (smFISH)[17] in U2OS cells, we found that PUBLINC localizes almost exclusively to the cytoplasm (**Figure 1C**) and similar cytoplasmic enrichment is observed in other cells lines (**Figure 1D**). The number of PUBLINC copies expressed in a cell is ~80 based on the RPKM data and 68±8 based on the smFISH experiments that we have performed on U2OS cells, while 94% of PUBLINC copies are located in the cytoplasm and 6% are in the nucleus.

PUBLINC is a bona fide noncoding RNA

PUBLINC is computationally predicted to be a noncoding RNA by the PhyloCSF (**Figure 1E**) and Pfam/HMMER pipelines, with CPAT[18] and CPC[19] giving it borderline scores due to the presence of an open reading frame (ORF) with >100aa (see below) and similarity to hypothetical proteins (encoded by PUBLINC homologs) in other primates. Therefore, we also examined whether PUBLINC contains any translated ORFs using Ribo-seq data[20]. When examining ribosome footprinting datasets from diverse human cell lines (MDA-MB-231[21], HEK-293[22], U2OS[23], and KOPT-K1[24]), we did no observe any substantial footprints over any of the ORFs in PUBLINC, including a poorly conserved 108 aa ORF (**Figure 1E**) found close to the 5' end of the human transcript. Interestingly, a substantial pileups of ribosome-

protected fragments was observed at the very 5' end of PUBLINC in all Ribo-seq datasets we examined (**Figure 1E** and **Supplementary Figure 2**), but those did not overlap any ORFs with neither the canonical AUG start codon nor any of the common alternative start codons (**Supplementary Figure 2**). Additionally, the region overlapping the protected fragments also does not encode any conserved amino acid sequences in any of the frames. We conclude that it is highly unlikely that PUBLINC is translated into a functional protein, and the footprints observed in Ribo-seq data result from either a ribosome stalled at the very beginning of a transcript, or from a contaminant footprint of a different ribonucleoprotein (RNP) complex, as such footprints are occasionally present in Ribo-seq experiments[22,25].

The middle part of the PUBLINC sequence contains at least 12 structured repeated units

When comparing the PUBLIC sequence to itself, we noticed a remarkable similarity among some parts of its sequence (**Figure 2A**). Manual comparison of the sequences revealed that the central ~3.5 kb of PUBLINC in human, mouse, and other mammalian species can be decomposed into twelve repeating units of ~300 nt each. Interestingly, these units appear to have resulted from a tandem sequence duplication that occurred at least 100 million years before the split of the eutherian mammals, as when performing pairwise comparisons among repeats from different species, units from different species were more similar to each other than to other units from the same species (data not shown). Overall, the sequences have diverged to a level where there are no sequence stretches that are strictly identical among all the repeats in human. At the core of the most conserved regions within the repeats we identify four sequence and structured motifs (**Figures 2C-E**), some combination of which appears in each of the repeats: (i) one or two Pumilio binding sites (defined by the consensus UGU[AG]UAUA); (ii) a short predicted stem-loop structure with four paired bases and a variable loop sequence, with importance of the structure supported by the preferential A→G and G→A mutations in the second stem-loop that would preserve the stem (**Figures 2D** and **Supplementary Figure 3**, also detected by EvoFold[26]); (iii) a U-rich stretch of 2–5 bases; (iv) a stem-loop structure with eight or nine predicted base pairs. Whenever these four core units appear, further sequence conservation is found upstream and downstream of them. Interestingly, some of the repeated units, namely 3–5 and 7–9, appear to be more constrained during

mammalian evolution than others (**Figure 2C**), and those units also tend to contain all of the repeat motifs, with more intact sequences and structures (**Figure 2E**).

PUBLINC contains multiple functional binding sites for the PUM2 protein

In order to identify potential protein binding partners of the repeating units and of other PUBLINC fragments, and to confirm binding of PUBLINC to Pumilio, we amplified the 8th repeat unit and a region from the 3' end of PUBLINC, transcribed *in vitro* the sense and antisense of the 8th repeat and the sense of the 3' end using the T7 polymerase with biotinylated UTP bases, incubated the labeled RNA with U2OS cell lysate, and subjected the resulting IP to mass spectrometry. Among the proteins identified as binding different regions of PUBLINC (**Supplementary Data 1**) we focus here on one that has predicted binding sites within the repeat units and for which supporting evidence for a functional interaction is available from other sources – PUM2, one of the two Pumilio proteins found throughout vertebrates[10], and which has predicted binding elements in eight of the repeat units. To test for support for a direct interaction between PUM2 and PUBLINC, we reanalyzed of PAR-CLIP data from HEK-293 cells[27] and found that that PUM2 binds at least 17 sites on PUBLINC. These experimentally verified sites (all exhibiting T→C mutations characteristic of PAR-CLIP and overlapping Pumilio binding sites) overlapped ten of the 11 PUM2 binding motifs within repeated units 2–10. It is notable that PUBLINC has an unusual density of Pumilio binding sites encoded in its sequence – there are 17 non-overlapping instances of the UGU[AG]UAUA motifs in the PUBLINC sequence compared to 0.38 expected by chance ($P<0.001$, see Methods). The number and density of Pumilio motifs within PUBLINC are higher than those found in all but one human gene (PLCXD1, which has 18 PUM2 binding sites mostly located in transposable elements, compared to 0.12 expected).

To test whether PUM2 binds to PUBLINC in U2OS cells, we also performed RNA Immuno-Precipitation (RIP) followed by qPCR, and found a striking enrichment of the PUBLINC transcript among the RNAs bound by PUM2 (Methods and **Figure 3**). Interestingly, the enrichment of PUBLINC among PUM2 targets was reduced after arsenite stress. We conclude that PUBLINC contains at least 17 confident binding sites for PUM2, most of which appear in conserved positions within the conserved repeated units, and that this binding might be regulated following stress and potentially other conditions.

PUBLINC knockdown and over-expression preferentially affects mRNA levels of predicted and experimentally verified PUM2 targets

As PUM2 is reported to affect mRNA stability[11,28], we next tested if changes in PUBLINC expression affect the levels of PUM2 targets. PUBLINC was perturbed using either one of two individual siRNAs or a pool of four siRNAs (Dharmacon), with a pool of four siRNAs yielding ~4-fold knockdown and individual siRNAs yielding ~2-fold knockdown (**Figure 4A**). In order to test the consequences of increased PUBLINC levels, we cloned PUBLINC into an expression vector where it was driven by a CMV promoter, and transfected the expression vector into U2OS and HeLa cells, which resulted in 2–16 fold up-regulation (**Figure 4A**). We obtained consistent effects with two independent siRNAs 48 hrs (**Supplementary Figure 4A, Supplementary Data 2**), with 51 genes consistently down-regulated by at least 20% after treatment with both siRNAs and 23 genes consistently up-regulated by at least 20%. The stronger knockdown using a pool of siRNAs (**Figure 4A**) resulted in more substantial changes in gene expression – 584 were consistently down-regulated by at least 30% in two replicats and 68 genes were consistently up-regulated (**Supplementary Data 2)**. Further, changes following PUBLINC down-regulation at 24 hr were strongly inversely correlated with the changes observed 24 h after PUBLINC over-expression (**Supplementary Figure 4B** and **Supplementary Data 2**, Spearman r =–0.54, $P<10^{-10}$), suggesting that the differential expression was indeed driven by changes in PUBLINC abundance. Strikingly, genes with enrichment of Pumilio motifs in their 3' UTRs were repressed more than controls when PUBLINC was downregulated, and their expression levels increased more than controls when PUBLINC was upregulated in both cell lines (**Figure 4B**). These differences remained significant after controlling for the increased lengths of the 3' UTRs of genes bearing Pumilio motifs (**Supplementary Figure 5A**) and when considering experimentally verified PUM2 targets from HEK-293 cells (these effects were strongest 48 hr after transfection, **Supplementary Figure 5B**). These results suggest that hundreds of genes regulated by PUM2 are sensitive to PUBLINC levels, with increased amounts of PUBLINC alleviating repression of PUM2 targets and decreased amounts increasing repression.

When we inspected the Gene Ontology annotations enriched in the different sets of genes responsive to PUBLINC perturbations, after correction for multiple testing

using TANGO[29], the only significantly enriched group were genes bound by PUM2 in the PAR-CLIP data and down-regulated 48 hr after PUBLINC knockdown. These genes were enriched with categories associated with cell cycle and mitosis, including "M phase of the cell cycle" (8 genes; $P = 6.4\times10^{-6}$) and "Spindle" (8 genes; $P = 1.2\times10^{-7}$) (**Figure 4C**). Interestingly, these genes were not affected at 24 hours after PUBLINC knockdown or over-expression (**Figure 4C**), and enrichments of PUBLINC targets were also significant when evaluated compared to all PUM2-bound targets, suggesting a cumulative, and perhaps cell-cycle-dependendt effect of PUBLINC perturbation on PUM2 targeting of genes involved in mitosis.

As Pumilio proteins may affect translation in addition to their effects on mRNA stability, we evaluated the translational consequences of PUBLINC perturbation after 48 hr using Ribo-seq[30]. Consistent with the RNA-seq data, the number of translating ribosomes on mRNAs with predicted or experimentally verified PUM2 target sites was reduced following PUBLINC KD (**Supplementary Figure 6A**). However, when normalizing for changes in mRNA levels, translation efficiency of PUM2 targets did not appear to be preferentially affected (**Supplementary Figure 6B**), suggesting that the main effects of PUBLINC on PUM2 targets are through effects on mRNA stability rather than translation.

PUBLINC-dependent regulation of Pumilio target 3' UTRs is dependent on the canonical Pumilio binding sites

In order to test whether regulation of PUM2 targets is dependent on canonical Pumilio binding sites, we utilized a luciferase reporter vector containing three strong Pumilio responsive elements (PREs) as well as a control reporter with mutated sites, in which the three UGUACAUA motifs were mutated to ACAACATA (mPRE)[11,28]. As expected, over-expression of PUM1 or PUM2 proteins in U2OS cells led to increased repression in a PRE-dependent manner (**Figure 4D**). Over-expression of PUBLINC, on the other hand, alleviated the repression of the PRE-containing luciferase mRNA, without affecting luciferase containing mRPE elements. Knockdowns of PUBLINC or PUM1/2 failed to yield a consistent effect on luciferase activity (**Figure 4D**), possibly because of the limited knockdown efficiency using siRNAs or through feedback regulation of PUM2 on its own mRNA (see Discussion). Overall, these

results indicate that the PUBLINC-dependent changes in abundance of PUM2 targets are likely mediated through canonical Pumilio binding sites.

Discussion

To our knowledge, we describe here the first example of a lncRNA that contains multiple highly conserved consensus binding sites for an RNA binding protein (RBP), and that is required for proper regulation of the RBP targets at physiological levels. One particularly interesting question for future studies is the functional importance and roles of the other conserved elements found in the PUBLINC repeats, in particular the two predicted hairpin structures. Conserved secondary structures are generally rarely detectable in lncRNAs[1], and so PUBLINC presents an opportunity for studying such structures and their potential functions. It is possible that these additional elements serve as binding sites for other RNA-binding proteins, whose binding may either facilitate the binding of PUM2 to PUBLINC or affect PUM2 protein stability or activity. Potentially interesting candidates for interacting with PUBLINC repeats that were identified in the mass spectrometry analysis are known RBPs such as IGF2BP1/2/3, XRN2, and PABPN1. In addition we observed that the interferon response pathway proteins IFIT1/2/3/5 and their downstream companion PKR could bind PUBLINC sequence. IFIT proteins were observed to bind the antisense of the PUBLINC 8th repeat unit, suggesting that they may recognize a structural element rather than a primary sequence within the repeat, whereas PUM2 bound only the sense sequence, consistent with its known sequence specificity. We were so far unable to substantiate interactions with IFIT1 and PKR by reciprocal pulldown experiments, but if this interaction is indeed specific it would link PUBLINC to the reported functions of Pumilio in viral response – PUM1 and PUM2 were shown to be functionally stimulated after migration into SG upon viral infection[31] – an event that induces the interferon pathway.

Further studies will be required in order to uncover the physiological consequences of the regulation of PUM2 targets by PUBLINC, but the enrichment of cytokinesis-related genes among the PUM2 targets that are sensitive to PUBLINC levels suggests that PUBLINC may modulate regulation of chromosomal segregation during mitosis by PUM2, and might even affect the conserved roles of Pumilio proteins in regulating

asymmetric cell divisions during embryonic development. An intriguing question is whether the relatively high levels of PUBLINC in U2OS cells correspond to a basal state, in which PUBLINC exerts a minimal effect on PUM2 that is increased when stimuli increase PUBLINC expression, or to a state where PUBLINC actively buffers substantial regulation by PUM2. Most results point to the former scenario, as relatively modest over-expression of PUBLINC resulted in stronger effects on Pumilio activity than its knockdown. Another possibility suggested by the enrichment of cell-cycle regulated genes among the most prominent PUBLINC/PUM2 targets is that the regulation of PUBLINC/PUM2 is cell-cycle dependent.

Another interesting question is whether PUBLINC affects PUM2 regulation through binding a substantial number of functional PUM2 proteins through its numerous binding sites or by transient binding that alters PUM2 stability or activity. Answering this question is complicated by the negative autoregulation of PUM2, which is binding their own 3' UTR[27]. We did not observe consistent and strong effects of PUBLINC perturbations on PUM1/2 mRNA or protein level but it is possible that those effects are masked by the feedback regulation. If, for instance, PUBLINC binding facilitates PUM2 degradation, we are expecting increased PUM2 production that may result in unaltered PUM2 protein levels but reduced availability of functional PUM2 in the cells. Importantly, PUBLINC is expressed at levels that are compatible with a substantial effect on PUM2 activity. We estimate that there are >50 of copies of PUBLINC in mammalian cells, with >100 copies in neuronal cells with each containing at least 17 potential PUM2 binding sites. Therefore, the PUM2 binding sites on PUBLINC constitute a relatively abundant binding platform that can modulate PUM2 activity.

Lastly, it is interesting to note that in all our experiments we saw PUBLINC preferentially interacting with PUM2 over PUM1. PUM1 is also highly expressed in U2OS cells (based on RNA-seq its mRNA levels are ~1/2 those of PUM2, and based on Ribo-seq, the number of ribosome protected fragments on PUM1 and PUM2 mRNAs are similar), and should bind the same sequence motifs as PUM2. However, while some RNAs are bound by both PUM1 and PUM2, many others specifically bind only one of these proteins [13], indicating that there may be other factors – such as

the other conserved elements in the PUBLINC repeats – that may govern the specificity of PUBLINC to bind and affect PUM2.

Materials and methods

Cell culture

Human cell line U2OS (osteosarcoma) and HeLa (cervical carcinoma) were routinely cultured in DMEM containing fetal bovine serum (10%) and 100 U penicillin / 0.1 mg/ml streptomycin at 37 °C in a humidified incubator with 5% CO_2. For inducing stress, Sodium Arsenite solution (Sigma) was added to the media at a final concentration of 500μM for 1h.

Plasmids and siRNAs

Plasmids transfections were performed using PolyEthylene Imine (PEI)[32] (PEI linear, M*r* 25000 from Polyscience Inc). In order to overexpress PUBLINC, the full transcript of the Linc was amplified from human genomic DNA (ATCC NCI-BL2126) using the primers TGCCAGCGCAGAGAACTGCC (Fv) and GGCACTCGGGAGTGTCAGGTTC (Rev) and cloned into a ZeroBlunt TOPO vector (Invitrogen) and then subcloned into the pcDNA3.1(+) vector (Invitrogen). PUM1 and PUM2 were over-expressed using pEF-BOS vectors[31] (a kind gift of Prof. Takashi Fujita). As controls in over-expression experiments we used pBluescript II KS+ (Stratagene). Plasmids were used in the amount of 0.1μg per 100,000 cells in 24 well plates for 24 h before cells were harvested. For luciferase experiments, the plasmids pGL4.13, psiCheck-1 containing 3X wild type PRE (Pumilio Recognition Element), which is underlined in the following sequence, 5'-TTGTTGTCGAAAATTGTACATAAGCCAA, psiCheck-1 containing 3X mutated PRE 5′-TTGTTGTCGAAAAT**ACA**ACATAAGCCAA and psiCheck-1 with no PRE were all as described[11,28] (a kind gift of Dr. Aaron Goldstrohm). pGL4.13 was used in the amount of 5ng per 20,000 cells in 96 well plates while the different psiCheck-1 plasmids were used in the amount of 15ng per 20,000 cells in 96 well plates. Transfection time was 48h prior to further experimental procedures.

Gene knockdown was done by siRNAs directed against PUBLINC, PUM1 and PUM2 genes (all from Dharmacon, **Supplementary Table 1**), while as control we used the mammalian non-targeting siRNA (Lincode Non-targeting Pool, Dharmacon) at final concentration of 50nM for 24h or 48h prior to further experimental procedures. The transfections were employed using the PolyEthylene Imine.

siRNA transfection into HeLa cells were doing using 100 nM siRNA and Dharmafect (Dharmacon) transfection reagent and using siRNA buffer only as a control, and transfection of pCDNA3.1-PUBLINC was into HeLa cells was peformed using Lipofectamine 2000.

Real-time PCR analysis of gene expression

Total RNA was isolated using TRI reagent (MRC), then reverse transcribed using an equal mix of oligo dT and random primers (Quanta), according to the manufacturer protocols. For determination of all genes levels real-time PCR was conducted using Fast SYBR qPCR mix (Life technologies). The primer sets that were used for the differenet genes are in the listed in **Supplementary Table 2**. The assays contained 10-50 ng sample cDNA in a final volume of 10 µl and were run on AB qRT-PCR system ViiA 7 (Applied Biosystems). All genes relative expression levels were normalized to GAPDH levels.

Fluorescent In-Situ Hybridization

Probe libraries were designed according to Stellaris guidelines and synthetized by Stellaris as described in Raj et al[17]. Libraries consisted of 48 probes of length 20bps, complementary to the PUBLINC sequence according to the Stellaris guidelines (**Supplementary Table 3**). Hybridizations were done overnight with Cy5 labeled probes. DAPI dye (Inno-TRAIN Diagnostik Gmbh) for nuclear staining was added during the washes. Images were taken with a Nikon Ti-E inverted fluorescence microscope equipped with a 100× oil-immersion objective and a Photometrics Pixis 1024 CCD camera using MetaMorph software (Molecular Devices, Downington, PA). The image-plane pixel dimension was 0.13 microns. Quantification was done on stacks of 3–12 optical sections with Z-spacing of 0.3 microns. Dots were automatically detected using a custom Matlab program, implementing algorithms described in Raj et al[17]. Briefly, the dot stack images were first filtered with a 3-dimensional Laplacian of Gaussian filter of size 15 pixels and standard deviation of 1.5 pixels. The number of connected components in binary thresholded images was then recorded for a uniform range of intensity thresholds and the threshold for which the number of components was least sensitive to threshold selection was used for dot detection. Automatic threshold selection was manually verified and corrected for errors. Background dots were detected according to size and by automatically

identifying dots that appear in more than one channel (typically <1% of dots) and were removed.

Biotin pulldown assay

Biotinylated transcripts were synthesized using PCR fragments ,prepared using the primer sets for PUBLINC repeat #8, as template (Fw: 5'-TTCTAATACGACTCACTATAGGGCACGTGCCTATATCCATCAGGT and Rv: 5'- GACAATGGTCAATGTGCCTCC) and for the 3' end (Fw: 5'- TAA TAC GAC TCA CTA TAG GG TCTAGAGGCGTGTTGCCATT and RV: 5'-CTGTGTGTAGGCACAACATCC). The 5' part of the Fw primers encoded the T7 RNA polymerase promoter sequence. PCR products were purified, transcribed in vitro over-night using Biotin RNA labeling mix (Roche) and T7 RNA polymerase (NEB) and treated with DnaseI (Quanta). Streptavidin-sepharose high performance beads (GE Healthcare) were blocked using 20 mM HEPES pH7.9, 100 mM KCl, 10 mM $MgCl_2$, 0.01% NP-40, 1mM DTT, 1mg/ml BSA, 200 µg/ml Glycogen and 200 µg/ml yeast tRNA for 2.5 hr 4°C in 7 rpm rotation. In addition, protein extract was prepared using 20 mM Tris-HCl at pH 7.5, 150 mM NACl, 1.5 mM $MgCl_2$, 2mM DTT, 0.5% Na-Deoxycholate, 0.5% NP-40, complete protease inhibitor cocktail (Sigma) and 100unit/ml RNase inhibitor (EURx) for 15 min on ice and further centrifuged at 13,000 RPM for 15 min at 4°C. Afterwards, 0.5-2mg of the protein extract was pre cleared by incubation with Streptavidin-sepharose beads for 3h at 4°C in 6RPM rotation. Then, 50 pmole of the in-vitro transcribed RNA was incubated with the pre-cleared protein extract for 3h at 4°C in 6RPM rotation. The formed RNA-protein complexes were isolated by incubation with the blocked beads for 1.5 h at 4°C in 7RPM rotation. Then, protein was isolated by incubation of the complex with 50µg/ml RnaseA for 15 min at 4°C in 50RPM. Then, the protein was precipitated by incubation with acetone for over-night at -20C, and washed 2 times with 80% sterile ethanol. SDS sample buffer was added to the protein pellet and boiled at 95C for 5 min. Then, protein was separated by 4-12% Express Page gradient gel (GeneScript) and was stained using silver nitrate procedure. Then, the whole lane of protein bands were analyzed using Mass spectrometry analysis as described[33].

RNA Immunoprecipitation (RIP)

Immunoprecipitation (IP) of endogenous RNP complexes from whole-cell extracts was performed as described described by Yoon et al[34]. In brief, cells were incubated with lysis buffer (20 mM Tris-HCl at pH 7.5, 150 mM NACl, 1.5 mM $MgCl_2$, 2mM DTT, 0.5% Na-Deoxycholate, 0.5% NP-40, complete protease inhibitor cocktail (Sigma) and 100unit/ml RNase inhibitor (EURx)) for 15 min on ice and centrifuged at 13,000 RPM for 15 min at 4°C. Part of the supernatants was collected for the total cell lysate input. Rest of the supernatants containing 1-2 mg protein extract were incubated for 2-3 hours at 4°C in gentle rotation with protein A/G magnetic beads (GeneScript) that were pre-washed and coated with antibodies against GAPDH (SantaCruz Biotechnology) and PUM2 (Bethyl) at 4°C in gentle rotation for over-night. As negative control we incubated the magnetic beads-antibodies complexes with lysis buffer. Afterwards, the beads were washed with the lysis buffer five times, each time separated by magnetic force. The remaining mixture of magnetic beads-antibodies-protein-RNA complexes were separated as half were mixed with sample buffer and boiled at 95C for 5 minutes for further analysis by Western blot. The other half was incubated with 1mg/ml Proteinase K for 30min at 37C with gentle shaking to remove proteins. The remaining RNA was separated by Trizol. The RNPs isolated from the IP materials was further assessed by RT-qPCR analysis.

Western blot analysis

The level of the desired proteins was analyzed by Western blot assay as previously described[35] using the antibodies against PUM2 (Bethyl) and GAPDH (Santa Cruz Biotechnology). The immune-complexes were detected by enhanced chemiluminescence (ECL) (Biological Industries).

Ribosome profiling

48h after transfection of the U2OS with siRNAs, cylcoheximide treatments were carried out as previously described[20]. Cells were lysed in lysis buffer (20mM Tris 7.5, 150mM KCl, 5mM MgCl2, 1mM dithiothreitol , 8% glycerol) supplemented with 0.5% triton, 30 U/ml Turbo DNase (Ambion) and 100µg/ml cycloheximide, ribosome protected fragments were then generated cloned and sequenced as previously described [20].

RNA-seq and data analysis

Strand-specific mRNA-seq libraries were prepared from U2OS cells using the TruSeq Stranded mRNA Library Prep Kit (Illumina) and sequenced on a NextSeq 500 machine to obtained at least 23 million 75 nt reads. Strand-specific mRNA-seq libraries for HeLa cells were prepared as described[36]. All sequencing data has been deposited to the SRA database (Accession SRPXXXX). Reads were aligned to the human genome (hg19 assembly) using STAR Aligner[37], and read counts for individual genes (defined as overlapping sets of RefSeq transcripts annotated with their Entrez Gene identifier) were counted using htseq-count[38] and normalized to reads per million aligned reads (RPF). Only genes with an average RPM of at least 50 normalized reads across the experimental conditions were considered and fold changes were computed after addition of a pseudo-count of 0.1 to the RPM in each condition.

Sequence analyses

Whole genome alignments were obtained from the UCSC genome browser. Expected numbers of Pumilio binding sites were computed by applying dinucleotide-preserving permutations to the sequences and counting motif occurrences in the shuffled sequences. 3' UTR-length-matched control targets were selected by dividing the genes into ten bins based on 3' UTR lengths and randomly sample the same numbers of genes not enriched with Pumilio target sites as the number of genes enriched with sites from each bin.

Luciferase assays

The activity of Pumilio was determined by Luciferase assay as described[39]. Briefly, 20,000 Cells were plated in a 96-well plate. After 24 hr cells were co-transfected with pGL4.13 as internal control and with the indicated psiCheck plasmids. In addition, the cells were transfected with the various siRNAs or plasmids (as described above). After 48 hr, luciferase activity was recorded using the Dual-Glo Luciferase Assay System (Promega) in the Micro plate Luminometer device (Veritas). A relative response ratio (RRR), from RnLuc signal/FFLuc signal, was calculated for each sample. Percent of change is relative to the control siRNA or control plasmid.

Statistics

All results are represented as an average ± SEM of at least 3 independent experiments. Statistics was performed as Student's t-test or Anova with Tuckey's post hoc test for 3 or more groups to be compared. In all results * $p<0.05$, ** $p<0.01$, *** $p<0.001$. Plots were prepared using custom R scripts. Gene Ontology enrichment analysis was performed using the WebGestalt server[40] and corrected for multiple testing using TANGO[29], using all the expressed genes as background set and Benjamini-Hochberg correction for multiple testing.

Acknowledgements

We thank members of the Ulitsky lab for useful discussions and comments on the manuscript. I.U. is incumbent of the Sygnet Career Development Chair for Bioinformatics and recipient of an Alon Fellowship. Work in the Ulitsky lab is supported by grants to I.U. from the European Research Council (Project "lincSAFARI"), Israeli Science Foundation (1242/14 and 1984/14), the I-CORE Program of the Planning and Budgeting Committee and The Israel Science Foundation (grant no 1796/12), the Minerva Foundation, the Fritz-Thyssen Foundation and by a research grant from The Abramson Family Center for Young Scientists.

Figures

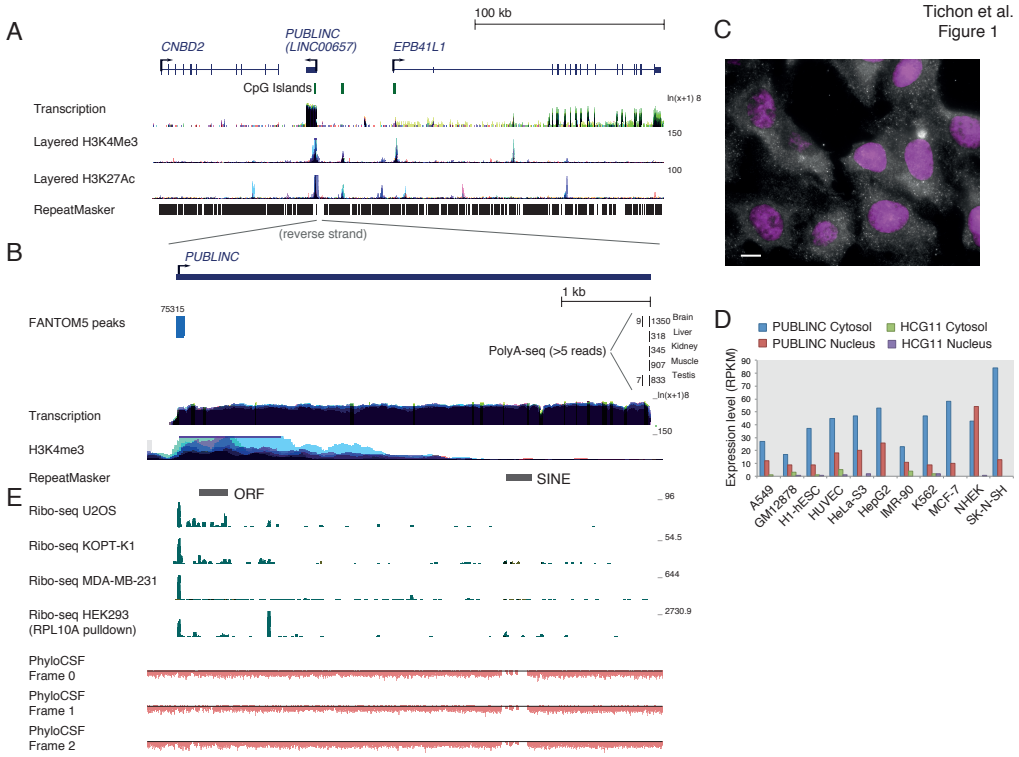

Figure1. Overview of the human PUBLINC locus. (A) Genomic neighborhood of PUBLINC. CpG island annotations and genomic data from the ENCODE project taken from the UCSC genome browser. **(B)** Support for the transcription unit of PUBLINC. Transcription start site information taken from the FANTOM5 project[41]. Polyadenylation sites taken from PolyA-seq dataset[42]. ENCODE datasets and repeat annotations from the UCSC browser. **(C)** Predominantly cytoplasmic localization of PUBLINC by smFISH. Scale bar 10µm. **(D)** Estimated copy number of PUBLINC in 22 independent cells from two independent experiments. **(E)** Expression levels of PUBLINC and HCG11 in the ENCODE cell lines (taken from the EMBL-EBI Expression Atlas (https://www.ebi.ac.uk/gxa/home). **(E)** Support for the noncoding nature of PUBLINC. Ribosome protected fragments from various human cell lines (MDA-MB-231[21], HEK-293[22], U2OS[23], KOPT-K1[24]) mapped to the PUBLINC locus and PhyloCSF[43] scores. All PhyloCSF scores in the locus are negative.

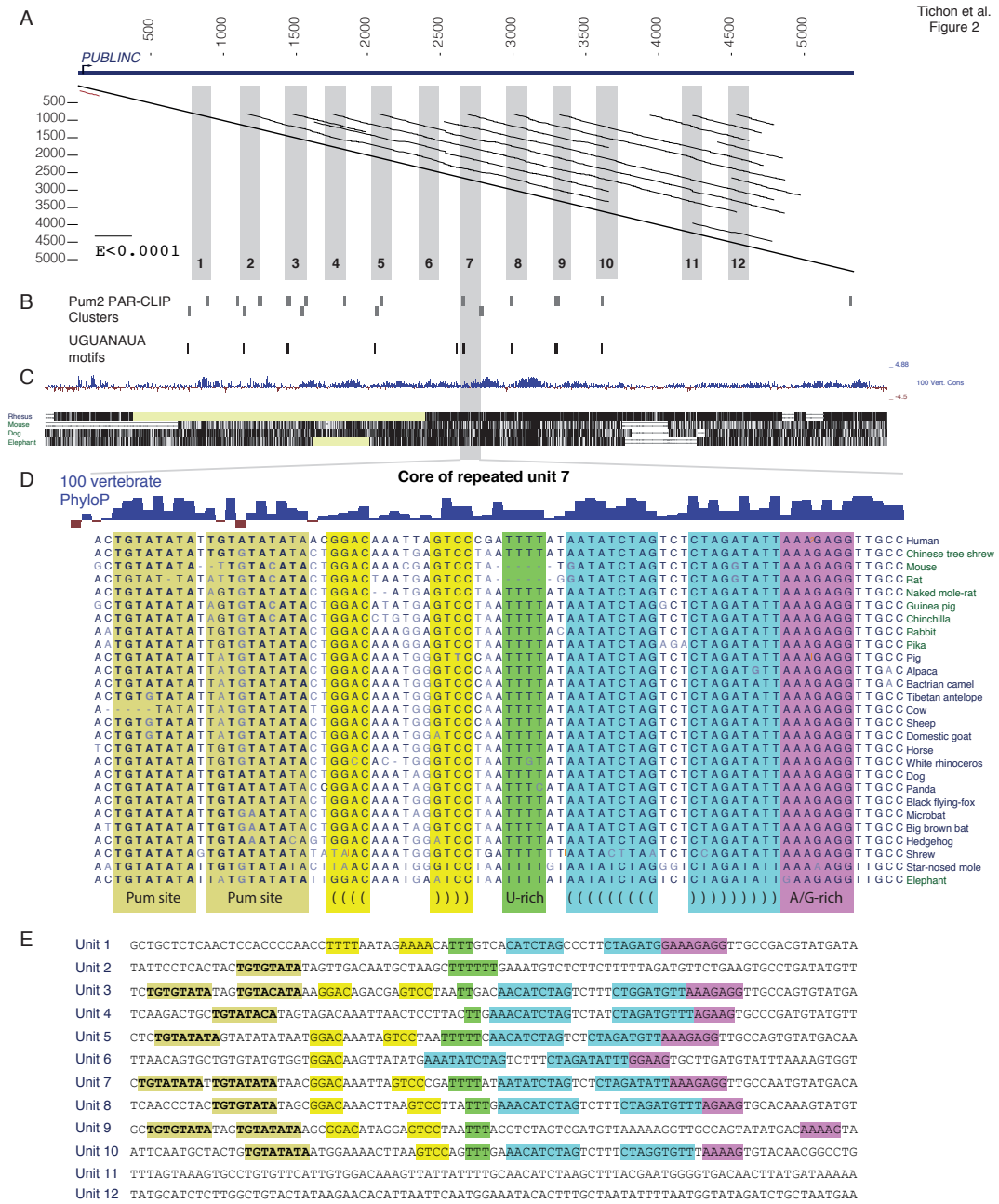

Figure 2. The repeated nature of the PUBLINC RNA. (A) A dotplot computed using plalign[44] (http://fasta.bioch.virginia.edu/) comparing PUBLINC with itself. The off-diagonal lines indicate high scoring local alignments between different parts of the sequence. Grey boxes indicate the core of the 12 manually annotated repeated units. **(B)** Clusters identified by PARalyzer[45] within the PUBLINC sequence using the PUM2 PAR-CLIP data[27] and positions of TGTRTATA motifs. **(C)** Sequence conservation of the PUBLINC locus, with PhyloP[46] scores for single-base-level conservation. **(D)** Detailed conservation of the seventh repeated unit. Shaded regions

indicate the five motifs present in most repeated units. **(E)** Core sequences of ten of the 12 repeated units, with the same shading as in D.

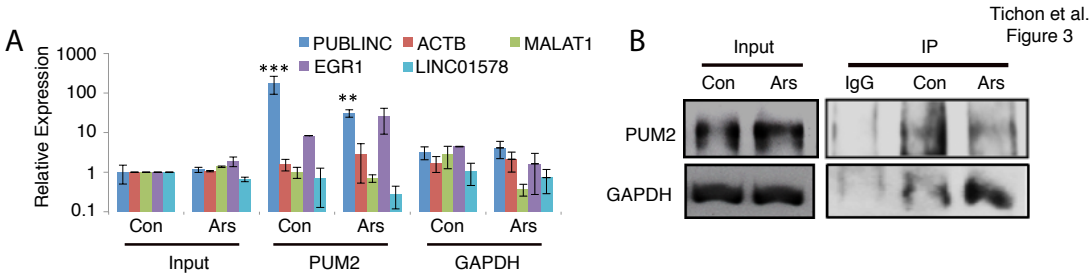

Figure 3. PUM2 binds PUBLINC. (A) Recovery of the indicated transcripts in the input and in the PUM2 or GAPDH IPs. All enrichments are normalized to GAPDH mRNA and to the input sample. (B) Western blots of the indicated factors in the input and IP samples.

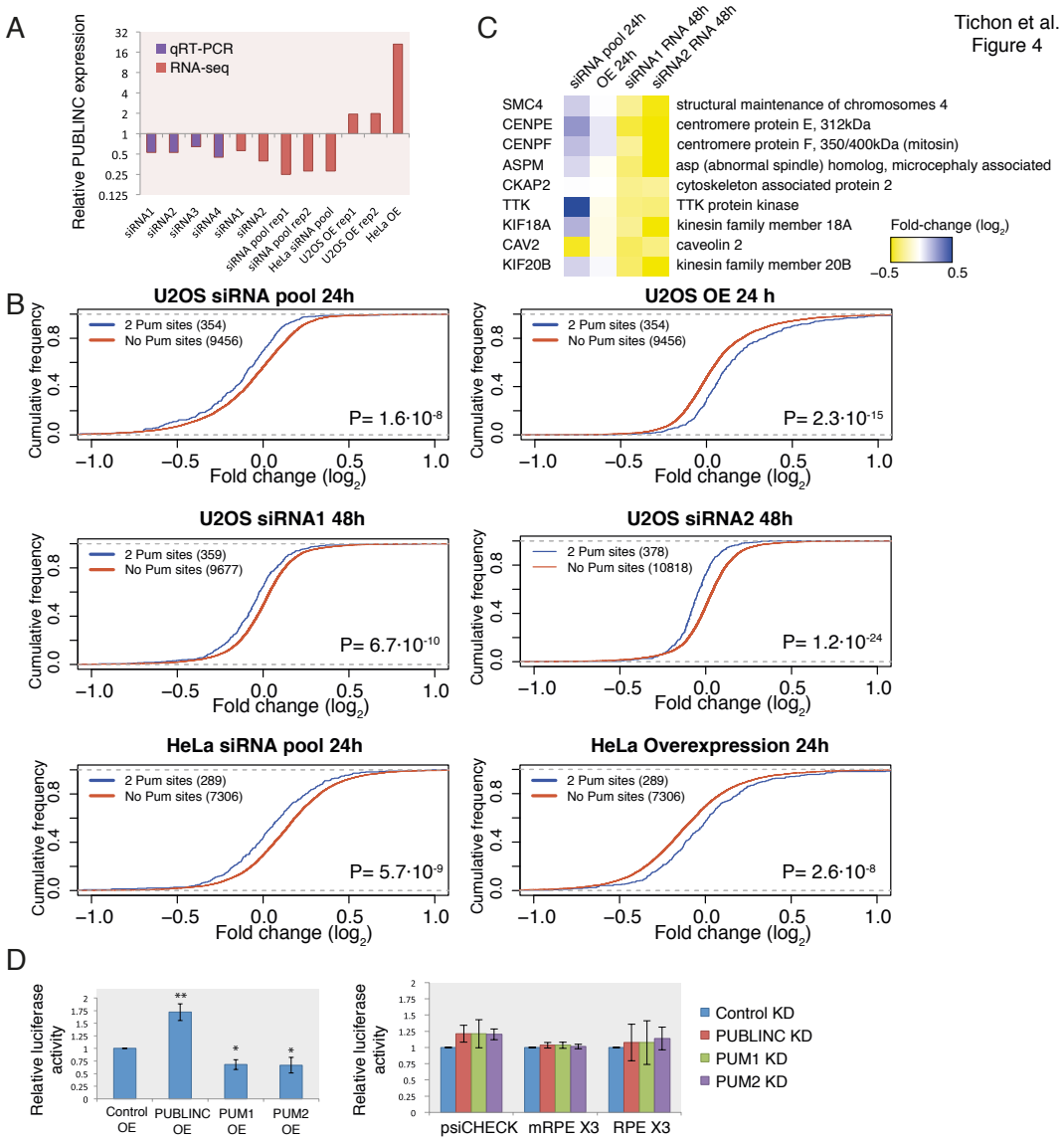

Figure 4. PUBLINC modulates expression of Pumilio targets. (A) Changes in PUBLINC expression following knockdown using siRNAs or overexpression (OE) from a plasmid, as measured by either qRT-PCR or RNA-seq. (B) Changes in expression of Pumilio targets compared to controls, following the indicated treatment. "2 Pum sites" are genes that contain at least two canonical Pumilio binding site over what is expected by chance in their 3' UTRs, "Controls" are those genes that do not contain more sites in their 3' UTRs than expected by chance. (C) Changes in expression of genes with annotated roles in the M phase of the cell cycle and/or the mitotic spindle following the indicated perturbations. (D-E) Changes in luciferase activity measured from the indicated vectors following knockdown (KD, panel D) or over-expression (panel E) of the indicated genes. In Panel (E), only the PRE-containing 3'UTR has been tested.

References

1 Ulitsky, I. & Bartel, D. P. lincRNAs: Genomics, Evolution, and Mechanisms. *Cell* 154, 26-46(2013).

2 Wapinski, O. & Chang, H. Y. Long noncoding RNAs and human disease. *Trends Cell Biol* 21, 354-361(2011).

3 Rinn, J. L. & Chang, H. Y. Genome regulation by long noncoding RNAs. *Annual review of biochemistry* 81, 145-166(2012).

4 Lee, J. T. & Bartolomei, M. S. X-inactivation, imprinting, and long noncoding RNAs in health and disease. *Cell* 152, 1308-1323(2013).

5 Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A. *et al.* Long non-coding antisense RNA controls *Uchl1* translation through an embedded SINEB2 repeat. *Nature* 491, 454-457(2012).

6 Gong, C. & Maquat, L. E. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature* 470, 284-288(2011).

7 Memczak, S., Jens, M., Elefsinioti, A., Torti, F. *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495, 333-338(2013).

8 Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B. *et al.* Natural RNA circles function as efficient microRNA sponges. *Nature* 495, 384-388(2013).

9 Guo, J. U., Agarwal, V., Guo, H. & Bartel, D. P. Expanded identification and characterization of mammalian circular RNAs. *Genome biology* 15, 409(2014).

10 Spassov, D. S. & Jurecic, R. The PUF family of RNA-binding proteins: does evolutionarily conserved structure equal conserved function? *IUBMB life* 55, 359-366(2003).

11 Weidmann, C. A., Raynard, N. A., Blewett, N. H., Van Etten, J. *et al.* The RNA binding domain of Pumilio antagonizes poly-adenosine binding protein and accelerates deadenylation. *RNA* 20, 1298-1319(2014).

12 Menon, K. P., Sanyal, S., Habara, Y., Sanchez, R. *et al.* The translational repressor Pumilio regulates presynaptic morphology and controls postsynaptic accumulation of translation factor eIF-4E. *Neuron* 44, 663-676(2004).

13 Galgano, A., Forrer, M., Jaskiewicz, L., Kanitz, A. *et al.* Comparative analysis of mRNA targets for human PUF-family proteins suggests extensive interaction with the miRNA regulatory system. *PloS one* 3, e3164(2008).

14 Vessey, J. P., Vaccani, A., Xie, Y., Dahm, R. *et al.* Dendritic localization of the translational repressor Pumilio 2 and its contribution to dendritic stress granules. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 26, 6496-6508(2006).

15 Lee, M. H., Hook, B., Pan, G., Kershner, A. M. *et al.* Conserved regulation of MAP kinase expression by PUF RNA-binding proteins. *PLoS genetics* 3, e233(2007).

16 Moore, F. L., Jaruzelska, J., Fox, M. S., Urano, J. *et al.* Human Pumilio-2 is expressed in embryonic stem cells and germ cells and interacts with DAZ (Deleted in AZoospermia) and DAZ-like proteins. *Proceedings of the National Academy of Sciences of the United States of America* 100, 538-543(2003).

17 Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. *et al.* Imaging individual mRNA molecules using multiple singly labeled probes. *Nature methods* 5, 877-879(2008).

18 Wang, L., Park, H. J., Dasari, S., Wang, S. *et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic acids research* 41, e74(2013).
19 Kong, L., Zhang, Y., Ye, Z. Q., Liu, X. Q. *et al.* CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic acids research* 35, W345-349(2007).
20 Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147, 789-802(2011).
21 Rubio, C. A., Weisburd, B., Holderfield, M., Arias, C. *et al.* Transcriptome-wide characterization of the eIF4A signature highlights plasticity in translation regulation. *Genome biology* 15, 476(2014).
22 Ingolia, N. T., Brar, G. A., Stern-Ginossar, N., Harris, M. S. *et al.* Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell reports* 8, 1365-1379(2014).
23 Eichhorn, S. W., Guo, H., McGeary, S. E., Rodriguez-Mias, R. A. *et al.* mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues. *Molecular cell* 56, 104-115(2014).
24 Wolfe, A. L., Singh, K., Zhong, Y., Drewe, P. *et al.* RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer. *Nature* 513, 65-70(2014).
25 Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S. *et al.* Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* 154, 240-251(2013).
26 Pedersen, J. S., Bejerano, G., Siepel, A., Rosenbloom, K. *et al.* Identification and classification of conserved RNA secondary structures in the human genome. *PLoS computational biology* 2, e33(2006).
27 Hafner, M., Landthaler, M., Burger, L., Khorshid, M. *et al.* Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 141, 129-141(2010).
28 Van Etten, J., Schagat, T. L., Hrit, J., Weidmann, C. A. *et al.* Human Pumilio proteins recruit multiple deadenylases to efficiently repress messenger RNAs. *The Journal of biological chemistry* 287, 36370-36383(2012).
29 Ulitsky, I., Maron-Katz, A., Shavit, S., Sagir, D. *et al.* Expander: from expression microarrays to networks and functions. *Nature protocols* 5, 303-322(2010).
30 Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324, 218-223(2009).
31 Narita, R., Takahasi, K., Murakami, E., Hirano, E. *et al.* A novel function of human Pumilio proteins in cytoplasmic sensing of viral infection. *PLoS pathogens* 10, e1004417(2014).
32 Durocher, Y., Perret, S. & Kamen, A. High-level and high-throughput recombinant protein production by transient transfection of suspension-growing human 293-EBNA1 cells. *Nucleic acids research* 30, E9(2002).
33 Isaac, R., Boura-Halfon, S., Gurevitch, D., Shainskaya, A. *et al.* Selective serotonin reuptake inhibitors (SSRIs) inhibit insulin secretion and action in pancreatic beta cells. *The Journal of biological chemistry* 288, 5682-5693(2013).

34 Yoon, J. H., Abdelmohsen, K., Srikantan, S., Yang, X. *et al.* LincRNA-p21 suppresses target mRNA translation. *Molecular cell* 47, 648-655(2012).
35 Tichon, A., Eitan, E., Kurkalli, B. G., Braiman, A. *et al.* Oxidative stress protection by novel telomerase activators in mesenchymal stem cells derived from healthy and diseased individuals. *Current molecular medicine* 13, 1010-1022(2013).
36 Guo, H., Ingolia, N. T., Weissman, J. S. & Bartel, D. P. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* 466, 835-840(2010).
37 Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15-21(2013).
38 Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166-169(2015).
39 Van Etten, J., Schagat, T. L. & Goldstrohm, A. C. A guide to design and optimization of reporter assays for 3' untranslated region mediated regulation of mammalian messenger RNAs. *Methods* 63, 110-118(2013).
40 Wang, J., Duncan, D., Shi, Z. & Zhang, B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic acids research* 41, W77-83(2013).
41 Consortium, F., the, R. P., Clst, Forrest, A. R. *et al.* A promoter-level mammalian expression atlas. *Nature* 507, 462-470(2014).
42 Derti, A., Garrett-Engele, P., Macisaac, K. D., Stevens, R. C. *et al.* A quantitative atlas of polyadenylation in five mammals. *Genome research* 22, 1173-1183(2012).
43 Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* 27, i275-282(2011).
44 Pearson, W. R. Flexible sequence similarity searching with the FASTA3 program package. *Methods in molecular biology* 132, 185-219(2000).
45 Corcoran, D. L., Georgiev, S., Mukherjee, N., Gottwein, E. *et al.* PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome biology* 12, R79(2011).
46 Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research* 20, 110-121(2010).