Widespread localisation of lncRNA to ribosomes: Distinguishing features and evidence for regulatory roles.

Juna Carlevaro-Fita[1,2,3], Anisa Rahim[4], Roderic Guigo[1,2,3], Leah A. Vardy[4,5,6], Rory Johnson[1,2,3,6]

1. Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain.
2. Universitat Pompeu Fabra (UPF), Barcelona, Spain.
3. Institut Hospital del Mar d'Investigacions Mèdiques (IMIM), 08003 Barcelona, Spain.
4. A*STAR Institute of Medical Biology, 8A Biomedical Grove, Immunos, Singapore 138648.
5. School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore.
6. Correspondence to RJ (rory.johnson@crg.eu) or LV (leah.vardy@imb.a-star.edu.sg).

Keywords: Long noncoding RNA; LncRNA; ribosome; translation; cytoplasm; transposable element; antisense.

Abstract

The function of long noncoding RNAs (lncRNAs) depends on their location within the cell. While most studies to date have concentrated on their nuclear roles in transcriptional regulation, evidence is mounting that lncRNA also have cytoplasmic roles. Here we comprehensively map the cytoplasmic and ribosomal lncRNA population in a human cell. Three-quarters (74%) of lncRNAs are detected in the cytoplasm, the majority of which (62%) preferentially cofractionate with polyribosomes. Ribosomal lncRNA are highly expressed across tissues, under purifying evolutionary selection, and have cytoplasmic-to-nuclear ratios comparable to mRNAs and consistent across cell types. LncRNAs may be classified into three groups by their ribosomal interaction: non-ribosomal cytoplasmic lncRNAs, and those associated with either heavy or light polysomes. A number of mRNA-like features destin lncRNA for light polysomes, including capping and 5'UTR length, but not cryptic open reading frames or polyadenylation. Surprisingly, exonic retroviral sequences antagonise recruitment. In contrast, it appears that lncRNAs are recruited to heavy polysomes through basepairing to mRNAs. Finally, we show that the translation machinery actively degrades lncRNA. We propose that light polysomal lncRNAs are translationally engaged, while heavy polysomal lncRNAs are recruited indirectly. These findings point to extensive and reciprocal regulatory interactions between lncRNA and the translation machinery.

Introduction

The past decade has witnessed the discovery of a tens of thousands of long non-protein coding RNAs (lncRNAs) in our genome, with profound implications for our understanding of molecular genetics, disease and evolution. Focus is now shifting to understanding the function to these molecules. We reason that such function is likely to be intimately linked to the location of lncRNA within the cell.

Following the first compelling discoveries of chromatin regulatory lncRNAs such as XIST (1)and HOTAIR (2), a paradigm was established for lncRNAs as nuclear-restricted, epigenetic regulatory molecules (3). However, it is not clear to what extent this is true for the >10,000 lncRNAs that remain uncharacterised (4-7). Indeed growing evidence points to lncRNA having diverse roles outside of the cell nucleus, including regulation of microRNA activity (8), protein sequestration(9), and mRNA translation(10).

Somewhat paradoxically, cytoplasmic lncRNA has recently been reported to interact with the ribosome. In footprinting experiments to map ribosome-bound transcripts genome-wide, the Weissman group identified a considerable number of lncRNAs directly engaged by the translation machinery (11), an observation subsequenttly corroborated in an independent study(12). These transcripts do not contain classical features of protein-coding sequence, and various analyses have argued that these lncRNAs are not productively translated in most cases (13,14). It is not yet clear whether ribosomal recruitment is a general property of all lncRNA in the cell. If not, it is of interest to understand what features distinguish ribosomal lncRNAs.

The biological significance of ribosomal lncRNA remains unclear. Two principle types of potential regulatory functions for ribosomal lncRNAs have been proposed: either sequence-specific regulation of mRNA translation or general regulation of ribosome function (15). LncRNA and mRNA arising from opposite genomic strands can form stable RNA-RNA hybrids that are localised in ribosomes (10). Through such “cis-antisense” interactions, lncRNA may specifically regulate stability and translation of their mRNA partner (10), although this has not yet been demonstrated at a genome-wide scale. The advent of ribosome footprinting technology has prompted the idea that lncRNA may non-specifically regulate translation through direct binding by ribosomes (15). Cryptic open reading frame (ORF) sequences within lncRNA may be recognised

by the ribosome, directly resulting in translational repression or else enabling recruitment of regulatory proteins. Other more mundane scenarios are also possible: ribosomes might be a default destination of all polyadenylated mRNA-like transcripts, where they are recognised as non-coding and processed by one of various known quality surveillance pathways.

In the present study we take these studies further by comprehensively mapping the entire known cytoplasmic and ribosomal lncRNA population of a human cell line. We show that the majority of cytoplasmic lncRNAs are robustly and verifiably associated with ribosomes. We show evidence that lncRNAs can be divided into classes based on subcellular location and distinguished by a variety of features. These classes likely serve distinct regulatory roles in translation. Finally we show that the translation machinery serves as the endpoint of the lncRNA life-cycle. We conclude that, rather than being an exception, ribosomal recruitment is frequently the destination of cytoplasmic lncRNAs.

Results

Creating a high confidence lncRNA catalogue

Our aim was to map the distribution of lncRNAs in the cytoplasm and on the polysomes of human cells. A potential confounding factor in any analysis of ribosome-bound RNAs is the possibility of misannotated protein-coding transcripts (16). These represent a non-negligible fraction of lncRNA annotation, due to the technical challenges of correctly identifying protein coding sequences with high sensitivity, as well as biological factors: a number of annotated lncRNAs have subsequently been found to encode peptides, including small "micropeptides", which were overlooked by conventional annotations (17,18).

We decided to implement the most stringent possible filtering to remove protein coding transcripts from our analysis, even at the expense of omitting some genuine non-coding transcripts. We first removed lncRNAs that could be unannotated extensions of protein-coding genes or pseudogenes. Remaining genes were filtered using a panel of methods for identifying protein coding sequence (Figure 1A and Materials and Methods). Altogether 9057 lncRNA transcripts (61.9%), 6763 genes (73.8%) were unanimously classified as non-coding - these we refer to as "filtered lncRNAs" (Figure 1A). The remaining genes of uncertain protein coding status are henceforth referred to as "potential protein coding RNAs" (4415 transcripts, 1878 genes). The complete sets of potential protein coding and filtered lncRNAs are available in Supplementary Table S1.

Mapping the cytoplasmic and ribosomal lncRNA population

We sought to create a comprehensive map of cytopasmic lncRNA localisation in a human cell. We chose as a model the K562 human myelogenous leukaemia cell line. Being an ENCODE Tier I cell, it has extensive transcriptomic, proteomic and epigenomic data publically available (19). We subjected cytoplasmic cellular extracts to polysome profiling, an ultracentrifugation method to identify ribosome-bound RNAs and distinguish transcripts bound to single or multiple ribosomes (Figure 1B) (20). Extracts were divided into three pools: "Heavy Polysomal", corresponding to high molecular

weight complexes cofractioning with >6 ribosomes; “Light Polysomal”, cofractioning with 2-6 ribosomes; and low molecular weight complexes corresponding to non-translated, cytoplasmic RNAs (Figure 1C). The latter contains free mRNAs found in the high peak in fraction 1, the 40 and 60S ribosomal subunits (fractions 2 and 3) and mRNAs that are bound by a single ribosome (fraction 4) - we define these as “Free Cytoplasmic” throughout the paper. It is important to note that although this fraction includes some RNAs bound by ribosomal subunits, or individual ribosomes, the majority of these are not considered to be efficiently translated (20).

Custom microarrays probing the entire Gencode v7 long noncoding RNA catalogue were used to analyse RNAs in the free cytoplasmic, light and heavy polysome fractions, in addition to total input RNA (see Materials and Methods)(5). Microarrays also contained probes targeting 2796 protein-coding genes. High positive correlation was observed between microarray RNA concentration measurements and RNA-sequencing of the same cells from ENCODE (Supplementary Figure S1)(19). Correlation between microarray results and RNAseq measurements of cytoplasmic RNA was higher than with either nuclear or whole-cell RNA from the same cells, attesting to the purity of these cytoplasmic extracts (Supplementary Table S2). Using stringent cutoffs we detected 10.6% of filtered lncRNA transcripts (962 transcripts, representing 665 or 9.8% of genes) and 52.8% of mRNAs (1476) in K562 cytoplasm (Figure 1D). An additional 292 transcripts (3.2%, representing 255 or 3.7% genes) were detected only in the nucleus. Altogether, 1254 filtered lncRNA transcripts (13.9%, representing 875 or 13.0% of genes) were detected.

We classified cytoplasmic lncRNAs according to their maximal ribosomal association, resulting in 347 (37.6% of cytoplasmic lncRNA transcripts) Free Cytoplasmic, 373 (40.4%) Light Polysomal, and 204 (22.1%) Heavy Polysomal transcripts (Figure 1D). Altogether, 62.5% of lncRNA transcripts detected in the cytoplasm have maximal detection in Light or Heavy Polysomal fractions. Two lines of evidence support this classification approach. First, 75% (959/1287) of protein-coding mRNAs are classified as Heavy Polysomal, consistent with their being actively translated and in accordance with previous studies (Figure 1E)(20,21). Second, protein abundance measurements show that Heavy Polysomal mRNAs are translated most

efficiently (Supplementary Table S3) (22). In contrast, potential protein-coding transcripts had a similar global ribosome-association profile to filtered lncRNA, suggesting that they are not translated efficiently and underlining the stringency of our lncRNA filtering (Figure 1E). Ribosomal lncRNA are not apparently enriched for those that produce small peptides (Supplementary Table S4).

Cytoplasmic and ribosomal localisation has previously been reported for a number of lncRNA. To test the degree of agreement between these and our data, we examined the 297 lncRNA transcripts (from 60 genes) from the LncRNA Database (23) that are also present in the Gencode v7 annotation. SNHG5 (5) and Gas5 (9) were detected in the cytoplasm and classified as Free Cytoplasmic transcripts, consistent with previous reports. The snoRNA host Gas5 has previously been reported as associated with ribosomes (24). Although we classified this gene as Free Cytoplasmic based on its maximal detection, 11 out of 16 transcript isoforms of Gas5 were also clearly detected in Light and Heavy polysomal fractions although with lower microarray probe intensities. SNHG1 is another snoRNA host reported to be bound by ribosomes (25), which we classify in the Light Polysomal fraction. For other known lncRNAs, we map their subcellular location for the first time: GNAS-AS1 (Nespas) and MEM161B-AS1 are specifically associated with the Light Polysomal fraction.

Independent evidence for ribosomal interaction of lncRNA

We next looked for additional evidence to support ribosomal interaction of lncRNA. During ultracentrifugation, it is possible that lncRNAs associated with non-ribosomal, high molecular weight complexes may co-sediment with polyribosomes and thus represent false positives. To investigate this, we repeated polysome profiling on cells treated with puromycin (puro), a drug that disrupts ribosomes, and profiled a set of candidate transcripts by volume-normalised RT-PCR (Figure 2A, B). Bona fide ribosome-bound transcripts are expected to relocalise to the free cytoplasmic fraction in response to puromycin. Eleven out of 16 (69%) ribosomal lncRNAs were validated in this way, similar to the 4/4 protein coding mRNAs tested. In contrast, 2/3 Free Cytoplasmic lncRNAs we examined showed minimal response to puro treatment. Thus

in the majority of cases, cosedimentation reflects a physical interaction between lncRNA and ribosomes.

We performed additional validation using fluorescence in situ hybridisation (FISH) to visualise the localisation of lncRNA at subcellular resolution. We tested three Light Polysomal lncRNAs (Figure 3). ENST0000504230 displays diffuse cytoplasmic localisation and exclusion from nucleoli. In addition to cytoplasmic localisation, the snoRNA precursor transcript ENST00000545440 (SNHG1) shows pronounced concentrations around the periphery of the nucleus, likely to be endoplasmic reticulum, and at three nuclear loci – possibly its site of transcription, given that the HeLa genome is predominantly triploid (26). Finally, ENST00000545462 (previously described as HEIH, a prognostic factor in hepatocellular carcinoma)(27), also has pronounced staining in the nuclear periphery, as well as within the nucleolus. Thus, both PCR and hybridisation methods support the interpretation from microarray data of ribosomal recruitment of lncRNA.

Evidence for conserved function of ribosomal lncRNAs

Purifying evolutionary selection represents powerful evidence of functionality. A number of studies have shown that lncRNAs are under weak but non-neutral purifying evolutionary selection (5,28,29). We sought to test if this holds true for cytoplasmic lncRNAs, and in particular whether different classes of cytoplasmic lncRNA described above might have experienced different strengths of selection. We extracted PhastCons measures of exonic conservation and compared lncRNAs of distinct subcellular origins (Figure 4). Ancestral repeats were treated as neutrally-evolving DNA for comparison. As expected, protein coding exons have highly elevated conservation. Free Cytoplasmic, Light Polysomal and nuclear lncRNAs exhibit similar rates of non-neutral evolution. In addition, Heavy Polysomal lncRNAs contain a subset (~10%) of transcripts with elevated conservation, second only to the potential protein coding transcripts, and higher than other expressed lncRNAs ($P = 0.002$, $OR = 2.40$ Fisher test, testing the top 10% of Heavy Polysomal lncRNA vs other lncRNAs pooled). Thus, cytoplasmic lncRNAs experience purifying evolutionary selection consistent with conserved function.

Ribosomal lncRNA are highly expressed and consistently localised across cell types

We next investigated the organismal and subcellular expression patterns of lncRNA, in addition to their post-transcriptional processing. The steady state expression levels of cytoplasmic lncRNA is similar across cytoplasmic classes in independent K562 whole cell RNAseq, similar to that of mRNAs and well above nuclear-specific lncRNA (Figure 5A)(19). A similar trend is observed in human tissues: mean RPKM across Human Body Map tissues for all three cytoplasmic classifications exceed nuclear RNA (P = 5e-5 / 2e-5 / 0.0009 for Light Polysomal / Heavy Polysomal / Free Cytoplasmic vs nuclear, Wilcox test) (Supplementary Figure S2).

LncRNAs have been reported to be more tissue specific than mRNAs (4,5). Analysis of ubiquity, an inverse measure of tissue-specificity, of lncRNA in human tissues was consistent with this (Figure 5B). Despite this similarity in expression profiles, we find Heavy Polysomal lncRNAs to be significantly more ubiquitous in their tissue expression profiles compared to other lncRNA classes (P= 1.2e-4, Fisher Test Heavy Polysomal vs Nucleus), and essentially the same as mRNAs (P=0.129, Fisher exact test).

Subcellular localisation of lncRNA reported by polysome profiling is consistent with similar analysis using ENCODE RNAseq (19). Transcripts we report as ribosomal or free cytoplasmic have significantly elevated cytoplasmic-nuclear ratios (Figure 5C) (P=0.027 OR 2.4, 1.8e-07 OR 4.4, 0.005 OR 2.6 for Heavy Polysomal, Light Polysomal and Free Cytoplasmic vs Nuclear transcripts, Fisher exact test). Indeed, Light Polysomal lncRNA have median cytoplasmic specificity that exceeds protein coding mRNAs. Heavy Polysomal transcripts have a more nuclear distribution, suggesting that while some transcripts are ribosomally bound, other copies are present in the nucleus. We next asked whether the observed subcellular localisation of lncRNA in K562 is conserved across other cell types (Figure 5D). Similar analysis on RNAseq from other cell types showed Light Polysomal and Free Cytoplasmic transcripts tend to have high cytoplasmic-nuclear distributions, often exceeding that of mRNAs, while Heavy Polysomal has a more mixed distribution that nevertheless differs from nuclear-specific transcripts. Protein-binding profiles of lncRNA yields a consistent picture, with lncRNA

tending to interact with proteins that localise to the same cellular compartment (Supplementary Figure S3). In summary, lncRNA subcellular localisation is consistent across cell types.

mRNA-like 5' regions distinguish ribosomally-bound lncRNAs

We next wished to identify factors that control the recruitment of lncRNA to ribosomes. The most obvious candidate feature is the ORF, especially given that lncRNAs contain abundant small ORF sequences that may be recognised by ribosomes. In protein, ORF length influences the number of ribosomes that can simultaneously bind, and hence the ribosomal fraction (compare mean sense ORF length for heavy and light polysome mRNA in Supplementary Figure S4)(12). However for lncRNA we could no evidence that ORFs determine ribosomal recruitment: neither their total ORF coverage, nor their number of ORFs, nor the length of their longest ORF is different from random sequence or correlates with ribosomal recruitment (Supplementary Figure S5). Nor apparently does gross gene structure or GC content, both clearly distinct between lncRNA and mRNA, appear to influence ribosomal recruitment (Supplementary Figure S6, S7).

We hypothesised that factors known to influence mRNA recognition by ribosomes may also apply to lncRNA. For mRNAs, a number of factors control the scanning and engagement by ribosomes, including 3' polyadenylation, RNA structures within the 5' UTR and 7-methylguanylate capping (30). To investigate whether polyadenylation influences ribosomal recruitment, we estimated the efficiency of polyadenylation of cytoplasmic and nuclear lncRNAs using ENCODE RNAseq on polyA+ and polyA- nuclear RNA. Although mRNA are more polyadenylated than lncRNA, we found no difference in polyadenylation efficiency between ribosomal and non-ribosomal lncRNAs (Supplementary Figure S8). We recently showed that splicing efficiency of lncRNAs is lower than mRNAs (31), but it does not distinguish ribosomal lncRNAs from other types (Supplementary Figure S9).

We next looked at the role of the 5' end in ribosomal recruitment. Although lncRNAs do not have identifiable ORFs and hence 5' UTRs, nevertheless they do contain abundant short "pseudo-ORFs": random occurrences of in-frame start and stop

277 codons. We defined the "pseudo-5'UTR" to be the region upstream of the first AUG
278 trinucleotide of the lncRNA sequence. Although secondary structures in the 5'UTR have
279 been shown to strongly influence translation of mRNAs (32), there is no overall
280 difference in structural propensity between ribosomal and other lncRNAs
281 (Supplementary Figure S10). However, the length of pseudo-5'UTRs does distinguish
282 ribosomal from non-ribosomal lncRNA. Similar to protein-coding transcripts, Light
283 Polysomal lncRNA have significantly longer 5'UTR regions than expected by chance
284 (here estimated from the transcript's reverse complement) (Figure 6A), while this effect
285 is essentially absent for other cytoplasmic lncRNAs. Thus long 5'UTR-like regions would
286 appear to contribute positively to ribosomal recognition of lncRNA.
287 Recognition of the 5' methyl-guanosine cap is required for mRNA scanning by
288 the 40S ribosomal subunit. Using CAGE (cap analysis of gene expression) data (19),
289 we examined the relationship between the ribosomal recruitment of lncRNA and
290 capping using logistic regression. As shown in Figure 6B, there is a strong positive
291 relationship between capping and recruitment to the Light Polysomal Fraction. In
292 contrast, this relationship is negative for Free Cytoplasmic and Heavy Polysomal
293 recruitment. This data suggests that capping of lncRNA is a driver of ribosomal
294 recruitment, at least to the light polysomal fraction.
295
296 **Endogenous retroviral fragments are negatively correlated with ribosomal**
297 **recruitment**
298 There is growing evidence that transposable elements (TEs) contribute functional
299 sequence to lncRNA (33,34). Taking all TE classes together, we observed an excess of
300 TE-derived sequence within Free Cytoplasmic lncRNAs (P=4e-14, compared to
301 remaining detected filtered lncRNAs, Wilcoxon test) (Figure 7A). Potentially protein
302 coding transcripts are significantly depleted for TEs (P=2e-16, compared to all detected,
303 filtered lncRNAs, Wilcoxon test). Given that protein coding transcripts are strongly
304 depleted for TE insertions (35), this latter observation supports the idea that a subset of
305 potential protein coding transcripts do indeed encode functional protein.
306 We were curious whether there exist TEs whose presence correlates with the
307 subcellular localisation of their host transcript (Figure 7B) (Materials and Methods).

Thus we systematically tested the relationship between subcellular localisation and TE class. We observed a relationship between the presence of Alu and transcript expression in K562: Alu are enriched amongst detected compared to undetected filtered lncRNAs (P=6e-7, Hypergeometric test), as recently described for human tissues (34). TcMar.Tigger, although rare, show evidence for preferential enrichment in polyribosomal lncRNAs (P=9e-4, Hypergeometric test). However the most obvious case is for the class of ERVL-MaLR, which are approximately two-fold enriched in free cytoplasmic lncRNAs compared to other expressed lncRNAs (Figure 7B). Closer inspection revealed that this effect is not due to a single repeat type, but rather to around a dozen subclasses of MST, MLT and THE endogenous retroelements (Figure 7C). We found no significant difference in the length of ERVL-MaLR insertions between lncRNA classes (Supplementary Figure S11). Rather it is the relative proportion of transcripts carrying an insertion that differs between groups. A selection of ERVL-MaLR containing lncRNAs are shown in Figure 7D.

Enrichment of ERVL-MaLR class elements in Free Cytoplasmic lncRNAs appears to be independent of cell type: using ribosome footprinting data from HeLa(36) we observe that ERVL-MaLR class TEs are specifically depleted from ribosome-bound lncRNAs (Figure 7E). Together these data suggest that endogenous retrovirus fragments may influence lncRNA trafficking in the cell.

Evidence for cis-antisense lncRNA-mRNA pairing in ribosomes

Several reports exist describing hybridisation of lncRNA to mRNA through complementary sequences, resulting in trafficking of the former to ribosomes. Antisense complementarity between lncRNA and mRNA could take one of two forms: more conventionally, the two transcripts may originate from opposite strands of the same genomic locus, thus sharing complementary sequence regions (here “exonic antisense”, also referred to as “cis-antisense”) (Figure 8A) (10). More recently, it was shown that lincRNA-P21 contains regions of complementarity to mRNAs, through which they hybridise and consequently localise together in the ribosome (37). Importantly, the genes for lincRNA-P21 and its targets are located in distinct genomic loci – these we define here as “trans-antisense” pairs.

We investigated whether either type of antisense may contribute to the observed recruitment of lncRNA to the ribosomes. We first hypothesised that exonic antisense lncRNAs would be more frequently localised in heavy polysomes, due to hybridisation to their corresponding (actively translated) mRNA. We classified all lncRNA by their genomic organisation with respect to protein-coding genes (5): intergenic (not overlapping), exonic antisense, intronic antisense, or intronic same sense. Consistent with our hypothesis, lncRNAs identified in heavy polysomes are significantly enriched for exonic antisense transcripts compared to those in other cellular compartments (P=4.2e-5, Fisher exact test) (Figure 8B). This finding is consistent with lncRNA / mRNA hybrids existing in human ribosomes. An example of such a cis-antisense pair is shown in Figure 8C. If this is the case, we would expect mRNAs bound by antisense heavy polysomal lncRNA to be more highly expressed than others. Examining RNAseq expression data we find this to be the case: mRNAs antisense to heavy polysomal lncRNA are significantly more highly expressed than mRNAs antisense to other lncRNA classes (P=7e-4, Wilcoxon test)(Figure 8D). In the course of this analysis, we also made the incidental observation that intronic same sense lncRNAs tend to be nuclear-specific (Supplementary Figure S12).

Trans-antisense hybridisation is another potential means by which lncRNA could interact with mRNA and be recruited to ribosomes. In this model, the transcripts share homology on opposite strands, but are not transcribed from the same genomic locus (Figure 8A). Using a BLAST approach, we compiled all sense-antisense homology relationships between intergenic lncRNA and mRNA (Figure 8E). As a control, we performed the same operation with size-matched, randomised genomic regions instead of lncRNA. This analysis resulted in two observations: first, lncRNA as a whole are more likely to have trans-antisense homology to mRNA compared to random genomic sequence (P=1e-14, Fisher test, comparing all lncRNA to all shuffled); second, this tendency was observed with statistical significance in ribosomal lncRNA (P=0.002, Fisher test for heavy and light lncRNA combined) but not in free cytoplasmic and nuclear lncRNA. These findings were consistent across a range of different BLAST settings. This data point to possible trans-antisense lncRNA-mRNA hybridisation as a general regulatory mechanism of ribosomal lncRNA.

370

371 **Degradation of lncRNA by the ribosome**

372 We were next curious whether recruitment to ribosomes had any effect on
373 lncRNA. It was proposed by Chew et al(38) that lncRNA at ribosomes are subject to
374 nonsense mediated decay (NMD), and indeed one report does exist of ribosome-
375 dependent degradation of a snoRNA host gene (35). We tested whether blocking
376 translation has any outcome on the stability of ribosomal lncRNA identified here. Using
377 the same candidate genes tested previously, we tested whether interfering with
378 ribosomal function through drug-induced stalling (cyclohexamide) influenced lncRNA
379 stability (Figure 9). In a number of cases we observed elongation-dependent
380 degradation of lncRNA, often decreasing lncRNA amounts by several fold over 6 hours.
381 This effect was highly heterogeneous, with other transcripts unaffected by
382 cyclohexamide treatment. Thus, ribosomal recruitment leads to degradation of many
383 lncRNAs.

384

Discussion

In order to gain clues as to lncRNA function at a global level, we have comprehensively mapped the ribosomal and cytoplasmic lncRNA populations of a human cell. The very substantial populations of lncRNA we discover in these fractions is at odds with existing paradigms of lncRNA function as principally nuclear molecules. We must now consider the possibility that lncRNA play more diverse roles outside the nucleus, including translational control, cellular metabolism or signal transduction.

One key challenge in the study of cytoplasmic lncRNAs is to rule out the possibility that they encode a cryptic, unannotated protein product. This question has been discussed in excellent reviews elsewhere (15), and has not yet been satisfactorily resolved. Indeed, it is likely that an extensive “grey zone” of transcripts with weak protein coding potential exists (and indeed may form the substrate for novel protein evolution (39)). It is also plausible that some or many transcripts do exist that function both as protein-coding and noncoding transcripts, although apart from the archetypal SRA1 (40), few concrete examples have so far been presented (41,42). In this study we took great pains to filter any transcripts with even minimal probability of encoding protein, in the process collecting many weakly coding (“potential protein coding”) transcripts that may be of rich scientific interest in future. We describe a set of 1867 annotated lncRNAs that have varying degrees of evidence for encoding protein. These transcripts have intriguing characteristics intermediate between coding and non-coding RNA: they are under higher evolutionary selection than lncRNA, are depleted for transposable elements, and tend to be cytoplasmically enriched – similar to mRNAs. In contrast, they have ribosomal association profiles and expression ubiquity similar to lncRNA. Finally, their gene structures and expression levels are intermediate between coding and noncoding sequences. It will be fascinating in future to study whether these transcripts represent an intermediate timepoint in the evolution of either new proteins from non-coding sequences, or the evolution of non-coding RNAs from formerly coding transcripts.

We cannot rule out the possibility that some of our filtered lncRNA produce a peptide product. Recent studies have revealed the potential for large volumes of unrecognised protein coding capacity in mammalian genomes, either as small peptides

(43) or non-canonical ORF translation (44,45). However, even if these lncRNAs give rise to peptides, it does not necessarily follow that all are functional: occasional nonsense translation of ribosomally-localised lncRNA may occur with no functional consequences - “translational noise”. Future intensive mass spectrometry studies of short peptides, such as that carried out by Slavoff et al (18), will hopefully allow us to further improve lncRNA annotations.

We present several lines of evidence that ribosomal lncRNA are a large functional gene class that genuinely interacts with the translation machinery: (1) ribosomal lncRNA are puromycin sensitive; (2) fluorescence in situ hybridisation indicates their cytoplasmic localisation; (3) they have elevated cytoplasmic-nuclear ratios by independent ENCODE RNAseq data across diverse cell types; (4) their sequence is under similar or even elevated evolutionary selection compared to nuclear lncRNAs. Thus these transcripts are functional and appear to have a regulated and consistent subcellular localisation.

Polysome profiling appears to distinguish lncRNAs with distinct properties. We have attempted to rather crudely classify transcripts according to their fraction of maximum detection, but most transcripts are detected at varying concentrations in all fractions. Nevertheless, through this classification we have managed to discover features that distinguish lncRNA and laid a foundation for predicting lncRNA localisation *de novo*. Similarly, a recent study discovered an RNA motif that predicts and appears to confer cytoplasmic localisation (46). We find that lncRNAs localised in the Light Polysomal fraction tend to have mRNA-like 5’ features, more specifically a non-randomly long pseudo-5’ UTR length and the presence of a cap structure. This is consistent with the importance of 5’ recognition in the initiation of translation. Other mRNA-like features such as polyadenylation, GC content or open reading frames do not appear to affect ribosomal interaction at a global level. In contrast, repetitive sequence features, and particularly human endogenous retrovirus fragments, are negatively associated with ribosomal recruitment. This is perhaps to be expected, given that mRNAs tend to be depleted of such repeats, in contrast to lncRNA (35). The mechanism by which hERV prevent lncRNA from ribosomal recruitment remains to be ascertained, although we proposed recently that such fragments may interact with

protein complexes that could antagonise ribosomal binding (33). In summary, these findings represent a starting point for discovering features that distinguish lncRNA classes and may eventually lead to useful models for predicting such classes.

Light Polysomal and Heavy Polysomal lncRNA appear to represent functionally distinct classes of lncRNA. It is generally considered that heavy polysomes tend to be actively translating, while light polysomes represent more weakly translated messengers, and this is supported by our mRNA data. We interpret the lncRNAs in the Light Polysomal fraction to be engaged by two or a few ribosomes that are not in the process of translating an mRNA. The above features of 5' processing distinguish Light Polysomal transcripts clearly from Free Cytoplasmic transcripts, but not Heavy Polysomal. The latter tend to be more nuclear and more strongly evolutionarily conserved. Some clues to the origin of these differences may be gleaned from the observation that cis-antisense transcripts are enriched in the heavy fraction. Cis-antisense transcripts have been studied for a number of years, and cases have been described where the antisense lncRNA hybridises with its sense mRNA and accompanies it to the ribosome (10). Thus we might posit that lncRNA in heavy polysomes are involved in active translational processes and include transcripts that exist as hybrids with their sense mRNA partner. Such recognition is sequence specific, and we may guess that this localisation occurs indirectly: the lncRNA is recruited through its binding to a translated mRNA, and not directly engaged by ribosomes. In contrast, given their mRNA-like 5' features, we propose that Light Polysomal lncRNA include cases that are directly engaged by ribosomes, resulting in non-specific translational repression and/or lncRNA degradation(15). This model is outlined in Figure 10.

Although it is tempting to propose that ribosomal lncRNA regulate protein translation, we must also seriously consider an alternative possibility: that the ribosome represents the default endpoint of the lncRNA lifecycle, and it is rather the non-ribosomal cytoplasmic transcripts that are exeptional. Indeed, it is perhaps not surprising that these mRNA-like transcripts - capped, polyadenylated and 100-10,000 nt long – should be recognised by the cell and trafficked accordingly. We here show evidence that, at least for a subset of transcripts, the result of ribosomal recruitment is

478 degradation. That is, the translation machinery is also responsible for lncRNA
479 clearance, and that the regulatory relationship between lncRNA and the translational
480 machinery is reciprocal.
481

Materials and Methods

Polysome fractionation

For polysome fractionations, 20 million K562 cells were incubated with 100 ug/mL of cycloheximide (Sigma, Cat C4859) for 10 min. Cell pellets were resuspended in 200ul RSB buffer (20 mM Tris-HCl, pH 7.4, 20 mM NaCl, 30 mM $MgCl_2$, 200ug/mL cycloheximide, 0.2mg/mL heparin (Sigma, Cat No. H4787), 1000 unit/mL RNasin), then lysed with an equal volume of Lysis Buffer (1X RSB, 1 % Triton X-100, 2% Tween-20, 200ug/ul heparin) with or without 1% Na deoxycholate. Following incubation on ice for 10 min, extracts were centrifuged at 13,000 x g for 3 min to remove the nuclei. Supernatants were further centrifuged at 13,000 x g for 8 min at 4°C. Equal OD units were loaded onto 10% to 50% linear sucrose gradients (prepared in 10 mM Tris-HCl pH 7.4, 75 mM KCl and 1.5mM $MgCl_2$), and centrifuged at 36,000 rpm for 90 min at 8° C in a SW41 rotor (Beckman Coulter). Twelve fractions were collected from the top of the gradient using a piston gradient fractionator (BioComp Instruments). A UV-M II monitor (BIORAD) was used to measure the absorbance at 254 nm. 110ul of 10% SDS and 12 uL of proteinase K (10 mg/mL Invitrogen) was added to each 1ml fraction and incubated for 30 min at 42°C. Fractions 1-5, 6-8 and 9-11 were pooled corresponding to groups Free Cytoplasmic (Free / Monosomal), LP (Light Polysome) and HP (Heavy Polysome), respectively. For puromycin-treated samples, cells were incubated in 100ug/ml puromycin for 15 minutes prior to processing and puromycin was used in place of cyclohexamide in all the buffers.

Unfractionated cytoplasmic RNA and pooled polysomal RNAs were purified using phenol chloroform isoamyl extraction followed by LiCl precipitation to remove the heparin. The integrity of the samples was monitored by Bioanalyzer. For qRT-PCR analysis equal volumes of RNA were used to synthesise cDNA using the Superscript III Reverse Transcriptase (Invitrogen) according to manufacturer's instructions. Two bacterial spike-in RNAs, Dap and Thr were added before RNA purification to equal volumes of each polysomal RNA pool. Gene specific primers were used with SYBR Green for qRT-PCR on an ABI PRISM 7900 Sequence Detection Systems. Candidate CT values were normalized to the spike in controls Dap and Thr that were present at

equal concentrations per pool. Relative RNA levels are presented as a percentage of the RNA present in each pool with 100% RNA calculated as the sum of the FM, LP and HP pools.

Microarray Design

This study was carried out using Agilent custom gene expression microarrays, in the 8x60k format with 60mer probes. Probes were designed using eArray software with standard settings: Base composition methodology / 60mer / 4 probes per target / sense probes / best probe methodology / 3' bias. Probes were designed for 14700 transcripts from the entire Gencode v7 lncRNA catalogue, in addition to 26 known lncRNAs from www.lncrnadb.org (23) and 90 randomly-selected protein-coding housekeeping genes. The array was then filled with probes targeting 2796 randomly-selected protein-coding gene probes. Microarray design details are available from the Gencode website (http://www.gencodegenes.org/lncrna_microarray.html).

Microarray Hybridization and Probe Quantification

100 ng of total RNA was labeled using Low Input Quick Amp Labeling kit (Agilent 5190-2305) following manufacturer instructions. mRNA was reverse transcribed in the presence of T7-oligo-dT primer to produce cDNA. cDNA was then in vitro transcribed with T7 RNA polymerase in the presence of Cy3-CTP to produce labeled cRNA. The labeled cRNA was hybridized to the Agilent SurePrint G3 gene expression 8x60K microarray according to the manufacturer's protocol. The arrays were washed, and scanned on an Agilent G2565CA microarray scanner at 100% PMT and 3um resolution. Intensity data was extracted using the Feature Extraction software (Agilent).

Raw data was taken from the Feature Extraction output files and was corrected for background noise using the normexp method(47). To assure comparability across samples we used quantile normalization(48). All statistical analyses were performed with the Bioconductor project (http://www.bioconductor.org/) in the R statistical environment (http://cran.r-project.org/) (49).

Preparation of filtered lncRNA gene catalogues

We first filtered the former set to remove any transcripts that potentially result from misannotated extensions or isoforms of protein-coding genes or pseudogenes. Any gene was discarded that has at least one transcript fulfilling one of the following conditions: overlapping on the same strand a Gencode v18 annotated pseudogene, overlapping on the same strand an exon of a protein-coding mRNA, or lying within 5 kb and on the same strand as an Gencode v18 protein-coding transcript or pseudogene (1169 transcripts, 521 genes). This resulted in a dataset of 13,472 lncRNA transcripts (8641 genes). Next, genes having at least one transcript predicted as protein coding by at least one method, were classified as “potential protein coding RNAs” (4415 transcripts, 1878 genes), while the remainder were classified as “filtered lncRNAs”. The four filtering methods used were: 1) PhyloCSF, a comparative genomics method based on phylogenetic conservation across species (50). The analysis was performed using 29 mammalian nucleotide sequence alignments and assessing the three sense frames. The alignment of each transcript was extracted from stitch gene blocks given a set of exons from Galaxy(51). Transcripts with score >95 were classified as potential protein coding, following the work of Sun et al (52) . 2) Coding Potential Assessment Tool (CPAT)(53), using the score threshold of 0.364 described by the authors. 3) Coding Potential Calculator (CPC), a support vector machine-based classifier based on six biological sequence features, using a cutoff of 1 (54). 4) Peptides: We used experimental mass spectrometry tag mappings from Pinstripe to identify any transcripts producing peptides (55). Any transcript having an exonic, same strand tag mapping were designated as “potential protein coding”. Collectively, sequence filters reduced the pool of analyzed transcripts to 9057 transcripts (6763 genes). The full table of classification data for all Gencode v7 lncRNA is available in Supplementary Table S1.

Microarray probe filtering

LncRNA transcripts were considered to be present in a sample when at least three out of four microarray probes were reliable and not absent. The expression intensity value for “present” transcripts was computed as the median of its present probes. Protein coding genes were considered “present” if at least one probe was reliable and not absent, and the intensity value was that of one of the present probes, chosen randomly.

Variances in probe intensity values within probesets were significantly different when comparing all probsets from present transcripts in a sample (Levene's test). To avoid non representative intensity values, 5% of transcripts (for each sample) with highest probeset variance were removed from our dataset. Applying these filters we define as cytoplasmically detected 962 filtered lncRNAs (665 genes), 906 potential protein-coding transcripts (382 genes) and 1476 protein-coding genes that are detected in K562 cytoplasm.

RNAseq correlation analysis

ENCODE RNA-sequencing quantifications (Gencode v10 annotation) from cytoplasmic fraction of K562 cells was used to check correlation with microarray data. Correlation was calculated only with transcripts present in both ENCODE data (considered as present when RPKM bio-replicates mean different to 0 and IDR < 0.1) and microarray data.

Classification of array transcripts

From the polysome profiling analysis, detected lncRNAs and mRNAs were classified according to the microarray sample (condition) where they displayed the highest transcript-level signal. Thus, present transcripts were classified into Heavy Polysomal (Heavy P.), Light Polysomal (Light P.) and Free Cytoplasmic transcripts (Free C.) transcripts. The remaining protein coding genes, which were not present in any microarray condition were considered not present. Remaining filtered lncRNA transcripts were subsequently checked in ENCODE K562 nucleus RNAseq. Those detected (defined as RPKM bio-replicates mean > 0 and IDR < 0.1) were classified as nuclear specific transcripts. Remaining transcripts, which are not present in cytoplasm nor in the nucleus are considered not present (NotPre).

Cytoplasmic-nuclear localisation using RNAseq data

Cytoplasmic and nuclear RNAseq data from six different cell lines (K562, Hela, NHEK, HepG2, GM12878, HUVEC) were obtained from ENCODE (19). For each cell line we calculated cytoplasmic-nuclear RPKM ratios for transcripts detected in both that cell line

606 and K562. RPKM was calculated as the mean of two available technical replicates, and
607 only transcripts with mean > 0 RPKMs and IDR < 0.1 were considered present. We
608 calculated log2 ratios of cytoplasmic expression versus nuclear expression (RPKM
609 units) for those transcripts present in both nucleus and cytoplasm.
610
611 **Tissue Expression Analysis**
612 We extracted tissue expression values for 16 human tissues from Human Body Map
613 (HBM) RNAseq data, downloaded from ArrayExpress under accession number E-
614 MTAB-513. These data were used to quantify Gencode v10 transcripts using the
615 GRAPE pipeline(56). Transcripts were defined as ubiquitous if they had >0.1 RPKM
616 expression in all 16 tissues.
617
618 **Transposable element analysis**
619 The 2013 version of RepeatMasker human genomic repetitive element annotations
620 were downloaded from UCSC Genome Browser, and was converted to BED format.
621 Using the tool IntersectBED, we calculated (1) the number of instances of intersection,
622 and (2) the number of nucleotides of overlap, between each lncRNA transcript and each
623 transposable element. This analysis was carried out for both transposable element
624 types, and transposable element classes.
625
626 **ORF analysis**
627 We mapped all possible canonical open reading frames (ORFs) in each of six frames in
628 lncRNA and protein coding transcripts from Gencode. If more than one start codon is in
629 frame with a stop codon, only the start codon for the longest ORF is considered.
630
631 **CAGE analysis of lncRNA capping**
632 5' cap analysis was performed on cap analysis gene expression (CAGE) tags from
633 ENCODE (19) for K562 cytoplasmatic poly+ RNA and mapped these tags to the
634 microarray region comprising between 100nt before and after transcription start sites of
635 lncRNA. In order to assess the relationship between cytoplasmic class and capping, we
636 compared CAGE tag presence to fractional occupancy in each class. The latter was

637 calculated by subtracting input cytoplasmic log2 microarray expression intensity values
638 from each of the three polysome profiling fractions intensity values (Free C., Light P. or
639 Heavy P.). We divided transcripts into (log2) fraction occupancy bins from -2 to 2 at 0.5
640 bins. Transcripts with values outside this range were pooled into the last corresponding
641 bin. Logistic regression was performed to assess the relationship between CAGE tag
642 presence and occupancy.
643
644 **BLAST analysis of trans-homology between lncRNA and mRNA**
645 Gencode v7 transcript-level FASTA files of mRNA (Gene type “protein coding”) and
646 lncRNA were downloaded from Gencode. Two control sets analogous to lncRNA were
647 also collected and processed in exactly the same way: first, Bedtools “shuffle” tool was
648 used to extract random regions identical in size to the lncRNAs. Second, the introns of
649 each lncRNA were concatenated, then a fragment of sequence identical in size to the
650 mature lncRNA was extracted at a random location within this sequence. All sequences
651 were repeat-masked using RepeatMasker with “sensitive” and “human” settings. A
652 BLAST library was created using default settings with the mRNA sequences. LncRNA
653 and control sequences were BLASTed against this library with maximum expectation
654 value of 20.
655
656 **RNA Stability Assay**
657 K562 cells were incubated with or without cyclohexamide (100ug/ml) for three hours
658 followed by treatment with actinomycin D (5ug/ml) for 6 hours. RNA samples were taken
659 at 0 and 6 hours following actinomycin D treatment to assess the stability of the RNA in
660 the absence of transcription. RNA was purified using Trizol and Qiagen RNeasy
661 columns. 1ug of RNA was used to make cDNA using RevertAid H Minus reverse
662 transcriptase. Luminaris Color HiGreen High ROX qPCR master mix was used with
663 gene specific primers for qRT-PCR on an ABI PRISM 7900 Sequence Detection
664 Systems. Expression levels were normalised to the housekeeping gene GAPDH by the
665 delta-delta Ct method.
666

667 **Acknowledgements**

668

669 We thank members of the Guigo lab and the CRG Bioinformatics and Genomics
670 Programme for many ideas and discussions, particularly Marta Melé, Joao Curado, and
671 Marc Friedlaender, in addition to Fatima Gebauer (CRG, Gene Regulation Stem Cells
672 and Cancer Programme). We would particularly like to thank Ferran Reverter for
673 invaluable statistical advice. Thomas Derrien (University of Rennes) and Giovanni
674 Bussotti (EBI) helped with evolutionary conservation analysis. We thank the CRG
675 Genomics Core Facility, particularly Anna Ferrer, Maria Aguilar, Sarah Bonnin and
676 Manuela Hummel, for array hybridisation and analysis. The following CRG colleagues
677 generously helped with FISH experiments: Francois Le Dily, Maria Sanz, Carme Arnan,
678 Maria Teresa Zomeño, as well as Timmo Zimmermann and Raquel García from CRG
679 Microscopy Facility.

680 **References**

681

682 1. Brown, C.J., Ballabio, A., Rupert, J.L., Lafreniere, R.G., Grompe, M., Tonlorenzi, R. and
683 Willard, H.F. (1991) A gene from the region of the human X inactivation centre is
684 expressed exclusively from the inactive X chromosome. *Nature*, **349**, 38-44.
685 2. Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough,
686 L.H., Helms, J.A., Farnham, P.J., Segal, E. *et al.* (2007) Functional demarcation of active
687 and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, **129**, 1311-
688 1323.
689 3. Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas,
690 K., Presser, A., Bernstein, B.E., van Oudenaarden, A. *et al.* (2009) Many human large
691 intergenic noncoding RNAs associate with chromatin-modifying complexes and affect
692 gene expression. *Proceedings of the National Academy of Sciences of the United States*
693 *of America*, **106**, 11667-11672.
694 4. Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A. and Rinn, J.L.
695 (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global
696 properties and specific subclasses. *Genes & development*, **25**, 1915-1927.
697 5. Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G.,
698 Martin, D., Merkel, A., Knowles, D.G. *et al.* (2012) The GENCODE v7 catalog of human
699 long noncoding RNAs: analysis of their gene structure, evolution, and expression.
700 *Genome research*, **22**, 1775-1789.
701 6. Hangauer, M.J., Vaughn, I.W. and McManus, M.T. (2013) Pervasive Transcription of the
702 Human Genome Produces Thousands of Previously Unidentified Long Intergenic
703 Noncoding RNAs. *PLoS genetics*, **9**, e1003569.
704 7. Managadze, D., Lobkovsky, A.E., Wolf, Y.I., Shabalina, S.A., Rogozin, I.B. and Koonin,
705 E.V. (2013) The vast, conserved mammalian lincRNome. *PLoS computational biology*,
706 **9**, e1002917.
707 8. Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M.,
708 Tramontano, A. and Bozzoni, I. (2011) A long noncoding RNA controls muscle
709 differentiation by functioning as a competing endogenous RNA. *Cell*, **147**, 358-369.
710 9. Kino, T., Hurt, D.E., Ichijo, T., Nader, N. and Chrousos, G.P. (2010) Noncoding RNA
711 gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid
712 receptor. *Science signaling*, **3**, ra8.
713 10. Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A., Zucchelli, S., Fedele, S., Pesce, E.,
714 Ferrer, I., Collavin, L., Santoro, C. *et al.* (2012) Long non-coding antisense RNA controls
715 Uchl1 translation through an embedded SINEB2 repeat. *Nature*, **491**, 454-457.
716 11. Ingolia, N.T., Lareau, L.F. and Weissman, J.S. (2011) Ribosome profiling of mouse
717 embryonic stem cells reveals the complexity and dynamics of mammalian proteomes.
718 *Cell*, **147**, 789-802.
719 12. van Heesch, S., van Iterson, M., Jacobi, J., Boymans, S., Essers, P.B., de Bruijn, E.,
720 Hao, W., Macinnes, A.W., Cuppen, E. and Simonis, M. (2014) Extensive localization of
721 long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. *Genome*
722 *biology*, **15**, R6.
723 13. Banfai, B., Jia, H., Khatun, J., Wood, E., Risk, B., Gundling, W.E., Jr., Kundaje, A.,
724 Gunawardena, H.P., Yu, Y., Xie, L. *et al.* (2012) Long noncoding RNAs are rarely
725 translated in two human cell lines. *Genome research*, **22**, 1646-1657.
726 14. Guttman, M., Russell, P., Ingolia, N.T., Weissman, J.S. and Lander, E.S. (2013)
727 Ribosome profiling provides evidence that large noncoding RNAs do not encode
728 proteins. *Cell*, **154**, 240-251.

15. Ulitsky, I. and Bartel, D.P. (2013) lincRNAs: genomics, evolution, and mechanisms. *Cell*, **154**, 26-46.
16. Dinger, M.E., Pang, K.C., Mercer, T.R. and Mattick, J.S. (2008) Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS computational biology*, **4**, e1000176.
17. Kondo, T., Plaza, S., Zanet, J., Benrabah, E., Valenti, P., Hashimoto, Y., Kobayashi, S., Payre, F. and Kageyama, Y. (2010) Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. *Science*, **329**, 336-339.
18. Slavoff, S.A., Mitchell, A.J., Schwaid, A.G., Cabili, M.N., Ma, J., Levin, J.Z., Karger, A.D., Budnik, B.A., Rinn, J.L. and Saghatelian, A. (2013) Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nature chemical biology*, **9**, 59-64.
19. Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101-108.
20. Zhang, D., Zhao, T., Ang, H.S., Chong, P., Saiki, R., Igarashi, K., Yang, H. and Vardy, L.A. (2012) AMD1 is essential for ESC self-renewal and is translationally down-regulated on differentiation to neural precursor cells. *Genes & development*, **26**, 461-473.
21. Beilharz, T.H. and Preiss, T. (2004) Translational profiling: the genome-wide measure of the nascent proteome. *Briefings in functional genomics & proteomics*, **3**, 103-111.
22. Khatun, J., Yu, Y., Wrobel, J.A., Risk, B.A., Gunawardena, H.P., Secrest, A., Spitzer, W.J., Xie, L., Wang, L., Chen, X. *et al.* (2013) Whole human genome proteogenomic mapping for ENCODE cell line data: identifying protein-coding regions. *BMC genomics*, **14**, 141.
23. Amaral, P.P., Clark, M.B., Gascoigne, D.K., Dinger, M.E. and Mattick, J.S. (2011) lncRNAdb: a reference database for long noncoding RNAs. *Nucleic acids research*, **39**, D146-151.
24. Smith, C.M. and Steitz, J.A. (1998) Classification of gas5 as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5'-terminal oligopyrimidine gene family reveals common features of snoRNA host genes. *Molecular and cellular biology*, **18**, 6897-6909.
25. Pelczar, P. and Filipowicz, W. (1998) The host gene for intronic U17 small nucleolar RNAs in mammals has no protein-coding potential and is a member of the 5'-terminal oligopyrimidine gene family. *Molecular and cellular biology*, **18**, 4509-4518.
26. Adey, A., Burton, J.N., Kitzman, J.O., Hiatt, J.B., Lewis, A.P., Martin, B.K., Qiu, R., Lee, C. and Shendure, J. (2013) The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature*, **500**, 207-211.
27. Yang, F., Zhang, L., Huo, X.S., Yuan, J.H., Xu, D., Yuan, S.X., Zhu, N., Zhou, W.P., Yang, G.S., Wang, Y.Z. *et al.* (2011) Long noncoding RNA high expression in hepatocellular carcinoma facilitates tumor growth through enhancer of zeste homolog 2 in humans. *Hepatology*, **54**, 1679-1689.
28. Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P. *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223-227.
29. Ponjavic, J., Ponting, C.P. and Lunter, G. (2007) Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome research*, **17**, 556-565.
30. Jackson, R.J., Hellen, C.U. and Pestova, T.V. (2010) The mechanism of eukaryotic translation initiation and principles of its regulation. *Nature reviews. Molecular cell biology*, **11**, 113-127.
31. Tilgner, H., Knowles, D.G., Johnson, R., Davis, C.A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T.R. and Guigo, R. (2012) Deep sequencing of

779 subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the
780 human genome but inefficient for lncRNAs. *Genome research*, **22**, 1616-1625.
781 32. Kudla, G., Murray, A.W., Tollervey, D. and Plotkin, J.B. (2009) Coding-sequence
782 determinants of gene expression in Escherichia coli. *Science*, **324**, 255-258.
783 33. Johnson, R. and Guigo, R. (2014) The RIDL hypothesis: transposable elements as
784 functional domains of long noncoding RNAs. *RNA*, **20**, 959-976.
785 34. Kelley, D. and Rinn, J. (2012) Transposable elements reveal a stem cell-specific class of
786 long noncoding RNAs. *Genome biology*, **13**, R107.
787 35. Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L., Bourque, G., Yandell,
788 M. and Feschotte, C. (2013) Transposable elements are major contributors to the origin,
789 diversification, and regulation of vertebrate long noncoding RNAs. *PLoS genetics*, **9**,
790 e1003470.
791 36. Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M. and Weissman, J.S. (2012) The
792 ribosome profiling strategy for monitoring translation in vivo by deep sequencing of
793 ribosome-protected mRNA fragments. *Nature protocols*, **7**, 1534-1550.
794 37. Yoon, J.H., Abdelmohsen, K., Srikantan, S., Yang, X., Martindale, J.L., De, S., Huarte,
795 M., Zhan, M., Becker, K.G. and Gorospe, M. (2012) LincRNA-p21 suppresses target
796 mRNA translation. *Molecular cell*, **47**, 648-655.
797 38. Chew, G.L., Pauli, A., Rinn, J.L., Regev, A., Schier, A.F. and Valen, E. (2013) Ribosome
798 profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding
799 RNAs. *Development*, **140**, 2828-2834.
800 39. Knowles, D.G. and McLysaght, A. (2009) Recent de novo origin of human protein-coding
801 genes. *Genome research*, **19**, 1752-1759.
802 40. Lanz, R.B., McKenna, N.J., Onate, S.A., Albrecht, U., Wong, J., Tsai, S.Y., Tsai, M.J.
803 and O'Malley, B.W. (1999) A steroid receptor coactivator, SRA, functions as an RNA and
804 is present in an SRC-1 complex. *Cell*, **97**, 17-27.
805 41. Ulveling, D., Francastel, C. and Hube, F. (2011) Identification of potentially new
806 bifunctional RNA based on genome-wide data-mining of alternative splicing events.
807 *Biochimie*, **93**, 2024-2027.
808 42. Marques, A.C., Tan, J., Lee, S., Kong, L., Heger, A. and Ponting, C.P. (2012) Evidence
809 for conserved post-transcriptional roles of unitary pseudogenes and for frequent
810 bifunctionality of mRNAs. *Genome biology*, **13**, R102.
811 43. Ma, J., Ward, C.C., Jungreis, I., Slavoff, S.A., Schwaid, A.G., Neveu, J., Budnik, B.A.,
812 Kellis, M. and Saghatelian, A. (2014) Discovery of human sORF-encoded polypeptides
813 (SEPs) in cell lines and tissue. *Journal of proteome research*, **13**, 1757-1765.
814 44. Zu, T., Gibbens, B., Doty, N.S., Gomes-Pereira, M., Huguet, A., Stone, M.D., Margolis,
815 J., Peterson, M., Markowski, T.W., Ingram, M.A. *et al.* (2011) Non-ATG-initiated
816 translation directed by microsatellite expansions. *Proceedings of the National Academy*
817 *of Sciences of the United States of America*, **108**, 260-265.
818 45. Vanderperre, B., Lucier, J.F., Bissonnette, C., Motard, J., Tremblay, G., Vanderperre, S.,
819 Wisztorski, M., Salzet, M., Boisvert, F.M. and Roucou, X. (2013) Direct detection of
820 alternative open reading frames translation products in human significantly expands the
821 proteome. *PloS one*, **8**, e70698.
822 46. Zhang, B., Gunawardane, L., Niazi, F., Jahanbani, F., Chen, X. and Valadkhan, S.
823 (2014) A novel RNA motif mediates the strict nuclear localization of a long noncoding
824 RNA. *Molecular and cellular biology*, **34**, 2318-2329.
825 47. Ritchie, M.E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A. and
826 Smyth, G.K. (2007) A comparison of background correction methods for two-colour
827 microarrays. *Bioinformatics*, **23**, 2700-2707.
828 48. Bolstad, B. (2001) Probe Level Quantile Normalization of High Density Oligonucleotide
829 Array Data.

49. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, **5**, R80.
50. Lin, M.F., Jungreis, I. and Kellis, M. (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, **27**, i275-282.
51. Goecks, J., Nekrutenko, A. and Taylor, J. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, **11**, R86.
52. Sun, K., Chen, X., Jiang, P., Song, X., Wang, H. and Sun, H. (2013) iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC genomics*, **14 Suppl 2**, S7.
53. Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.P. and Li, W. (2013) CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic acids research*, **41**, e74.
54. Kong, L., Zhang, Y., Ye, Z.Q., Liu, X.Q., Zhao, S.Q., Wei, L. and Gao, G. (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic acids research*, **35**, W345-349.
55. Gascoigne, D.K., Cheetham, S.W., Cattenoz, P.B., Clark, M.B., Amaral, P.P., Taft, R.J., Wilhelm, D., Dinger, M.E. and Mattick, J.S. (2012) Pinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes. *Bioinformatics*, **28**, 3042-3050.
56. Knowles, D.G., Roder, M., Merkel, A. and Guigo, R. (2013) Grape RNA-Seq analysis pipeline environment. *Bioinformatics*, **29**, 614-621.
57. Li, J.H., Liu, S., Zhou, H., Qu, L.H. and Yang, J.H. (2014) starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic acids research*, **42**, D92-97.
58. Lorenz, R., Bernhart, S.H., Honer Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F. and Hofacker, I.L. (2011) ViennaRNA Package 2.0. *Algorithms for molecular biology : AMB*, **6**, 26.

Figure Legends

Figure1: Discovery and classification of ribosome-associated lncRNAs.
(A) Numbers of Gencode v7 lncRNA genes filtered by protein-coding prediction methods. Genes (and all their constituent transcripts) having at least one transcript identified as protein-coding by at least one method were designated “Potential Protein Coding”. Remaining genes with no evidence for protein coding potential were defined as “Filtered LncRNAs”. (B) Outline of the subcellular mapping of K562 lncRNA by polysome profiling and microarray hybridisation. (C) Definition of the pooled fractions from sucrose ultracentrifugation used in this study. (D) Summary of the numbers of genes and transcripts classified in subcellular fractions. (E) Heatmaps show the relative microarray intensity measured for each RNA sample. The colour scale runs from blue (low detection) through white to red (high detection). “Protein coding” refers to the 2796 probes for protein-coding mRNAs included on the microarray, “Known lncRNAs” are those filtered transcripts that also belong to the lncRNAdb database(23).

Figure 2: Validation of selected ribosome-associated lncRNA candidates.
(A) We individually validated nine predicted ribosome-associated lncRNAs in independent ribosome profile experiments. In each case, two replicate experiments each were carried out with control K562 (red) and cells treated with puromycin (blue), for three distinct RNA fractions: (from left to right) free cytoplasmic, light polysomal, heavy polysomal. RNA levels are normalized to absolute levels of an RNA spiked into equal volumes of RNA sample. The first four panels represent protein coding mRNAs. Transcript IDs and classifications are shown above each panel. (B) Genomic map of ENST00000423918, a ribosome-associated transcript validated in this way.

Figure 3: Fluorescence in situ hybridisation of ribosomal lncRNA in HeLa. Left Panel: DAPI staining of DNA; Middle: FISH probe; Right: merged. The actively translated housekeeping mRNA GAPDH was tested as a positive control for cytoplasmic localisation.

Figure 4: Ribosomal and cytoplasmic lncRNA are under purifying selection. Cumulative distribution of the mean PhastCons nucleotide-level conservation for the exons of the indicated transcript classes. PhastCons scores for ancestral repeats regions are also included to represent neutral evolutionary rates.

Figure 5: Intra- and Sub-cellular expression of ribosomal lncRNAs. (A) Expression in K562 whole cell by RNAseq. (B) Percent of transcripts having ubiquitous expression across human tissues defined by Human Body Map RNAseq. (B) Percent of ubiquitously expressed transcripts in each class. (C) Log2 cytoplasmic/nuclear RPKM ratios calculated from ENCODE RNAseq for indicated RNAs in K562 (whole cell, polyA+). For potential protein coding transcripts and mRNAs, data is only shown for detected transcripts. Numbers indicate median value. (D) Subcellular localisation of lncRNA amongst different cell lines. Colours reflect median cytoplasmic / nuclear RPKM values.

Figure 6: Ribosomal lncRNAs have mRNA-like 5' ends. (A) The pseudo 5' UTR was defined to be the distance from the start to the first AUG trinucleotide (top row). As a control, we calculated the same measure on the antisense strand (bottom row). Shown is the distribution of these lengths for each set of transcripts - protein coding mRNA (left), followed by cytoplasmic lncRNA classes. The red line indicates the mean value. P-values are for comparison of sense and antisense distributions using the Kolmogorov-Smirnov test. (B) Capping efficiency positively correlates with Light Polysome localisation of lncRNA. We defined every transcript to be capped if it has a K562 cytoplasmic polyA+ CAGE tag within 100bp upstream or downstream of its transcription start site. LncRNA were binned according to their relative enrichment in each of the three cytoplasmic fractions (x axis). In each bin, the percent of capped transcripts is shown in the y axis. Logistic regression was used to assess the relationship between these variables.

Figure 7: Transposable element composition of lncRNAs. (A) The fraction of each transcript covered by annotated repeat sequence from RepeatMasker. (B) The heatmap

shows the normalised frequency of insertion for RepeatMasker-defined classes, ie the number of insertions per class divided by the length of each transcript. (C) As in (B) but showing data only for MLT-type repeats. (D) The repeat composition of a selection of Free Cytoplasmic, MLT-containing lncRNAs. The direction of the arrows indicates the annotated strand of the repeat with respect to the lncRNA. The colours represent the repeat class. (E) As in (B), except showing data for HeLa derived from ribosome footprinting experiments(36). For practical reasons, the lncRNAs are divided into three classes, see Materials and Methods for more details.

Figure 8: Cis- and Trans-antisense lncRNA-mRNA pairs and ribosomal recruitment.

(A) Cartoon illustrating the definition of cis- and trans-antisense lncRNA-mRNA pairs. Red boxes indicate regions of opposite-strand homology. (B) The percent of each subcellular lncRNA class defined as exonic-antisense (cis-antisense) to a protein coding gene. (C) Example of a cis-antisense lncRNA-mRNA pair. ENST00000529247 (forward strand) is a heavy polysomal lncRNA transcribed antisense to the EEF1D gene (reverse strand), encoding a subunit of the translation elongation factor 1 complex. (D) Whole cell K562 polyA+ steady state levels of mRNAs that are antisense to lncRNA in the indicated subcellular classes. (E) The percent of lncRNA (blue bars) or size-matched random genomic fragments, having an antisense trans homology match to at least one mRNA.

Figure 9: Changes in lncRNA stability in response to ribosome stalling.

Bars represent mean detection in cells treated with cyclohexamide (CHX) for 6 hours and control cells (0h). Experiments were performed with three biological replicates. Bars show mean and standard deviation. Statistical significance was calculated by one-sided t-test (* $P<0.05$, ** $P<0.01$, *** $P<0.001$)

Figure 10: A model of lncRNA targeting within the cytoplasm.

Supplementary Data Files

Table S1: Gencode v7 lncRNA classification. Rows represent lncRNA transcript from Gencode v7.

Columns:

TransID: ENST ID for transcripts.

GeneID: ENSG ID for the corresponding gene.

Chr: Chromosome.

Trans_Start: Transcript start position.

Trans_End: Transcript end position.

Strand

Cellular_Localization: Classification of the transcripts into 5 different categories: 1: Present in cytoplasm (from polysome profiling experiment, K562 cell line); 2: Present in nucleus (from ENCODE nucleus RNAseq data, K562 cell line), but not in cytoplasm (polysome profiling experiment); 3: Not present in cytoplasm (polysome profiling experiment) nor in nucleus (ENCODE nucleus RNAseq data, K562 cell line); 4: Transcripts classified as potential protein coding transcripts; 5: Discarded transcripts. See Materials and Methods for details.

FreeC_intensity: Log2 intensity value for Free C. condition (NA if not present in this condition).

LightP_intensity: Log2 intensity value for light P. condition (NA if not present in this condition).

HeavyP_intensity: Log2 intensity value for Heavy P. condition (NA if not present in this condition).

WholeC_intensity:Log2 intensity value for whole cytoplasmic fraction (NA if not present in this condition).

Ribosomal_Classification: Classification for transcripts present in the cytoplasm: 1: Free C.; 2: Light P.; 3: Heavy P.

CPAT: CPAT score.

PhyloCSF: PhyloCSF score.

CPC: CPC score.

MS: information about presence of peptides in Mass Spectrometry analysis for this transcripts: 0: No peptide associated; 1: Peptide associated.

Table S2: Correlation of gene expression quantification between microarray K562 cytoplasmic measurements, and ENCODE RNAseq data from K562 cellular fractions (19).

Table S3: Heavy Polysome mRNAs are most actively translated. Shown are the numbers of ENCODE K562 mass spectrometry tags originating from ribosome-profiled mRNAs.

Table S4: Small peptides originating from lncRNA. Shown are the numbers of known small peptides discovered by mass spectrometry that map to Gencode v7 lncRNA (43).

Figure S1: Comparison of lncRNA microarray and RNAseq quantifications. Steady state values for K562 cytoplasmic RNA were analysed. RNAseq data was obtained from ENCODE. Only transcripts detected in both experiments are shown.

Figure S2: Mean tissue expression across 16 human tissues from Human Body Map. In the cases of potential protein coding and protein coding transcripts, data is only shown for those transcripts detected in K562.

Figure S3: Protein interactions related to subcellular compartmentalisation. Heatmap depicts lncRNA genes that interact with the indicated proteins, as defined by the Starbase database (57). Interactions of “Low stringency” were used in all cases. The colour scale indicates the percent difference of the actual to expected number of overlaps. The rows show lncRNA gene sets assigned to the four subcellular lncRNA categories, and the columns represent various proteins for which CLIPseq binding sites were analysed. Arrows indicate the reported subcellular localisation of the protein, identified by manual curation of the literature. We found a number of cases where the

localisation of lncRNA corresponded with the known distribution of the protein to which they are bound: Nuclear-associated lncRNAs showed elevated binding to nuclear-localised proteins, including hnRNPC (P=0.007, Fisher's exact test), U2AF65 (P=0.002), and eIFAIII (P=0.0002). In contrast, lncRNAs bound by the cytoplasmic acting IGFBP1 were significantly enriched in the Light Polysomal and free cytoplasmic fractions (P=0.007). Light Polysomal lncRNAs are enriched for binding by the TAF15 protein (P=0.033). A general depletion of Heavy Polysomal lncRNA was observed in the protein binding data.

Figure S4: Association of ORF length with polysome density for protein coding transcripts. Shown are histograms for % coverage of transcripts by their longest ORF. Top row: sense strand ORFs; Bottom row: antisense strand ORFs (control). Data are shown for mRNAs included in microarray design and classified by ribosomal occupancy. P values compare sense / antisense distributions in each case. Red line indicates mean ORF coverage percentage. Note the difference in mean value between Heavy and Light Polysomal means.

Figure S5: Association of ORF length with polysome density for lncRNA transcripts. Shown are histograms for % coverage of transcripts by their longest ORF. Top row: sense strand ORFs; Bottom row: antisense strand ORFs (control). Data are shown for all mRNAs together, for comparison. P values compare sense / antisense distributions in each case. Red line indicates mean ORF coverage percentage. Note the lack of difference in mean value between Heavy and Light Polysomal means.

Figure S6: Gene structure characteristics of lncRNAs. (A) Exon length distributions. (B) Intron length distributions. (C) Exon number per transcript. (D) Mature (processed) transcript length.

Figure S7: GC content of coding and noncoding transcripts.

Figure S8: Polyadenylation of lncRNAs. Using ENCODE data (19), we calculated the ratio of RPKM for PolyA+ / PolyA- K562 cytoplasmic RNAseq. RPKM values were averaged across the two available technical replicates, and only transcripts with non-zero mean values in both RNA samples retained. No statistically significant differences were found between Free Cytoplasmic lncRNAs and either of the ribosomal groups using either the Kolmogorov-Smirnov or Wilcoxon tests.

Figure S9: Splicing efficiency of lncRNAs. Using ENCODE data, we calculated RPKM values separately for the exons and introns of all lncRNAs. Shown are the log10 ratios of exon/intron values for all sets of transcripts. No statistically significant differences were found between Free Cytoplasmic lncRNAs and either of the ribosomal groups using either the Kolmogorov-Smirnov or Wilcoxon tests.

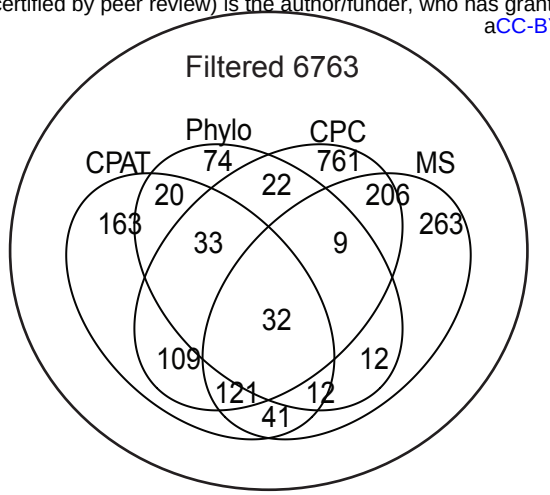
Figure S10: Comparison of 5' RNA folding energy. Using the Vienna RNAfold programme (58) with default settings, we estimated the free energy of folding of the first 50nt of lncRNA and mRNA. While mRNA have more stable folding on average than expressed lncRNA (P=2.2e-16, Wilcoxon test), we could find no difference between either Heavy Polysomal (P=0.8) or Light Polysomal (P=0.7) and Free Cytoplasmic lncRNAs.

Figure S11: ERVL-MaLR insertion length distributions.

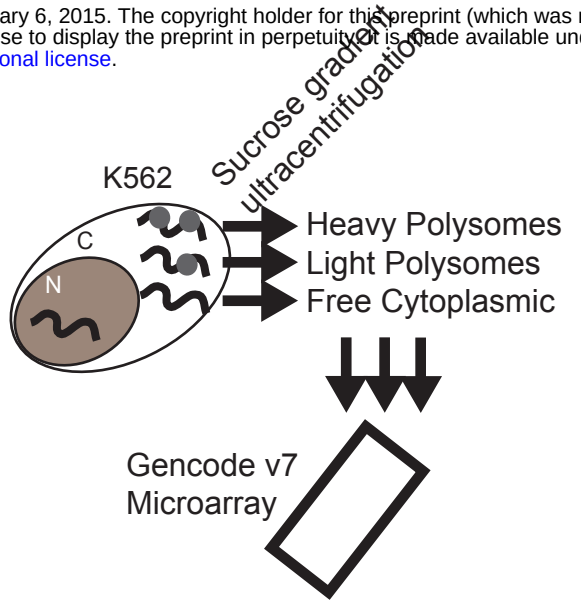
Figure S12: The association between sense-intronic lncRNAs and nuclear localisation. Shown is the percent of transcripts in each subcellular category that are located within the intron of a same-strand protein coding gene.

Figure 1

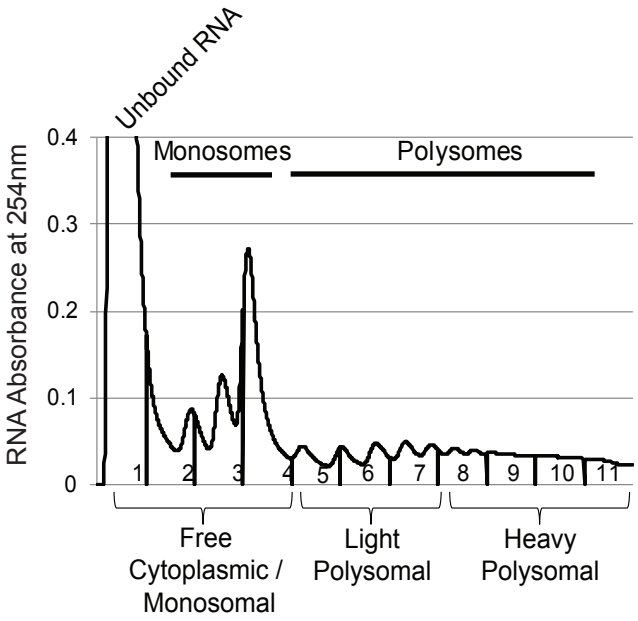
A



B



C



D

	Transcripts	% cytoplasmic transcripts	Genes
Heavy Polysome	204	22.1	177
Light Polysome	373	40.4	248
Free Cytoplasmic	347	37.5	248
Nucleus	292		255
Not Present	7754		6033
Potential Coding	4415		1878
Discarded	1169		521

E

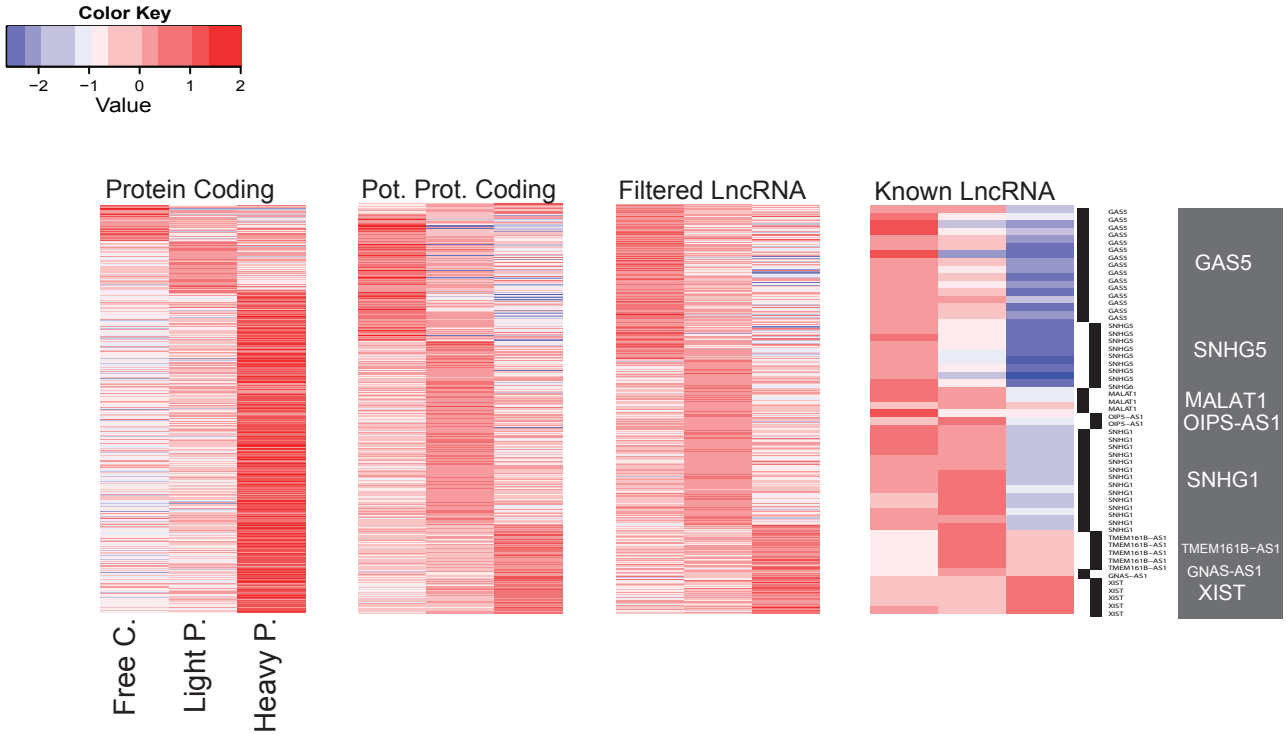
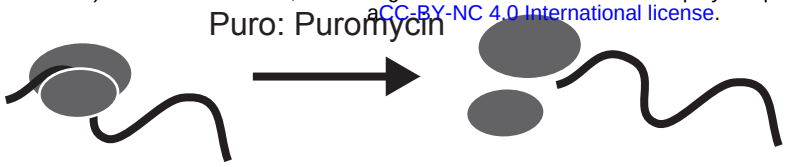
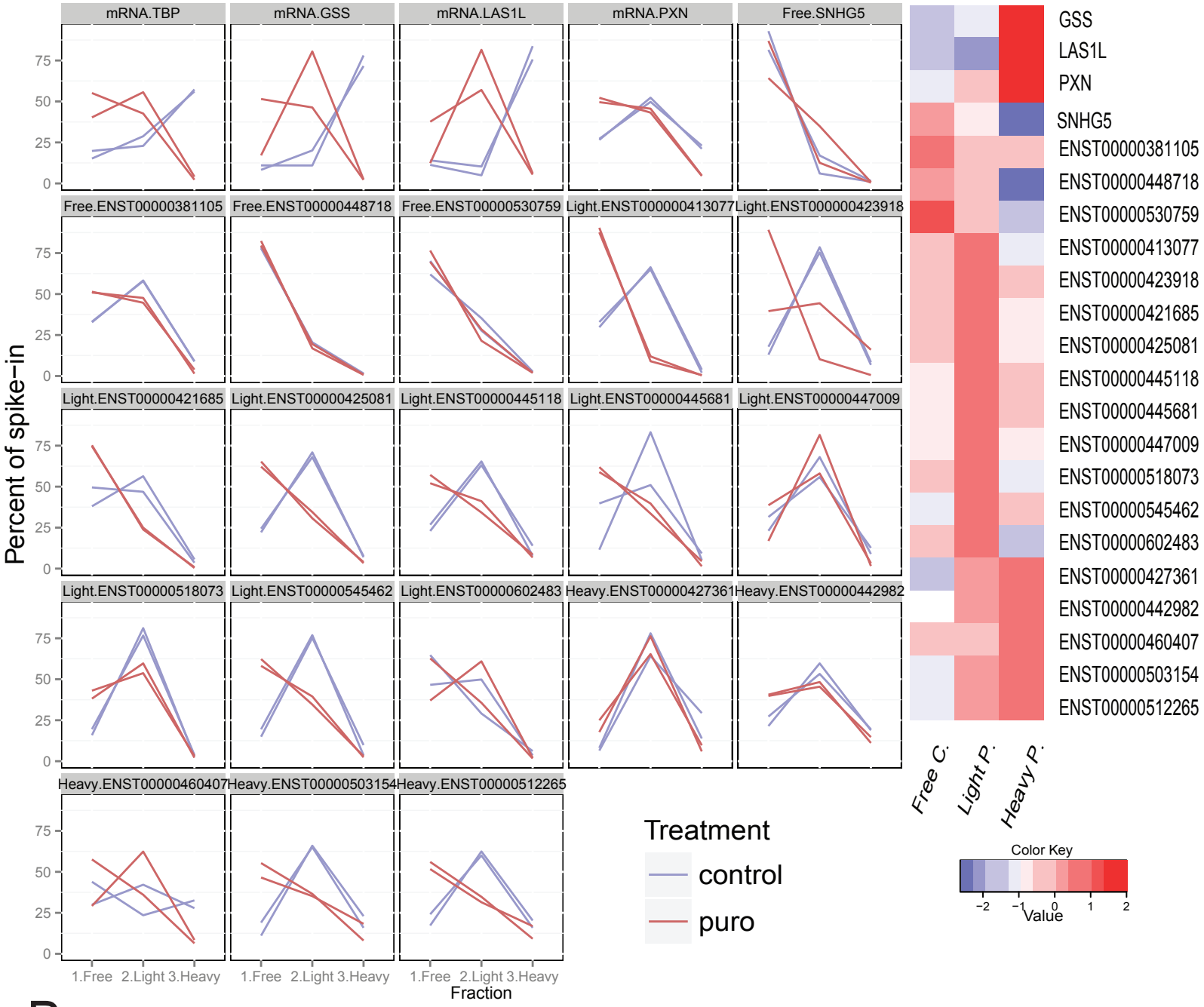


Figure 2

A





B

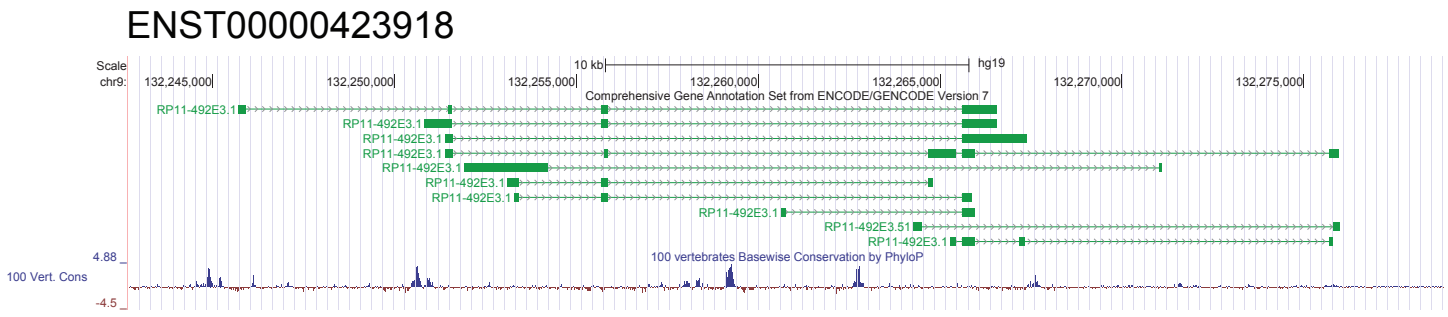
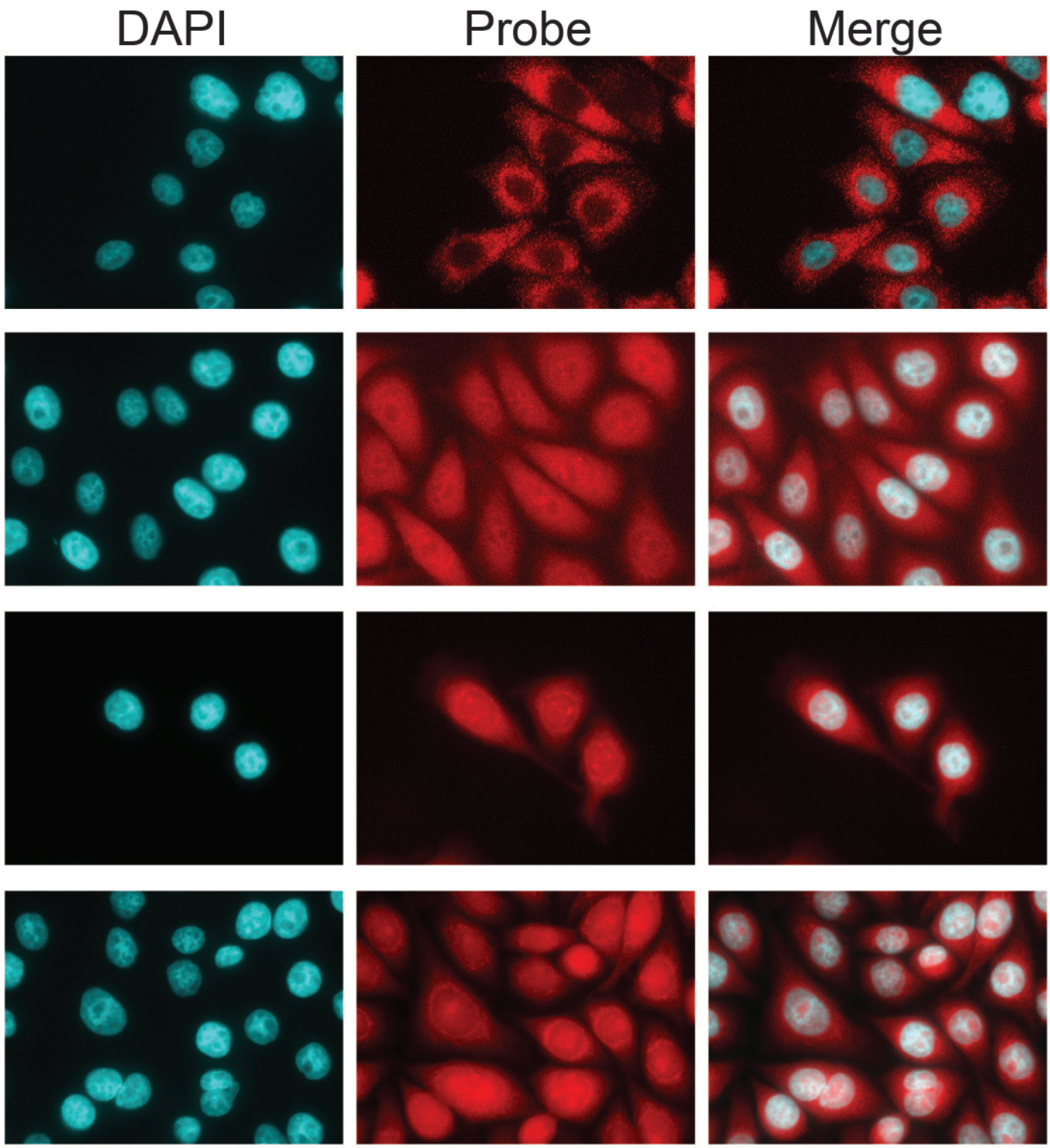


Figure 2



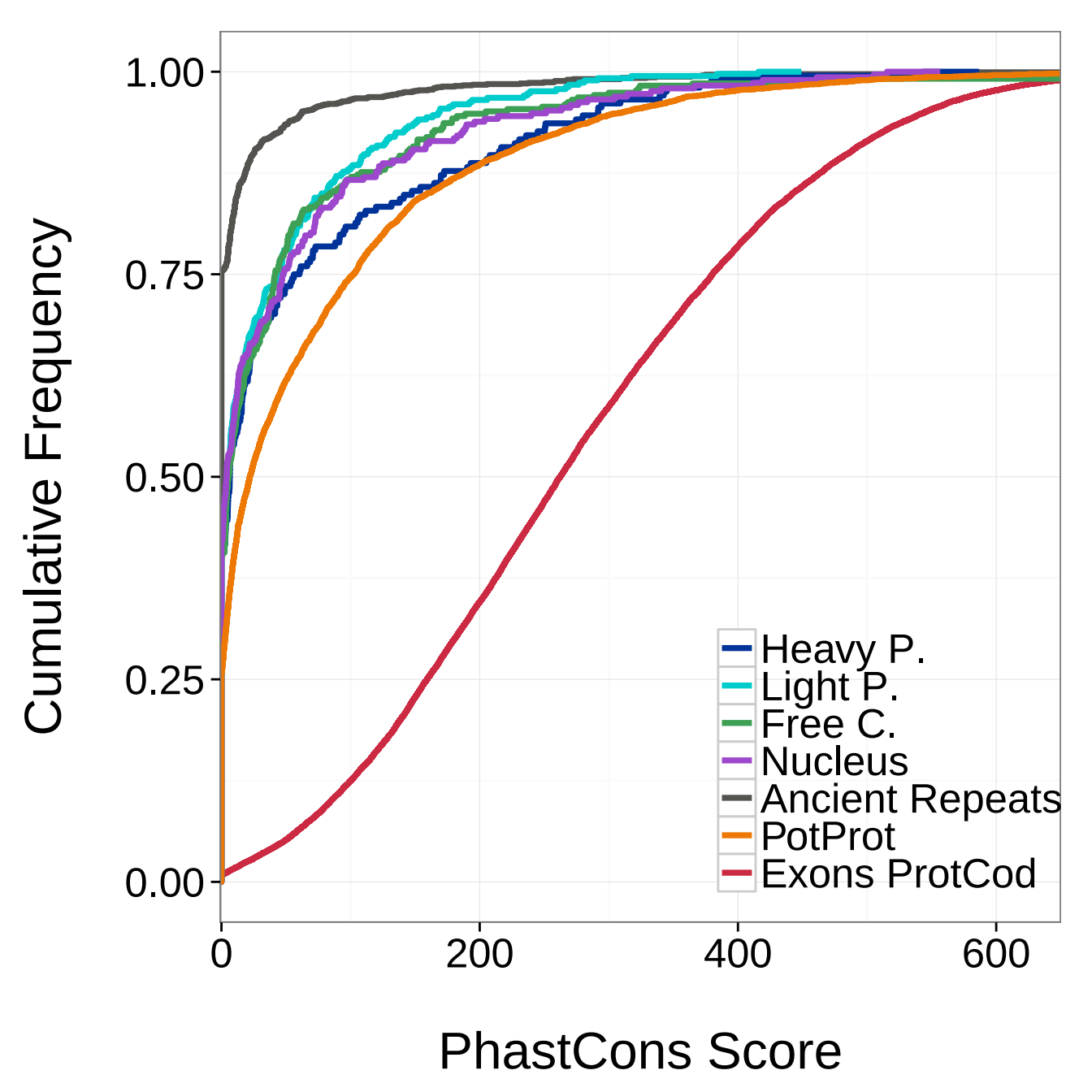

Cumulative Frequency
1.00
0.75
0.50
0.25
0.00
0
200
400
600
PhastCons Score
Heavy P.
Light P.
Free C.
Nucleus
Ancient Repeats
PotProt
Exons ProtCod

Figure 5

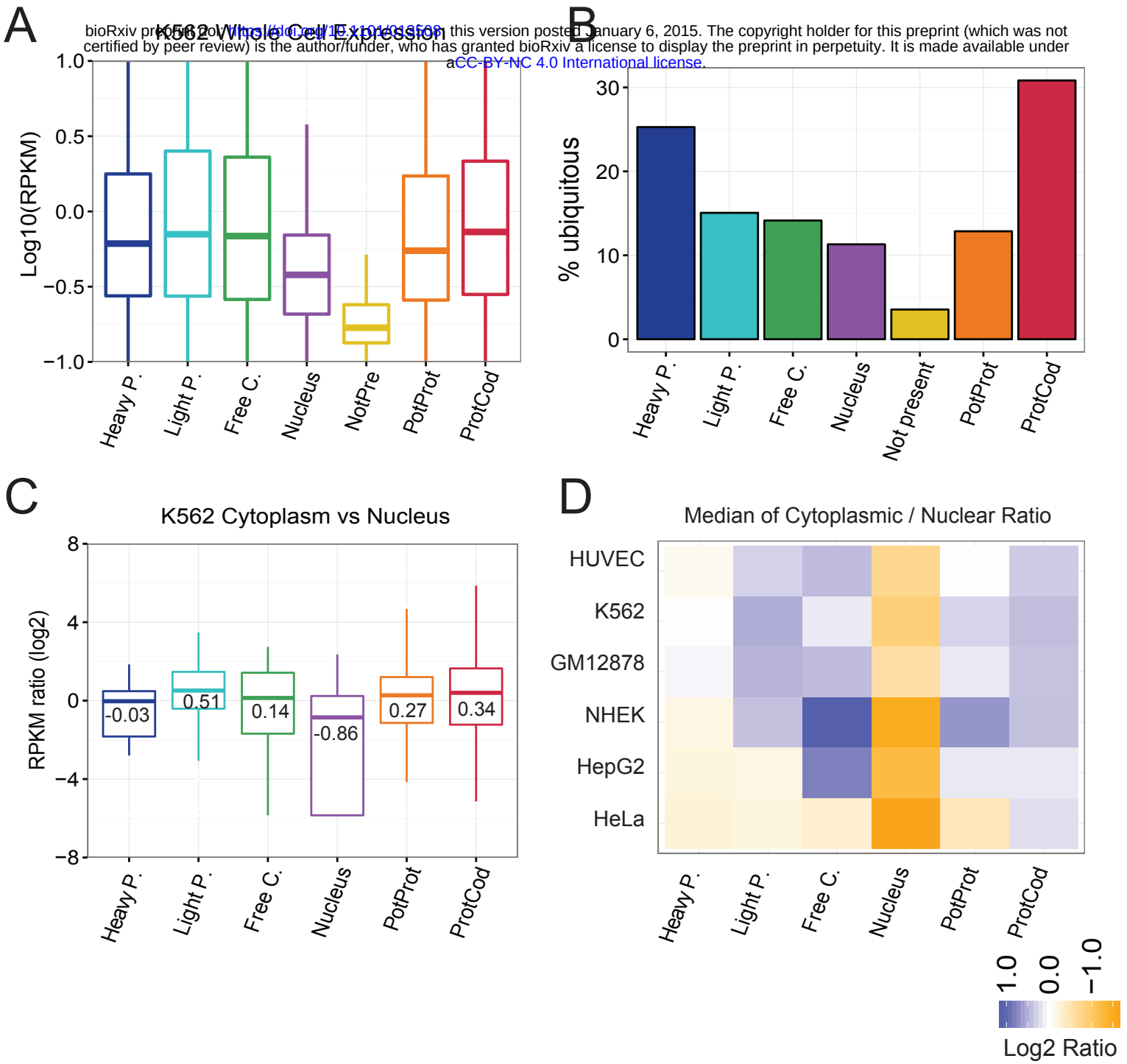
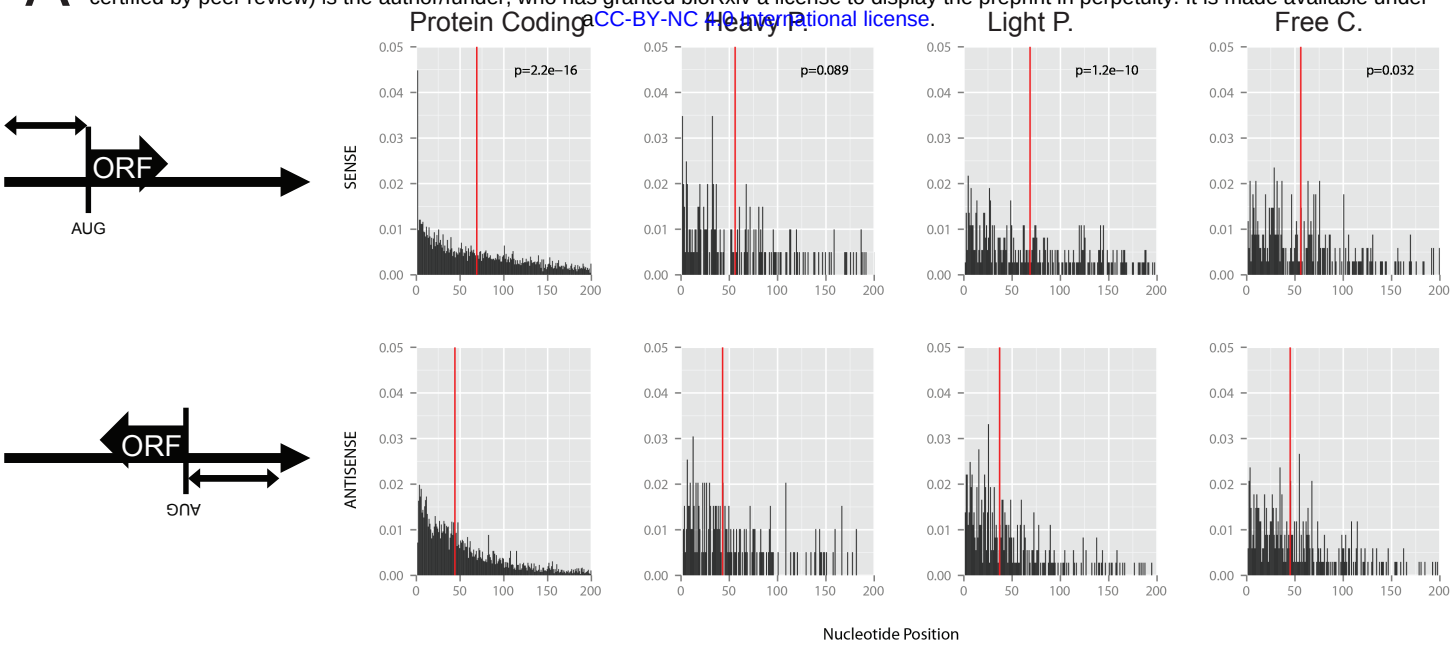


Figure 6

A



B

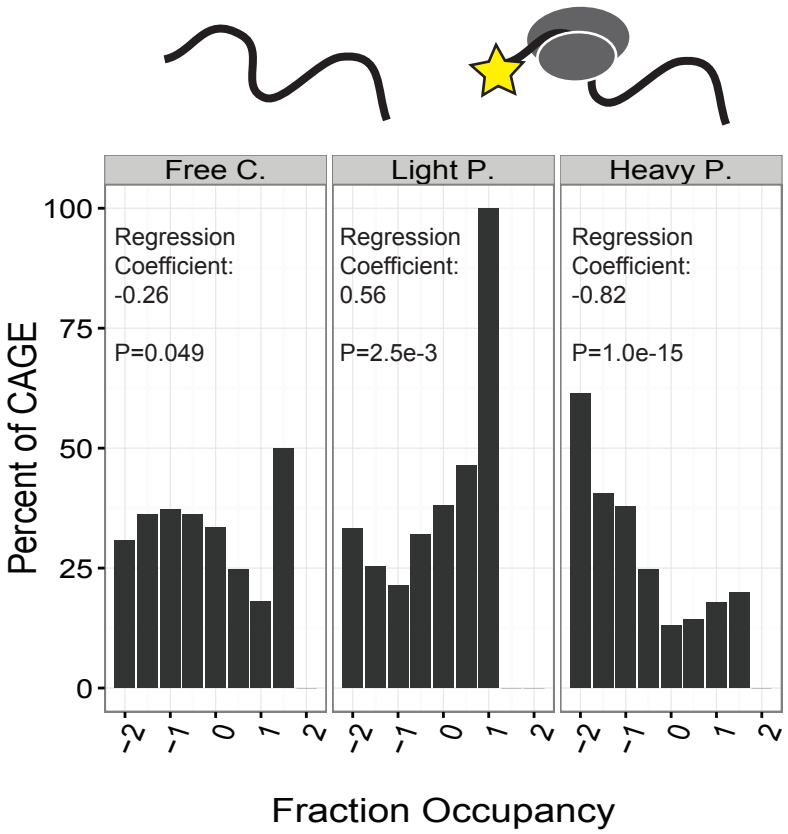


Figure 7

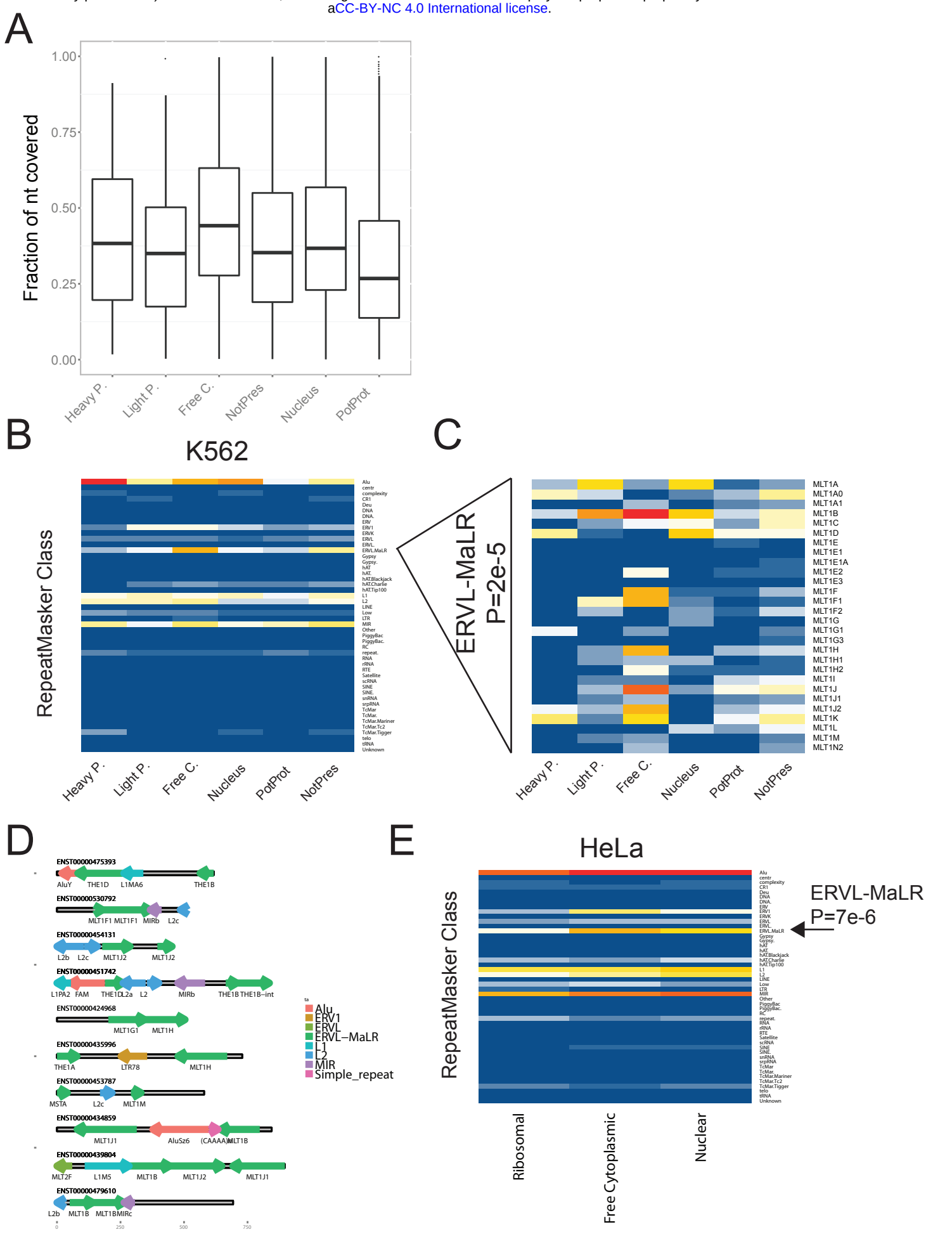
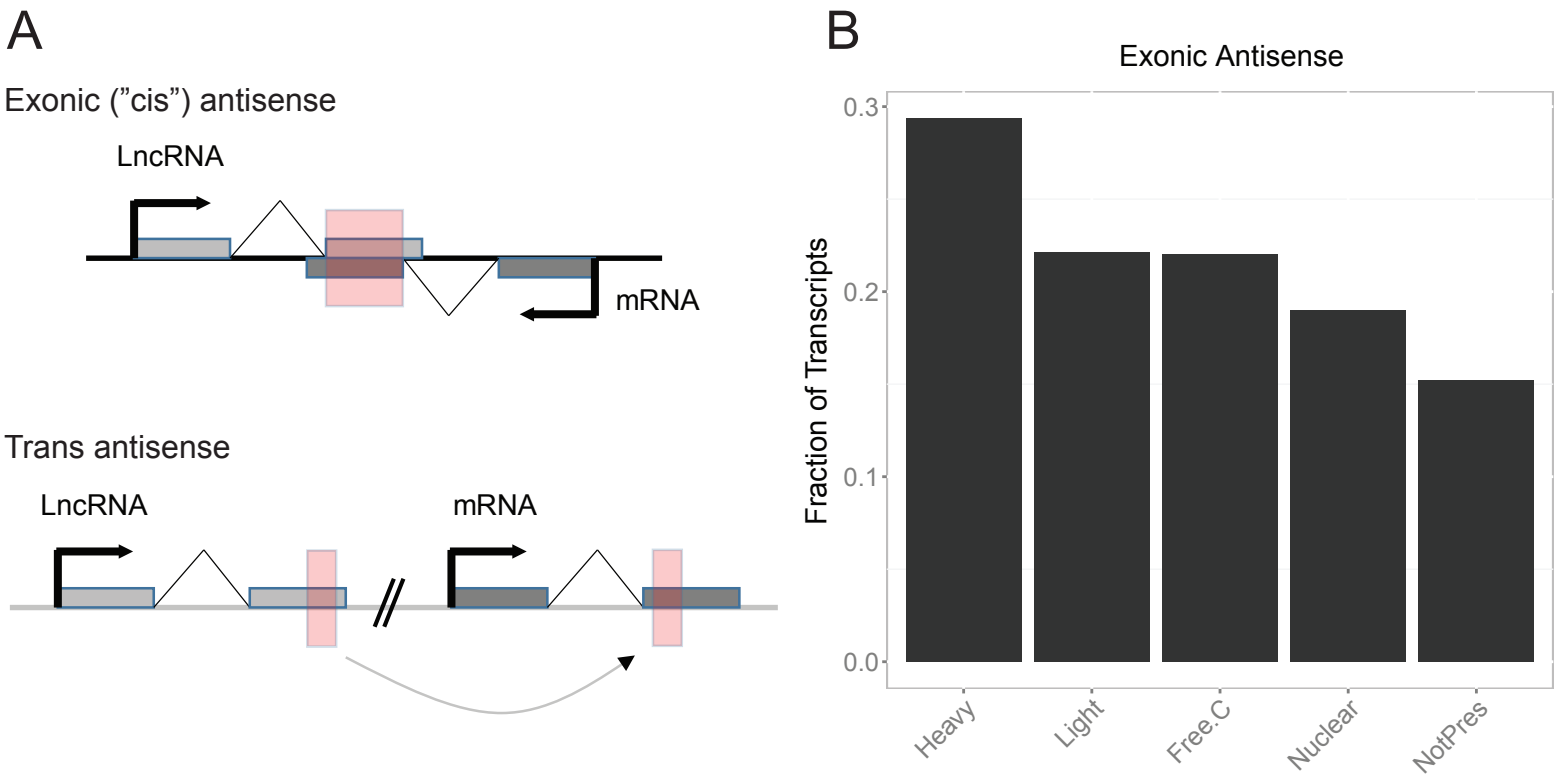
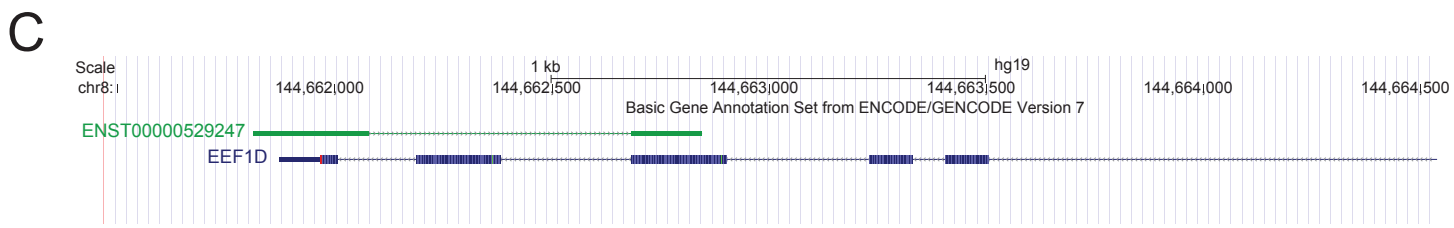
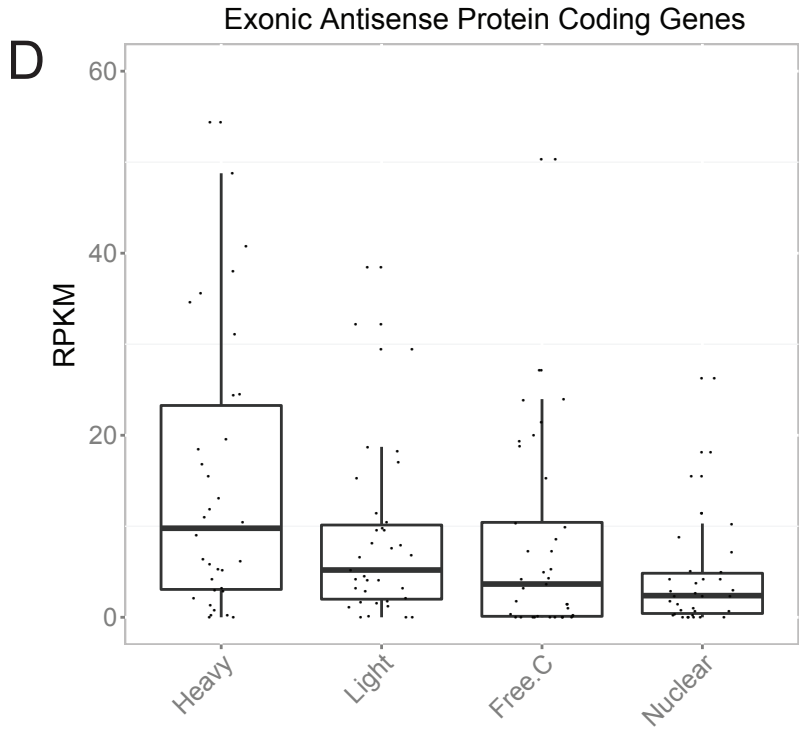


Figure 8







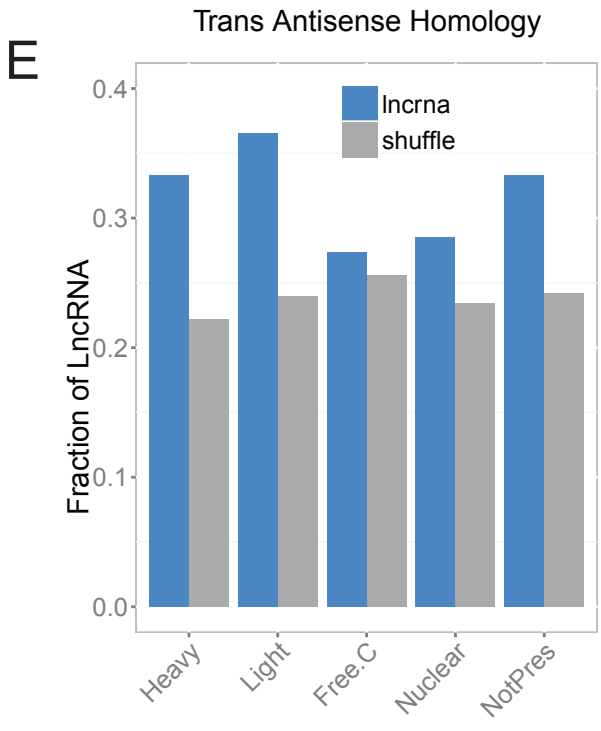


Figure 9

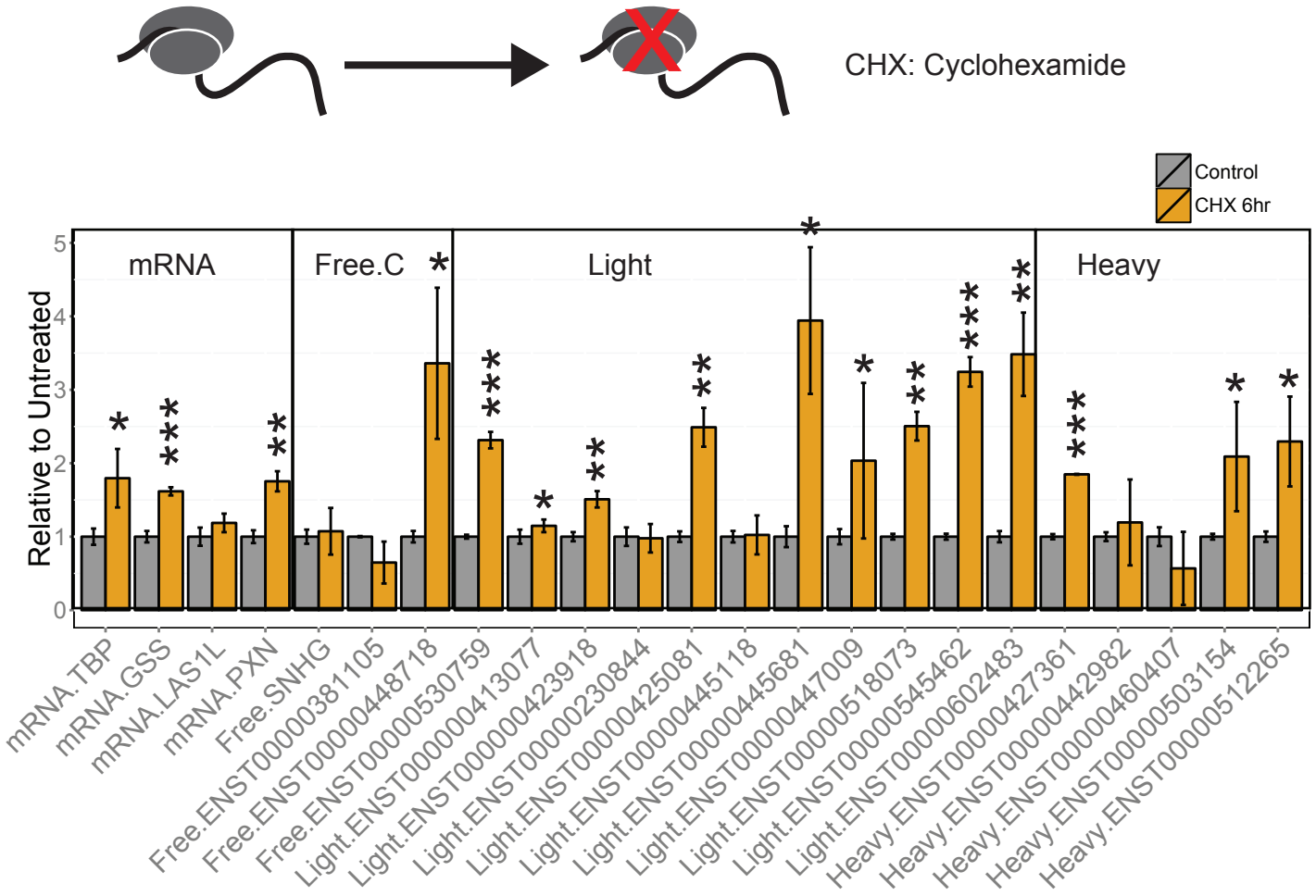


Figure 10