A ubiquitous GC content signature underlies multimodal mRNA regulation by DDX3X

Ziad Jowhar[1,10,11,#], Albert Xu[1,10,11,#], Srivats Venkataramanan[1], Mariah L Hoye[2], Debra L Silver[2,3,4,5,6] Stephen N Floor[1,7*] and Lorenzo Calviello[8,9*]

[1]Department of Cell and Tissue Biology, UCSF, San Francisco, United States
[2]Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, United States
[3]Department of Cell Biology, Duke University Medical Center, Durham, United States
[4]Duke Regeneration Center, Duke University Medical Center, Durham, United States
[5]Department of Neurobiology, Duke University Medical Center, Durham, United States
[6]Duke Institute for Brain Sciences, Duke University Medical Center, Durham, United States
[7]Helen Diller Family Comprehensive Cancer Center, San Francisco, United States
[8]Functional Genomics Research Centre, Human Technopole, Milan, Italy
[9]Computational Biology Research Centre, Human Technopole, Milan, Italy
[10]Medical Scientist Training Program, University of California, San Francisco, San Francisco, CA 94158, USA
[11]Biomedical Sciences Graduate Program, University of California, San Francisco, San Francisco, CA 94158, USA

[#]These Authors contributed equally to this work
[*]Correspondence: Lorenzo.Calviello@fht.org, Stephen.Floor@ucsf.edu

Abstract

The road from transcription to protein synthesis is paved with many obstacles, allowing for several modes of post-transcriptional regulation of gene expression. A fundamental player in mRNA biology is DDX3X, an RNA binding protein that canonically regulates mRNA translation. By monitoring dynamics of mRNA abundance and translation following DDX3X depletion, we observe stabilization of translationally suppressed mRNAs. We use interpretable statistical learning models to uncover GC content in the coding sequence as the major feature underlying RNA stabilization. This result corroborates GC content-related mRNA regulation detectable in other studies, including hundreds of ENCODE datasets and recent work focusing on mRNA dynamics in the cell cycle. We provide further evidence for mRNA stabilization by detailed analysis of RNA-seq profiles in hundreds of samples, including a *Ddx3x* conditional knockout mouse model exhibiting cell cycle and neurogenesis defects. Our study identifies a ubiquitous feature underlying mRNA regulation and highlights the importance of quantifying multiple steps of the gene expression cascade, where RNA abundance and protein production are often uncoupled.

Introduction

The cytoplasmic fate of RNA molecules is impacted their subcellular localization, RNA binding partners, and engagement with the ribosomal machinery. These aspects are strongly interconnected[1], which poses a great challenge, as it increases the number of variables and experimental approaches needed to answer many questions in mRNA biology. To this end, many protocols couple biochemical isolation, or metabolic labeling, of RNA with high throughput sequencing technologies, thus providing a snapshot of the transcriptome at specific stages of the mRNA life cycle, with high throughput and sensitivity. For example, high-throughput sequencing protocols, when coupled to ribosome isolation, such as in Ribo-seq[2], metabolic labeling strategies in SLAM-seq[3], immunoprecipitation of RNA binding proteins (RBP) as in CLIP-seq[4] and many others, have shed light on many regulatory mechanisms pertaining to different aspects of post-transcriptional gene regulation.

DDX3X is a multifunctional RNA helicase that is highly expressed in many tissues and able to unwind structured RNA to influence cytoplasmic post-transcriptional gene regulation[5]. Together with its ability to bind initiating ribosomes, DDX3X has been often described as a translation regulator, specifically promoting translation of RNA with structured 5'UTRs[6,7]. However, as mentioned above, cytoplasmic processes like translation or mRNA decay are intertwined, and connection between the two processes encompass different molecular mechanisms, such as mRNA surveillance mechanisms like nonsense-mediated decay (NMD)[8], ribosome-collision dependent mRNA cleavage[9], and others. In order to understand when and how such processes are coupled, it is important to study the dynamics of such mechanisms. For instance, it has been proposed that miRNA can first trigger translation suppression and then mRNA deadenylation and decapping leading to RNA degradation[10].

Mutations in *DDX3X* are associated with a variety of human diseases including cancers and developmental delay[11]. Variant types are disease selective in *DDX3X*, with cancers ranging from primarily loss-of-function alleles in NK-TCL and other blood cancers to nearly exclusively missense variants in medulloblastoma[11]. In *DDX3X* syndrome, missense variants are phenotypically more severe than loss-of-function. Previously, we used functional genomics approaches to identify mechanistic differences between depletion of DDX3X and expression of missense variants[7]. We found that *DDX3X* missense variants predominantly affect ribosome occupancy while DDX3X depletion impacts both ribosome occupancy and RNA levels. However, it is unclear whether the changes in RNA levels constituted a cellular response to translation suppression, often described as "buffering"[12].

mRNA regulation has been linked to neurogenesis during development, where multiple RNA binding factors, including DDX3X, ensure correct protein synthesis as cells transition between different fates and states[13]. To that end, it is important to think about the dynamics of gene expression, as complex dynamics of cell proliferation and differentiation ensure correct developmental patterning.

In order to access such complex interplays of a multitude of factors which shape gene expression, large-scale consortia have provided a great resource for investigations into gene regulation. While historically devoted to promoting investigation into transcriptional regulation, recent efforts started to provide precious information into post-transcriptional mechanisms, with hundreds of RBPs profiled in terms of both binding and function, by means of CLIP-seq, and knockdown followed by RNA-seq[14]. As in biology many molecular processes

are interconnected, large-scale datasets and data amenable to re-analysis are at the very heart of many research efforts[15].
Here, we identify how inactivation of DDX3X evolves over time to lead to acute and long-term changes to post-transcriptional gene regulation. We here employ different analytical approaches applied to newly generated experimental data and many previously published studies related to mRNA regulation, to show that GC content is associated with mRNA stability changes following DDX3X depletion. Our analyses indicate that this effect is widespread and is associated with cell cycle changes in mRNA regulation, including RNA stability. This further reinforces roles for DDX3X in RNA stability in addition to translation. Together, our work represents a significant advancement in the understanding of a fundamental regulator, which sits at the very heart of the gene expression cascade.

Results

Time-resolved gene expression regulation by DDX3X.

To characterize the dynamics of DDX3X-dependent changes in the gene expression cascade, we employed a previously validated auxin-degron system to efficiently deplete DDX3X protein in the HCT116 colorectal cancer cell line[16]. We profiled RNA levels and translation using RNA-seq and Ribo-seq along a time-course of DDX3X depletion, at 4, 8, 16, 24 and 48 hours after auxin or DMSO control treatment. (Figure 1A). Efficiency of DDX3X depletion, together with quality control and general statistics of the generated libraries, can be found in Supplementary Figure 1 and Supplementary Table 1. As expected, the number of differentially expressed genes increased along the time-course, with most changes supporting the role of DDX3X as a positive regulator of translation (Figure 1B). Changes in translation were negatively correlated with changes in mRNA levels, which together contributed to many changes in Translation Efficiency (TE), calculated using Ribo-seq changes given RNA-seq changes (Methods). At a closer look, we observed how “TE_down” mRNAs undergo translation suppression in the early time point after DDX3X depletion, with their mRNA levels increasing in the later time points (Figure 1C). The opposite behavior is observed for “TE_up” mRNAs, exhibiting higher ribosome occupancy first, and lower mRNA levels later. Such behavior was more evident when showing time-point specific changes and binning mRNAs in a 2D grid on the Ribo-seq/RNA-seq coordinate plane (Figure 1D, Methods), which highlighted a common regulatory mode, with early translation regulation followed by changes in mRNA levels.
This analysis shows the time-resolved dynamics of mRNA regulation by DDX3X, with hundreds of mRNAs changing in their steady-state levels albeit showing the opposite directionality in translation rates.

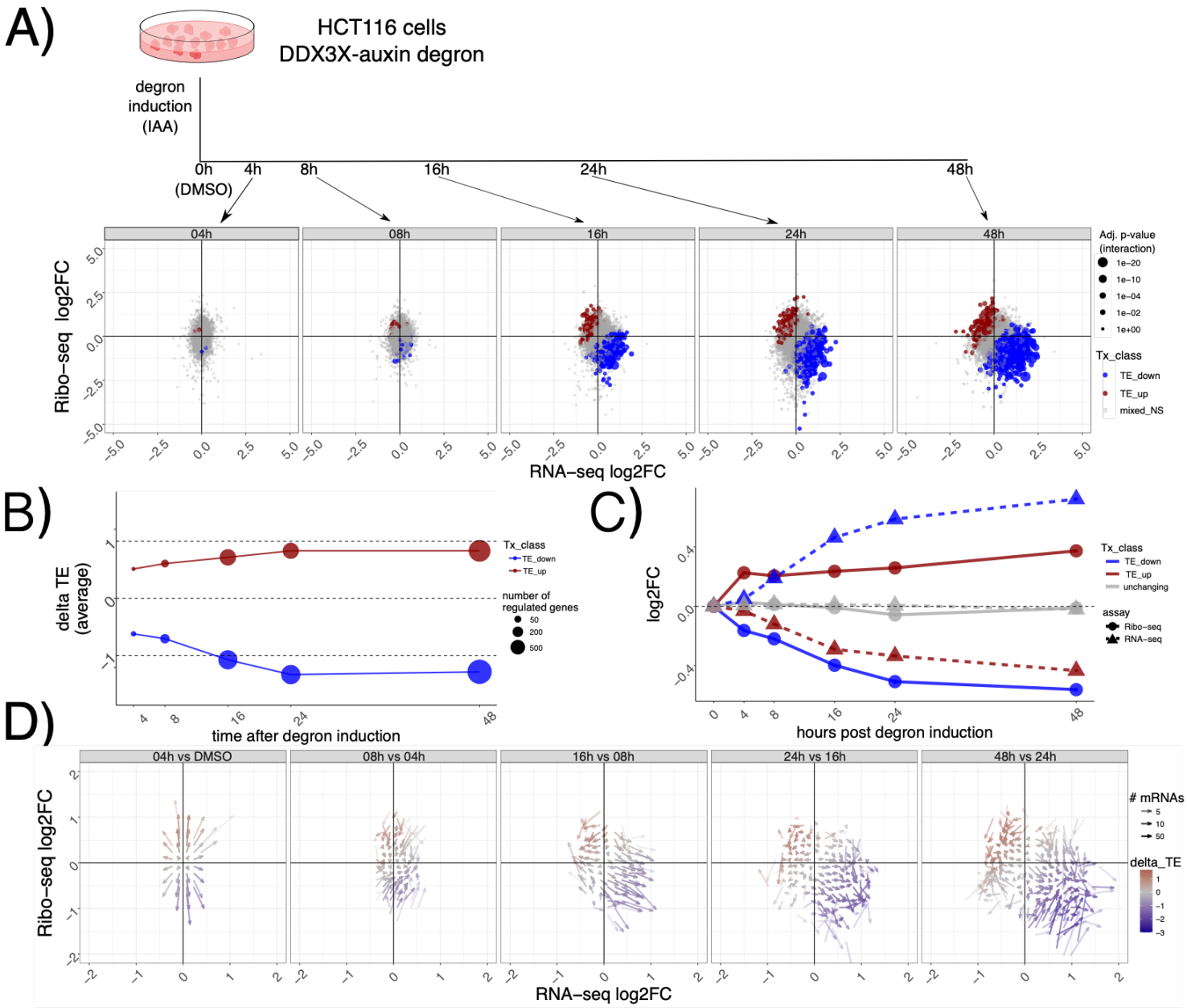

Figure 1: Dynamics of mRNA regulation by DDX3X

In A) a description of the experimental design. Below Ribo-seq and RNA-seq fold changes at different time points. Different regulated classes are shown in different colors. The size of the dots indicates the adjusted P-values for differential translation efficiency test (Methods). TE: translation efficiency, NS: not significant. In B) average delta TE values for each class along the time course. The size of the dots indicates the number of significantly changing mRNAs. C) progression along the time course for mRNA regulated 48h post degron induction. RNA-seq and Ribo-seq fold changes are shown at each time point. D) Differences in Ribo-seq or RNA-seq fold changes between each time point and the previous one, shown as a vector plot. Magnitude of changes shown as a color gradient, while transparency of the vectors indicates the number of mRNAs in each coordinate bin (Methods).

Translation suppression by DDX3X is coupled with mRNA stabilization.

Changes to transcript levels can result from changes in transcription rates or post-transcriptional regulation. To identify the relative contribution of different processes to RNA levels, we used our time-course dataset to calculate changes in transcription, processing and stability using INSPEcT[17]. Briefly, INSPEcT uses the proportion of intronic versus exonic reads to identify nascent vs. mature transcripts, and uses a system of ODEs to infer rates of RNA synthesis, processing and decay. Compared to non-regulated mRNAs, regulated mRNAs

showed modest changes in transcription rates, suggesting transcription changes are not the major contributor to RNA level changes following DDX3 depletion, In contrast, we found more pronounced changes in mRNA stability as evidenced by TE down transcripts (Figure 2A). As our initial RNA-seq protocol was not designed to capture pre-mRNA molecules, we validated our estimated mRNA stability changes by employing the 4sU metabolic labeling SLAM-seq protocol[3] in our degron system after 8 hours of DDX3X depletion, in a way to detect changes in mRNA stability at early time points. As expected, we observed a drop in T>C harboring reads after 24h of RNA labeling with 4sU, which reflects the decay rates of the nascent transcriptome (Supplementary Figure 2). We then applied SLAM-seq in our degron system, using DMSO as a control dataset. As shown in Figure 2B, after a labeling time of 24 hours, the percentage of reads harboring T>C mutations significantly changed for the regulated categories (Methods) after only 8 hours of degron induction, indicating differences in mRNA half-life upon DDX3X depletion. While the modest depth of our SLAM-seq dataset (Supplementary Figure 2) couldn't allow for more detailed insights on mRNA changes, it represented an important validation of mRNA stability regulation by DDX3X.

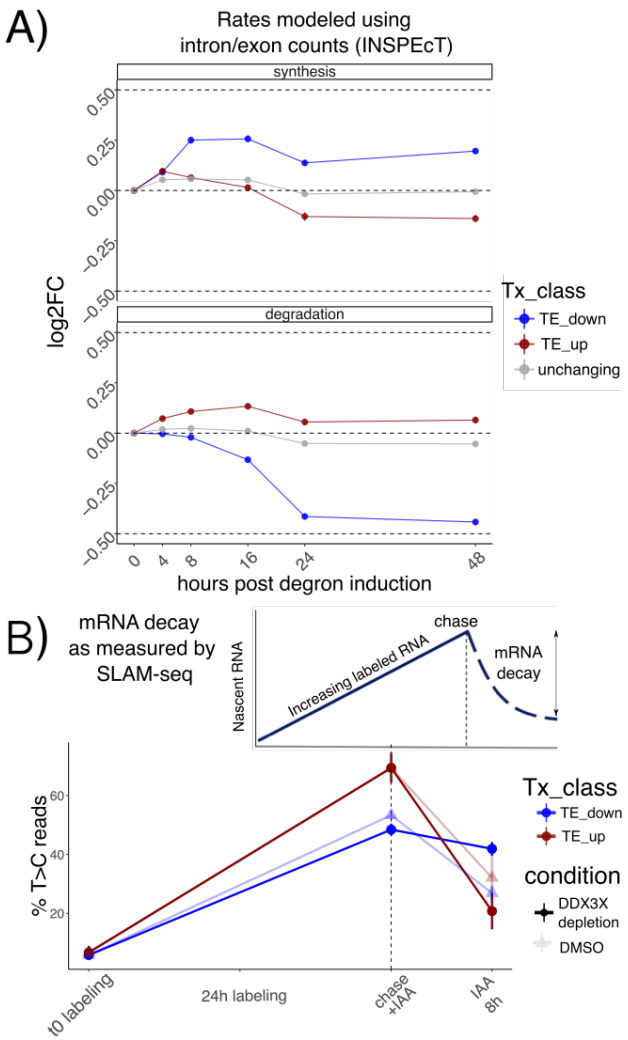

Figure 2: Stabilization of untranslated mRNAs

A) Synthesis and decay as inferred by INSPEcT: different regulated classes in different colors along the time course. Log2FC of estimated rates with respect to control are shown on the y axis. B) Schematic of a SLAM-seq experiment (above). Real data shown at the bottom: percentage of T>C-containing reads on the y axis after labeling and chase. DDX3X degron (using DMSO as a control) was triggered together with the chase reaction to monitor differences in decay rates upon DDX3X depletion.

By profiling ribosome occupancy, steady state transcript levels, and mRNA decay, this analysis shows that DDX3X depletion triggers multiple modes of post-transcriptional regulation, involving translation suppression and a subsequent wave of mRNA stabilization.

GC-rich coding sequences underlie mRNA regulation by DDX3X.

With hundreds of mRNAs post-transcriptionally regulated after DDX3X depletion, we aimed to identify specific features belonging to up- or downregulated targets. We therefore built regression models to quantitatively predict levels of TE changes (Methods, Supplementary Table 2). We used different biophysical properties of genes and mRNAs, (e.g. length and GC content) and several gene and transcript features (e.g. introns, 3'UTR, etc.., Methods) as features for a Random Forest regression model. Given the extensive literature on codon-mediated mRNA stability regulation, we added codon frequencies and previously validated codon optimality calculations[18]. Also, we added measured GC-content at 1st, 2nd or 3rd codon position, as it was recently shown to potentially play a role in mRNA stability regulation[19,20]. In addition, to pinpoint features predictive of mRNA stability changes rather than translation changes exclusively, we divided transcripts according to their position in the Ribo-seq/RNA-seq coordinate system, to capture mRNAs where changes between assays agreed or not (Figure 3A, Methods).

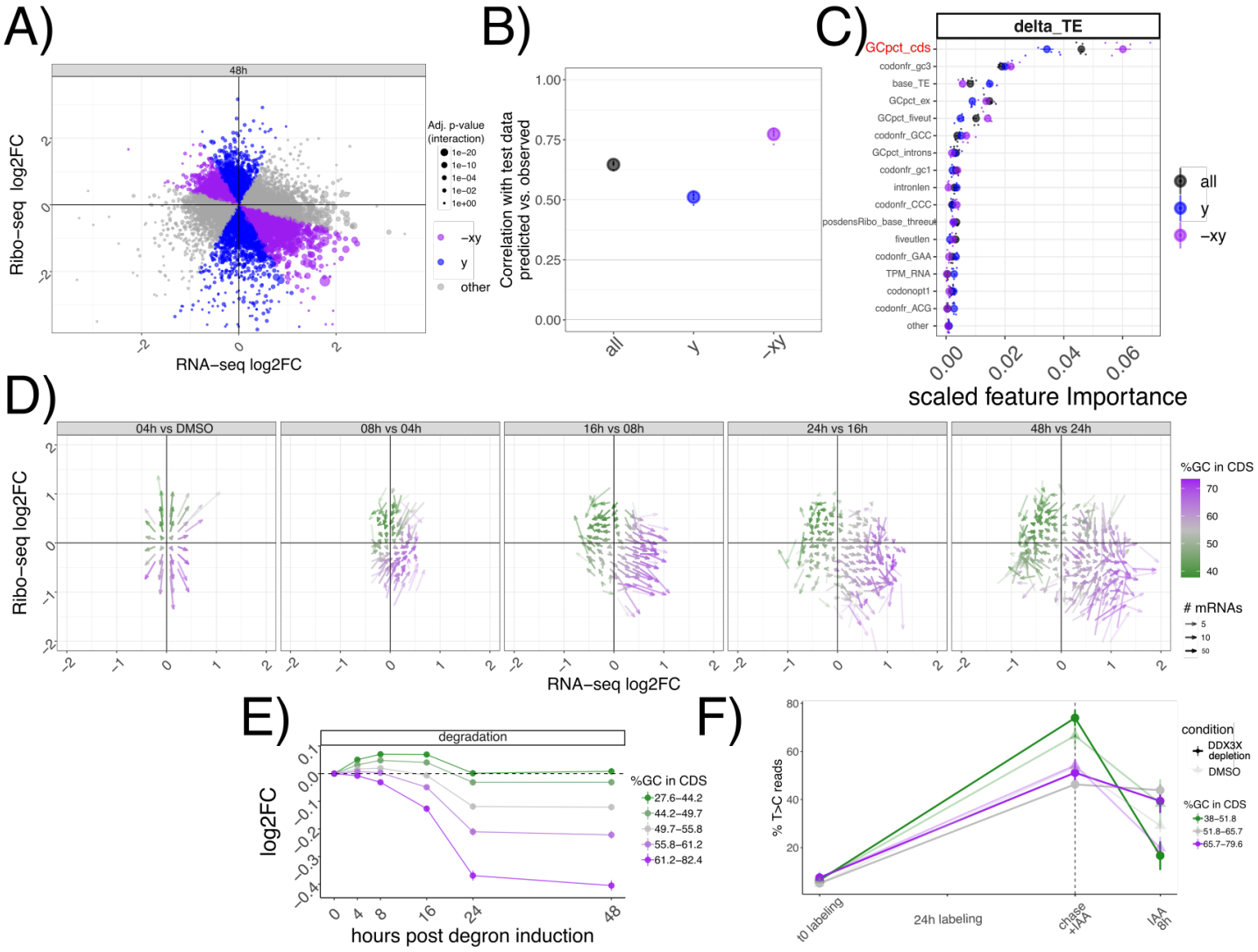

Figure 3: GC content in the coding sequence predicts regulation by DDX3X

A) Classification of different mRNAs according to their change in mRNA levels or translation. In B) performance on unseen test data of the random forest regression model for transcript classes as defined in A). C) Predictive power of different features, with their importance values (Methods) plotted on the x axis. D) Vector plot as in Figure 1D, highlighting GCcds values. Partition of inferred degradation rate (D) or SLAM-seq profiles (E) for mRNAs partitioned by GCcds values.

As shown in Figure 3B, the Random Forest model predicted TE changes with high precision, especially in cases where mRNA stability and translation were anti-correlated. In addition, this model calculated the predictive power of each input feature (Figure 3C, Methods), which highlighted GC content in the coding sequence (which we will refer to as GCcds) as the most important feature. Feature selection is a very important method to select predictive features, especially when facing high levels of multicollinearity (Supplementary Figure 3). To validate the results from the Random Forest regression, we used Lasso regression (Methods), another well-known method for feature selection. Results from the Lasso regression were very similar, and also identified GC content in the coding sequence as the most relevant feature in predicting TE changes (Supplementary Figure 4).

In the light of these results, we tested whether GCcds was associated with the DDX3X-dependent transcriptome dynamics reported above. As shown in Figure 3D, mRNAs partitioned on the Ribo-seq/RNA-seq coordinate system based on their GCcds value. Moreover, stability values from both INSPEcT and SLAM-seq partitioned according to GCcds values (Figure 3E-F). A similar, albeit weaker, separation was observed for predicted transcription and processing rates (Supplementary Figure 5).

By using multiple analytical approaches, we here show how GCcds, not just GC content in general, or in other sections of the transcriptome, is a predominant feature of stabilized, yet untranslated, mRNAs following DDX3X depletion.

GC content in the coding sequence is a ubiquitous signal in mRNA regulation.

Given the extensive connections between different aspects of mRNA regulation by thousands of regulators, we tested the breadth of the influence of features such as GCcds in other studies of RNA regulators. We re-analyzed >2000 RNA-seq samples (Methods) from the recent ENCODE RBPome[14] study encompassing >200 RBP knockdowns, and performed differential analysis followed by predictive modeling using the same methods and features as described in the previous section, this time aiming at predicting changes in mRNA levels (Figure 4A).

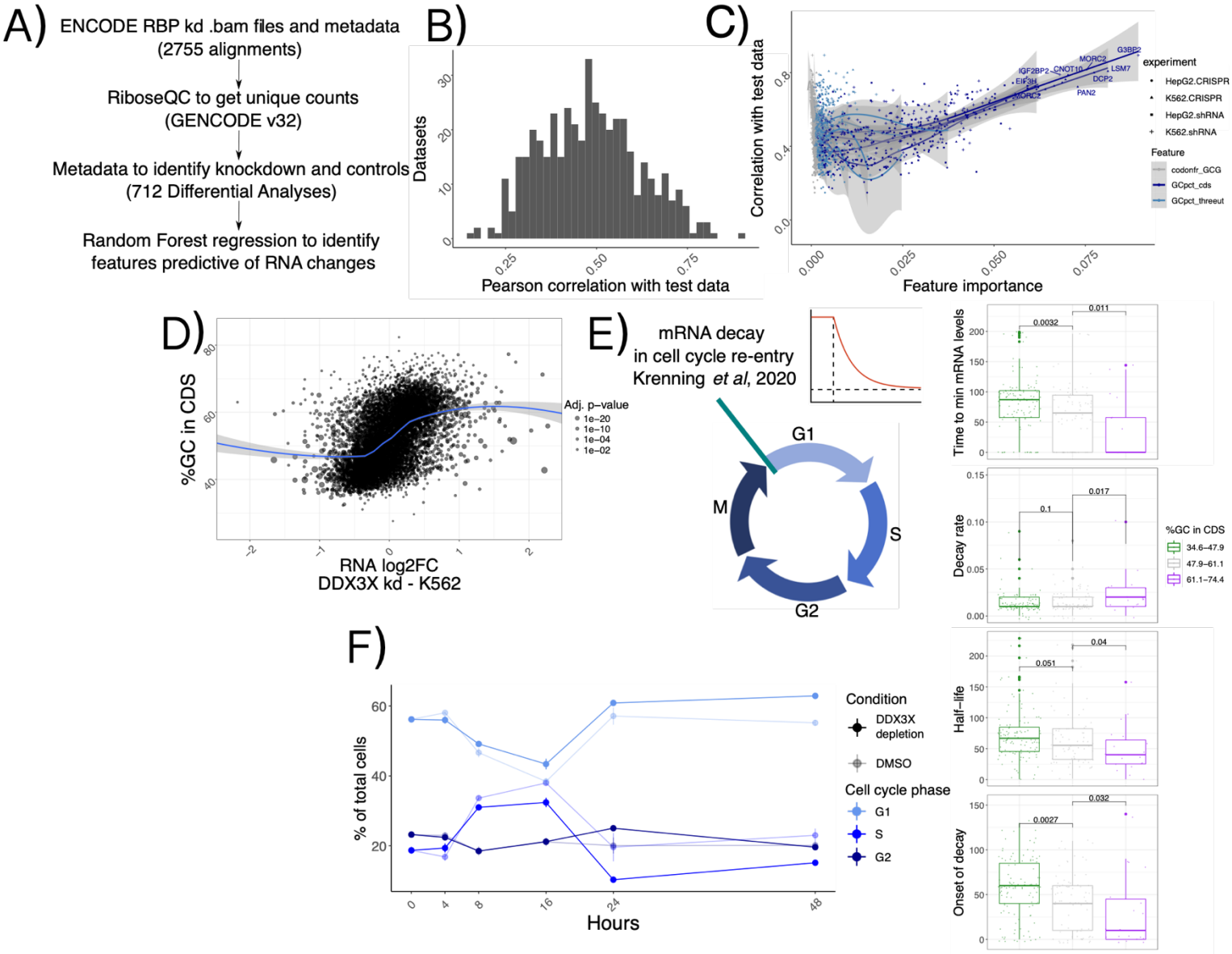

Figure 4: A ubiquitous feature in mRNA regulation

A) Schema describing the ENCODE analysis strategy. B) Histogram representing overall model performance across datasets. C) Model performance on the y axis vs. importance of different variables; note the linear relationship between GCcds importance and performance. D) mRNA level changes against GCcds values in a DDX3X knockdown experiment in the ENCODE dataset. E) Schematics of the cell cycle data used. Values for different kinetic parameters were partitioned according to GCcds values of their mRNAs and tested for significant differences. F) Percentage of cells in different stages of the cell cycle along the degron time course. Values for degron and control are shown with different transparency values.

We first grouped datasets according to knockdown efficiency, which varied according to knockdown method and cell line (Supplementary Figure 6, Methods). We selected the sample with the highest knockdown efficiency for each RBP and called feature importance using our analytical pipeline. Predictive power of our Random Forest regression strategy varied across different datasets (Figure 4B). Once again, the strongest predictor of mRNA changes was GCcds, whose predictive power dominated over other variables (Figure 4C). As expected, changes upon DDX3X knockdown in the ENCODE data also exhibited a clear dependency over GCcds (Figure 4D), albeit to a lower degree compared to our degron dataset, likely due to differences in DDX3X depletion strategies and, importantly, to our translation profiling

dataset, which allowed us to distinguish between specific classes (i.e. “TE_down”) of regulated mRNAs (Discussion).
Given the widespread relevance of GCcds as a predictor of post-transcriptionally regulated targets, we reasoned that a major cellular process might mediate the observed mRNA changes. We re-analyzed data from a recent study[21] focused on mRNA clearance during cell cycle re-entry, where the authors used a FUCCI cell system coupling RNA labeling, scRNA-seq and single-molecule imaging techniques to find extensive decay differences among different transcripts, potentially related to poly-A tail mediated regulation. Despite a lower throughput when compared to sequencing-based experiments, kinetic parameters estimated from their data showed significant differences when partitioned by GCcds values (Figure 4E). mRNAs rich in GCcds showed lower half-life values, and fast decay kinetics at cell cycle re-entry, with the opposite trend exhibited by mRNAs poor in GC content in their coding sequence. Motivated by this finding, we decided to investigate differences in cell cycle dynamics in our degron system, by using EdU incorporation followed by FACS analysis (Methods, Supplementary Figure 7). As shown in Figure 4F, DDX3X depletion resulted in cells staying more in G1 and less in S phase when compared to controls, throughout the time course.
By re-analysis of thousands of RNA-seq samples, these results show the prevalence of GCcds in post-transcriptional regulation and RBP functions, with a potential role for cell-cycle dependent mRNA dynamics in shaping such a regulatory phenomenon.

A shift in 5’-3’ RNA-coverage as a hallmark of mRNA stabilization.

In addition to gene-level aggregate measures of abundance, we investigated whether changes in decay could be identified by taking advantage of the high resolution of RNA-seq experiments across gene bodies, which has previously been employed to inform about mRNA decay[19]. We leveraged our time-resolved degron dataset to investigate changes in 5’-3’ coverage, a known hallmark of RNA degradation often employed to verify overall integrity of cellular mRNAs or to estimate transcript-level decay. We calculated 2 different metrics, using the strategy illustrated in Figure 5.

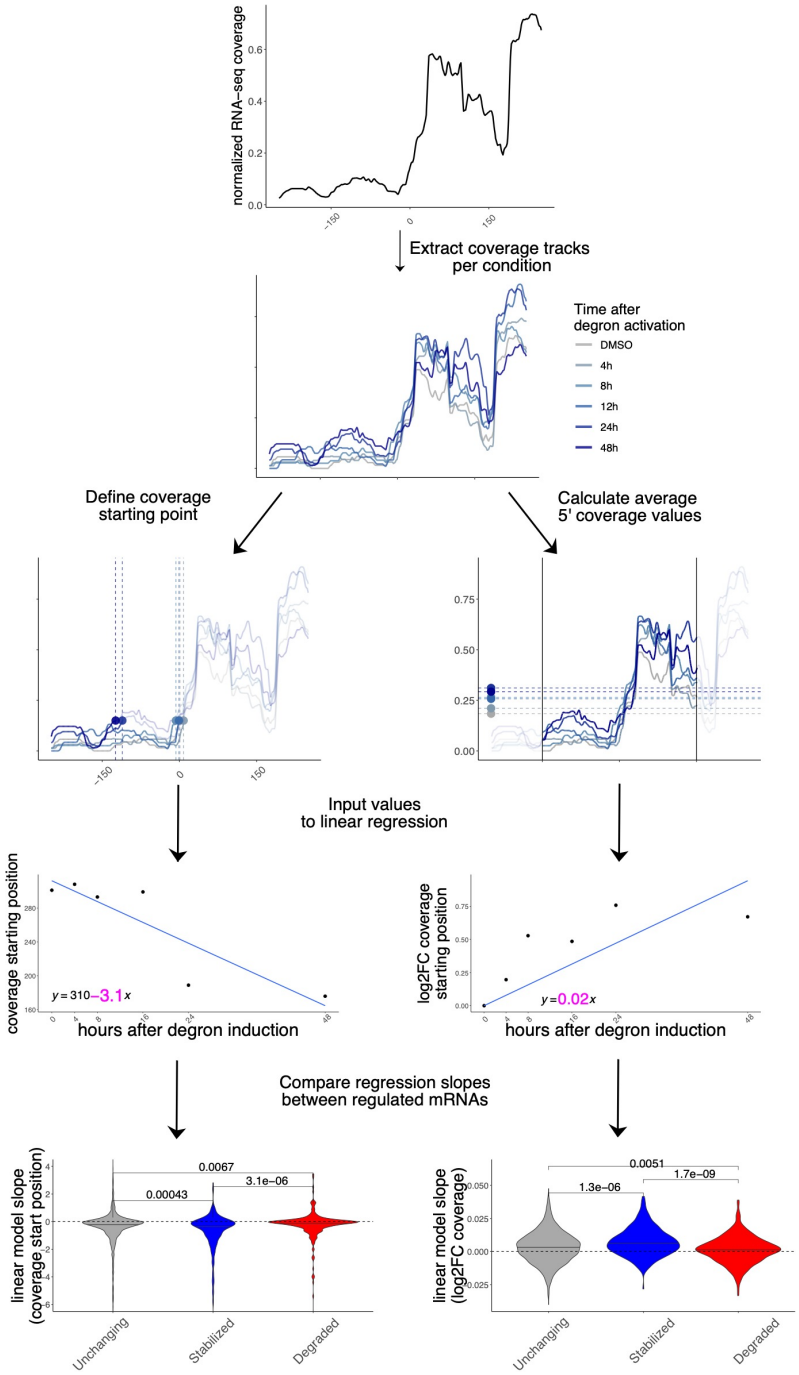

Figure 5: Coverage analysis of regulated mRNAs reveals changes in 5'-3' decay

Coverage analysis strategy in the degron dataset using a practical example (*CSRNP2* gene): coverage starting point is first identified using pooled data, then coverage tracks for each experiment are extracted. Coverage starting points and coverage values are calculated for each time point and input to a linear model. The beta coefficient (in pink) for each model is then extracted for each mRNA and values are compared across different classes (stabilized vs unchanging vs degraded). More details are available in the Methods section.

Initially, we pooled all samples to identify the major isoform for each gene (Methods), and the first position at 15% of the maximum coverage. We then calculated such position for each time point. Importantly, coverage values were normalized for each transcript, thus controlling for expression level changes. Also, we did not observe similar changes at the 3' end of transcripts (Supplementary Figure 8). We then used coverage starting points as input for linear regression. The regression coefficient was extracted and compared across the top 250 stabilized, degraded, and control mRNAs, alongside 1500 control transcripts. As shown in

Figure 5, coverage values on stabilized mRNAs started as an earlier position in the transcripts, indicating a lower 5'-3' decay along the DDX3X degron time course. The opposite trend was observed for degraded transcripts. Similarly, we calculated average coverage values in a window of 300nt around the coverage start and applied a similar strategy: 5' coverage values increased along the time course, confirming the accumulation of translationally suppressed mRNA species otherwise destined for degradation. Results were similar when using different cutoffs for the definition of coverage starting point (Supplementary Figure 9).

To test whether the suppression of 5'-3' decay of untranslated transcripts by DDX3X occurs *in vivo*, we re-analyzed recent RNA-seq/Ribo-seq dataset in a conditional *Ddx3x* (cKO) mouse model[13](Figure 6), where cell cycle and neurogenesis defects are evident when *Ddx3x* is depleted in neuronal progenitors. After applying our analytical pipeline, we observed that the

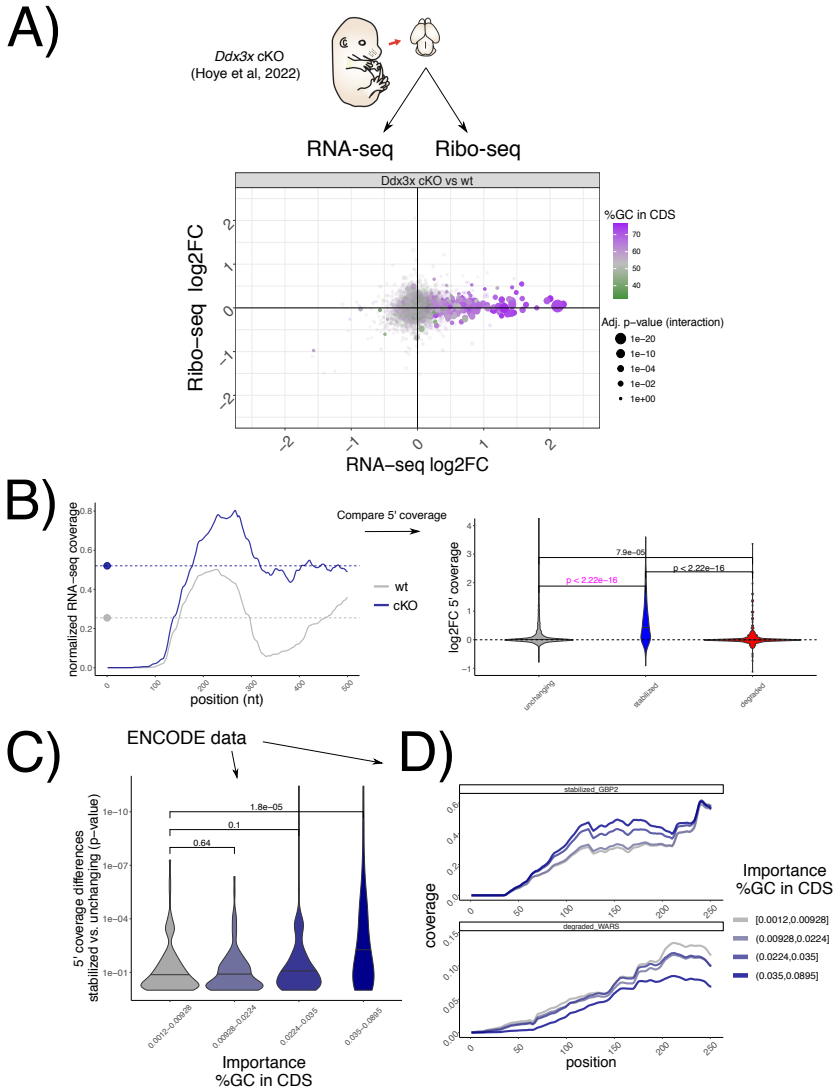

Figure 6. GCcds - mediated mRNA stabilization is detectable in vivo and across the ENCODE RBP database.

A) Changes in Ribo-seq and RNA-seq levels in a conditional *Ddx3x* mouse model, as in Figure 1A, showing GCcds values. B) Strategy for coverage analysis in the mouse *Ddx3x* cKO experiment, shown for the *Ctxn1* gene. Differences in coverage values are extracted and compared across regulated mRNAs. In C) same strategy as in Figure 5A applied to each differential analysis followed by RBP knockdown in the ENCODE dataset. Differences in coverage values between stabilized and unchanging mRNAs (shown by p-values, as calculated as in panel B), in pink color) are plotted against GCcds importance (x axis). D) Example mean coverage on 2 mRNAs (1 stabilized and 1 degraded), partitioning RBP knockdown datasets by their GCcds importance. An increase in coverage can be observed for the stabilized mRNA, while the opposite trend is visible for a degraded transcript.

accumulation of untranslated transcripts is even more evident in this *in vivo* model, as is its relationship with GCcds values (Figure 6A). Analogous to the strategy presented in Figure 5, 5' coverage values, as well as coverage starting points (Supplementary Figure 10), differed significantly between wild type and *Ddx3x* cKO animals (Figure 6B) in regulated transcripts, while no difference was detected at the 3'end (Supplementary Figure 11).

Leveraging again the power of hundreds of RNA-seq experiments, we examined 5' coverage profiles in the ENCODE dataset, partitioning experiments by their dependency on GCcds values. Differences between stabilized and control mRNAs are greater as the GCcds signature is more predominant (Figure 6C). Aggregating different experiments according to their GCcds dependency for example transcripts (Figure 6D) confirm this phenomenon, where both coverage starting position and coverage values changed across different datasets, indicative of mRNA decay regulation.
Taken together, we provide evidence for *in vivo* DDX3X-mediated stabilization of untranslated transcripts, its dependence on GCcds values, and a high-resolution RNA-seq coverage analysis strategy to investigate GCcds-related mRNA decay regulation, with support from hundreds of post-transcriptionally perturbed transcriptomes.

Discussion

The multifaceted role of DDX3X, described as involved in different molecular processes, often hinders the ability to understand its functions, especially considering the interconnected nature of molecular processes in the cell. Multiple mRNA features might underlie different modes of regulation, as we previously showed and experimentally validated 5'UTR dependencies underlying DDX3X translation regulation[7]. This outlines an unmet need for studies linking multiple aspects of the gene expression cascade.
In addition to profiling RNA levels and translation, we further dissected dynamics of cytoplasmic regulation by DDX3X, by employing a time course of efficient DDX3X depletion (Figure 1A). Akin to previous studies observing translation suppression preceding mRNA changes during miRNA-mediated regulation[10], we observed an accumulation of translationally suppressed RNAs. This highlights the importance to profile not only mRNA abundance but also translation levels, which, in absence of quantitative estimates of regulated protein levels, can greatly help researchers understanding the functions of many cryptic regulators often involved in multiple processes, like DDX3X and other RBPs[22].

By employing multiple techniques for feature selection, we identified a major feature underlying mRNA regulation by DDX3X, as well as by many other post-transcriptional regulators. An important area of investigation for the future is to employ more unbiased approaches, akin to recent Natural Language Processing-inspired methods in transcription regulation[23], in mRNA biology to accurately estimate the relevant features directly from data rather than specified by potentially biased approaches. In our hands, the relevance of GCcds is clearly picked up by both the Random Forest and the Lasso (Supplementary Figure 4). Importantly, we included similar features, such as overall GC content[24], in UTRs, introns etc., alongside codon frequencies[20] and previously estimated values of codon optimality.

Our study suggests that data-driven approaches to functional transcriptomics are highly needed, where data from multiple experiments are routinely re-analyzed to test hypotheses and provide new insights into the complex world of mRNA biology. However, while profiling translation allowed us to focus on specific mRNA classes and their features, no large-scale translation profiling study exists yet, with few, precious small atlases recently appearing in the literature[25]. The current ENCODE RBP series is of great value to many mRNA biology researchers worldwide and it has been an invaluable resource for many recent studies[26,27], yet an extension of these approaches which includes other aspects of post-transcriptional regulation, such as translation and stability, is in great need.

In the original ENCODE RBP study[14], gene expression estimates were GC-corrected for each sample, as GC content has been often reported as a confounder, especially when comparing across sequencing technologies and labs. Our degron time course analysis detected GC content specifically in coding sequence as a feature underlying regulation, and this region-specific effect is not consistent with a general confounding role for GCcds. Moreover, our analysis focused on differences upon a perturbation under a single sequencing platform and laboratory settings, which are likely to have similar GC-related confounders, should there be any. Important confirmation of the relevance of GCcds and its relationship to mRNA dynamics also came from: employing SLAM-seq to estimate differences in stability (Figure 2), re-analysis of *in vivo Ddx3x* cKO RNA-seq/Ribo-seq (Figure 6), re-analysis of hundreds of RBP perturbations in human cell lines (Figure 4), and by analyzing kinetics extracted by transcriptome dynamics in cell-cycle specific states (Figure 4).

Together with well-established differential analysis statistical methods, which allowed us to robustly identify different classes of regulated mRNAs, we exploited the high resolution offered by RNA-seq to analyze differences in 5'end coverage for thousands of individual transcripts (Figure 5), as an additional metric reflecting active regulation of mRNA decay mechanisms. We posit that popular analysis strategies for -omics techniques, despite their popularity over more than a decade, often obscures information with regards to mRNA processing and other molecular mechanisms, which can be uncovered by dedicated computational methods. Importantly, such dynamics are invisible (or, worse, can significantly distort quantification estimates) when performing gene-level analyses.

The mechanism, or mechanisms, by which GC content in coding regions shapes mRNA dynamics is still to be determined. We speculate that complex RNA structures in the coding sequence can form in the absence of active translation elongation, and such structure may mediate degradation, helped by RNP complexes in the cytoplasm. However, recent literature focused on the role of different codons in mediating such effect[18]. In our hands, codon-mediated effects seem to be negligible when considering the overall GCcds values, but more work needs to be done to identify cases where one or the other, or a mix of the two, can mediate mRNA decay on different transcripts. The involvement of mRNA dynamics during the cell cycle (Figure 4) suggests a model where, during cell cycle - dependent translation suppression, mRNAs are able to fold structures in the coding sequence promoting decay, and, when such processes are misregulated (e.g., by depleting multifunctional RNA helicases such as DDX3X), this process is less efficient. Further work needs to be done to refine the exact function, together with the subcellular localization, of regulated mRNAs. For instance, mRNA retention in the nucleus might be an additional underappreciated mode of gene expression

control[28], and is in line with our observation about the untranslated status of regulated transcripts. However, we identified GC content in the coding sequence as the hallmark feature for stabilized transcripts, a feature which is defined by translation in the cytoplasm. While RBP binding data remains an important starting point from which we can build testable hypothesis, simple binding-to-function paradigms might also create bias when trying to explain complex phenotypes arising from RBP misfunction. This study lays the ground for future work, to focus on identification of RNA secondary structures in different sections of mRNAs and their functions in DDX3X-mediated regulation.

Previous work implicated DDX3X in mediating cell cycle dynamics by a variety of mechanisms[1], including a direct regulation of cyclin E1 translation[29], which however was not among the most regulated mRNAs in our dataset (Supplementary Table 2). More work needs to be done to accurately quantify mRNA dynamics and RBP functions in the cell cycle, where translation regulation mechanisms[30,31] ensure controlled rates protein synthesis. The connection between cell cycle, sequence content and mRNA regulation is reinforced by the *in vivo* data, adding to the importance of studying post-transcriptional regulation along the neurogenesis axis[32,33], where the equilibrium between proliferation, apoptosis and differentiation[34] shapes the complexity of the developing brain.

Methods

Ribo-seq and RNA seq experimental protocol

HCT116 cells with inducible degradation of DDX3X (as previously described[16]), were plated in 15cm plates at 20% confluency (~3.5x10 6 cells/plate). 48 hours post plating, when the cells were at ~ 70% confluency, the media was changed and fresh media with 500 μM IAA (Indole-3-acetic acid, the most common naturally occurring Auxin hormone) (Research Products International, cat: I54000-5.0) or DMSO was added to cells. Cells were harvested at 0, 4, 8, 16, 24, and 48 hours post IAA addition. Cell number did not appreciably increase over the 48 hours of the experiment. To quantify DDX3X protein, we used an anti-DDX3X antibody described in previous work[7] normalized to an anti-GAPDH antibody (Rockland Immunochemicals, cat: 600-401-A33S).
Cells were treated with 100 μg/ml cycloheximide (CHX), washed with PBS containing 100 μg/ml CHX, and immediately spun down and flash frozen. Once all time-points were collected, the cells were thawed and lysed in ice-cold lysis buffer (20 mM TRIS-HCl pH 7.4, 150mM NaCl, 5 mM MgCl2, 1mM DTT, 100 μg/ml CHX, 1 % (v/v) Triton X-100, 25 U/ml TurboDNase (Ambion)). 240 μl lysate was treated with 6 μl RNase I (Ambion, 100 U/μl) for 45 minutes at RT with gentle agitation and further digestion halted by addition of SUPERase:In (Ambion). Illustra Microspin Columns S-400 HR (GE healthcare) were used to enrich for monosomes, and RNA was extracted from the flow-through using Direct-zol kit (Zymo Research). Gel slices of nucleic acids between 24-32 nts long were excised from a 15% urea-PAGE gel. Eluted RNA was treated with T4 PNK and preadenylated linker was ligated to the 3' end using T4 RNA Ligase 2 truncated KQ (NEB, M0373L).
Linker-ligated footprints were reverse transcribed using Superscript III (Invitrogen) and gel-purified RT products circularized using CircLigase II (Lucigen, CL4115K). rRNA depletion was performed using biotinylated oligos as described[35] and libraries constructed using a different reverse indexing primer for each sample.
For the RNA-seq, RNA was extracted from 25 μl intact lysate (non-digested) using the Direct-zol kit (Zymo Research) and stranded total RNA libraries were prepared using the TruSeq Stranded Total RNA Human/Mouse/Rat kit (Illumina), following manufacturer's instructions.
Libraries were quantified and checked for quality using a Qubit fluorimeter and Bioanalyzer (Agilent) and sequenced on a HiSeq 4000 sequencing system.

Slam-seq experimental protocol

SLAM-seq was performed at 60-70% confluency for DDX3X-mAID tagged HCT116. Media was changed and fresh media with 100μM 4-thiouridine (4sU) was added to cells and changed every 3 hours for 24 hours. 8 hours prior to collection, growth medium was aspirated and replaced. Uridine chase was performed where cells were washed twice with 1X PBS and incubated with media containing 10 mM uridine and DMSO or 100μM IAA for 0 or 8 hours to induce degradation of DDX3X. At respective time points, cells were harvested followed by total RNA extraction using TRIzol (Ambion) following the manufacturer's instructions (SLAMseq Kinetics Kit – Catabolic Kinetics Module, Lexogen). Total RNA was alkylated by iodoacetamide for 15 min and RNA was purified by ethanol precipitation. 200ng alkylated RNA were used as input for generating 3'-end mRNA sequencing libraries using a

commercially available kit (QuantSeq 3' mRNA-Seq Library Prep Kit FWD for Illumina, Lexogen).

Ribo-seq data analysis

Reads were stripped of their adapter, collapsed, and UMI sequences were removed. Clean reads were then mapped to rRNA, tRNA, snoRNA and miRNA sequences using bowtie2[36] using sequences retrieved from UCSC browser and aligning reads were discarded. Remaining reads were mapped to the genome and transcriptome using STAR[37] v2.7.9a supplied with the GENCODE v32 GTF file. STAR parameters were: *--outFilterMismatchNmax 3 --outFilterMultimapNmax 50 --chimScoreSeparation 10 --chimScoreMin 20 --chimSegmentMin 15 --outFilterIntronMotifs RemoveNoncanonicalUnannotated --alignSJoverhangMin 500 --outSAMmultNmax 1 --outMultimapperOrder Random.*

SLAM-seq data analysis

Reads were mapped to the genome and transcriptome using same RNA-seq parameters, except for *--outFilterMismatchNmax 10*. Reads containing T>C mutations were extracted from the BAM file using *GenomicAlignments* and *GenomicFiles* Bioconductor[38] packages.

RNA-seq data analysis

Reads were mapped to the genome and transcriptome using STAR with same Ribo-seq parameters. Synthesis, processing, and degradation rates were obtained using INSPEcT[17] v1.17, using default settings. Genes significantly changing in their dynamics at a p-value cutoff of .05 were used for subsequent analysis.

Differential analysis

Unique counts on different genomic regions were obtained using *RiboseQC*[39]. 5' end coverage values were inspected using Bioconductor[38] packages such as *GenomicFeatures*[40] and *rtracklayer*[41]. *DESeq2*[42] was used to obtain RNA-seq, Ribo-seq, and TE regulation, as described previously[7]: changes in translation efficiency were calculated using DESeq2 by using assay type (RNA-seq or Ribo-seq) as an additional covariate. Translationally regulated genes were defined using an FDR cutoff of 0.05 from a likelihood ratio test, using a reduced model without the assay type covariate, e.g. assuming no difference between RNA-seq and Ribo-seq counts. A similar strategy was used to define significant changes in DDX3X-mediated stability from SLAM-seq: count tables with T>C reads were built and analyzed using labeling (4sU/DMSO) and degron status (8h. vs DMSO) as the two variables of interest; regulation in stability was defined using a reduced model without the degron type covariate, e.g. assuming no difference between DMSO and degron activation.
Translationally regulated genes (as defined by Ribo-seq/RNA-seq) and stability regulated genes (as defined by SLAM-seq) were defined using a p-value cutoff of .05.
For Figure 1D and 3D, the coordinate system was divided into 70 bins on each axis. GCcds values (for Figure 3D), or Ribo-seq and RNA-seq fold changes between each time point and the previous one (for Figure 1D) were averaged across genes in the same bin. Only mRNAs

with significant changes in translation efficiency at 48h post degron induction were considered.

Random Forest and Lasso regression

The Random Forest regression was run using the *randomForest*[43] package with default parameters. Lasso regression was performed on scaled variables using the *glmnet*[44] package. While the entire feature table is available in Supplementary Table 2, a short description of the input features follows:
TPM values using RNA-seq (in log scale). Baseline TE levels, defined as ratio of Ribo to RNA reads. Baseline RNA mature levels, defined as length-normalized ratio of RNA-seq reads in introns versus exons. GC content, length (in log scale) and Ribo-seq/RNA-seq density in: 5' UTRs, a window of 25nt around start and stop codons, CDS regions, non-coding internal exons, introns, and 3' UTRs. Codon frequencies. Measures of gene-specific codon optimality, previously calculated from a recent study[18]. GC-content at first, second, or third codon position.

Feature importance (measured by mean decrease in accuracy for the random forest model and by the lasso coefficients) and correlation between predicted and measured test data were calculated on a 5-fold cross-validation scheme.

Analysis of cell cycle - dependent mRNA dynamics

Estimated mRNA decay kinetics at cell cycle re-entry were deposited as supplementary files of the original study[21]. Genes were partitioned cutting their GCcds values into 3 groups given the low number of quantified genes (total n=220).

Cell cycle staging

To measure DNA replication and cell cycle stage, EdU (5-ethynyl-2´-deoxyuridine) was added to cells at 10nM for 1.5 hrs before harvesting. 1 confluent well of a 6-well plate of HCT116 cells were harvested and processed as per manufacturer's instructions for the Click-iT™ Plus EdU Alexa Fluor™ 647 Flow Cytometry Assay Kit (Thermo Fisher cat: C10634). Per manufacturer's instructions, FxCycle Violet DNA content stain (Thermo Fisher cat: F10347) was added after the Click-iT reaction at 1:1,000 dilution before quantifying on a BD LSR Dual Fortessa flow cytometer. Alexa Fluor 647 was measured in the 670-30 Red C-A Channel and FxCycle Violet Stain was measured in the 450-50 Violet F-A Channel. Analysis was performed using FACS DIVA and FlowJo V10 (FlowJo, LLC) software.

5'end coverage analysis

Computation on single-nucleotide coverage values was performed using *rtracklayer*[41]. For each differential analysis, we extracted the most 250 stabilized and the most 250 degraded genes ranking P-values from RNA-seq differential analysis. 1500 control RNAs were randomly sampled from non-regulated genes, using p-values >.2 and TPM values > 3. Coverage values were 0-1 (min/max) normalized and the first position at value >.15 was identified as coverage

starting position. In addition, a general coverage starting point was selected by pooling all samples, and a window of 250nt around such position was used to calculate average coverage values around the coverage start. Log2 fold change with respect to the control condition were then calculated.
For degron data, starting position and log2fc coverage values were extracted and used as input for linear regression. For coverage values, intercept was omitted, as the first value was 0. Beta coefficients were then extracted and compared between stabilized, degraded, and control mRNAs.
For mouse Ddx3x cKO and ENCODE data, differences between starting position (knockdown vs wt) and log2FC (knockdown vs wt) in coverage values were used to compare stabilized, degraded and control mRNAs, bypassing the regression step (2 values were calculated, as only wt or knockdown conditions were present).

Data and code availability

Raw sequencing data for Ribo-seq, RNA-seq and SLAM-seq can be found under GEO accession GSE218433, with token "ujmtquoulnirpgx". Encode accession numbers can be found in Supplementary Table 3. *Ddx3x* knockout Ribo-seq and RNA-seq were analyzed from accession number GSE203078, processed data can be found in Supplementary Table 4. Code to reproduce all figures, together with processed data, can be found at https://github.com/calviellolab/DDX3X_GC_paper .

Author contributions

Z.J. performed SLAM-seq and provided support for analysis of stabilized mRNAs. A.X. performed cell cycle analysis along the degron time course. S.V. performed Ribo-seq and RNA-seq. M.H. and D.L.S. provided crucial input with the *Ddx3x* mouse model dataset. S.N.F. supervised the initial part of the project, secured funding, and provided significant input, together with all authors, in writing the manuscript. L.C. conceived the project, performed all computational analyses, prepared the figures, and wrote the manuscript.

Acknowledgements

L.C. wants to thank Piero Angela (1928-2022) for inspiring him, and countless other kids in Italy, to pursue a career in academic science, in constant awe of the beautiful mechanisms shaping the natural world. This work was supported by the National Institutes of Health DP2GM132932 and R01NS120667 (to S.N.F.). S.N.F. is a Pew Scholar in the Biomedical Sciences, supported by The Pew Charitable Trusts. ZJ was supported by the UCSF Moritz-Heyman Discovery Fellow Program. AX was supported by NIH F30 Ruth L Kirschstein National Research Service Award HD110250. ZJ and AX were supported by the UCSF Medical Scientist Training Program (T32GM007618). Flow cytometry data was generated at the UCSF Parnassus Flow CoLab (RRID:SCR_018206) and sequencing was performed at the UCSF CAT, supported by UCSF PBBR, RRP IMIA, and NIH 1S10OD028511-01 grants.

Supplementary information

Supplementary Figures are provided separately.

Supplementary Table 1. Read mapping statistics for the Ribo-seq RNA-seq DDX3X time course dataset.

Supplementary Table 2. Input to the Random Forest model for the DD3X3 time course dataset.

Supplementary Table 3. Accession codes for the analyzed ENCODE datasets, with information for each differential analysis. Multiple accession can be technical replicate of a biological replicate.

Supplementary Table 4. Input to the Random Forest model for the cKO *Ddx3x* mouse dataset.

References

1. Shoemaker, C. J. & Green, R. Translation drives mRNA quality control. *Nat Struct Mol Biol* **19**, 594–601 (2012).
2. Ingolia, N. T., Brar, G. a, Rouskin, S., McGeachy, A. M. & Weissman, J. S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc* **7**, 1534–1550 (2012).
3. Herzog, V. A. *et al.* Thiol-linked alkylation of RNA to assess expression dynamics. *Nature Methods 2017 14:12* **14**, 1198–1204 (2017).
4. Hafner, M. *et al.* CLIP and complementary methods. *Nature Reviews Methods Primers 2021 1:1* **1**, 1–23 (2021).
5. Sharma, D. & Jankowsky, E. The Ded1/DDX3 subfamily of DEAD-box RNA helicases. *Critical Reviews in Biochemistry and Molecular Biology* Preprint at https://doi.org/10.3109/10409238.2014.931339 (2014).
6. Oh, S. *et al.* Medulloblastoma-associated DDX3 variant selectively alters the translational response to stress. *Oncotarget* (2016) doi:10.18632/oncotarget.8612.
7. Calviello, L. *et al.* DDX3 depletion represses translation of mRNAs with complex 5' UTRs. *Nucleic Acids Res* **49**, 5336–5350 (2021).
8. Chang, Y.-F. F., Imam, J. S. & Wilkinson, M. F. The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem* **76**, 51–74 (2007).
9. D'Orazio, K. N. *et al.* The endonuclease Cue2 cleaves mRNAs at stalled ribosomes during No Go Decay. *Elife* **8**, (2019).
10. Bazzini, A. A., Lee, M. T. & Giraldez, A. J. Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. *Science* **336**, 233–7 (2012).

11. Lennox, A. L. *et al.* Pathogenic DDX3X Mutations Impair RNA Metabolism and Neurogenesis during Fetal Cortical Development. *Neuron* **106**, 404-420.e8 (2020).
12. Ingolia, N. T. Ribosome Footprint Profiling of Translation throughout the Genome. *Cell* **165**, 22–33 (2016).
13. Hoye, M. L. *et al.* Aberrant cortical development is driven by impaired cell cycle and translational control in a DDX3X syndrome model. *Elife* **11**, (2022).
14. Van Nostrand, E. L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature 2020 583:7818* **583**, 711–719 (2020).
15. Hon, C. C. & Carninci, P. Expanded ENCODE delivers invaluable genomic encyclopedia. *Nature 2021 583:7818* **583**, 685–686 (2020).
16. Venkataramanan, S., Gadek, M., Calviello, L., Wilkins, K. & Floor, S. N. DDX3X and DDX3Y are redundant in protein synthesis. *RNA* **27**, rna.078926.121 (2021).
17. De Pretis, S. *et al.* INSPEcT: a computational tool to infer mRNA synthesis, processing and degradation dynamics from RNA- and 4sU-seq time course experiments. *Bioinformatics* **31**, 2829–2835 (2015).
18. Medina-Muñoz, S. G. *et al.* Crosstalk between codon optimality and cis-regulatory elements dictates mRNA stability. *Genome Biol* **22**, 1–23 (2021).
19. Courel, M. *et al.* Gc content shapes mRNA storage and decay in human cells. *Elife* **8**, (2019).
20. Hia, F. *et al.* Codon bias confers stability to human mRNAs. *EMBO Rep* **20**, e48220 (2019).
21. Krenning, L., Sonneveld, S. & Tanenbaum, M. Time-resolved single-cell sequencing identifies multiple waves of mRNA decay during the mitosis-to-G1 phase transition. *Elife* **11**, (2022).
22. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat Rev Genet* **15**, 829–845 (2014).
23. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods 2021 18:10* **18**, 1196–1203 (2021).
24. Thomas, A. *et al.* RBM33 directs the nuclear export of transcripts containing GC-rich elements. *Genes Dev* **36**, 550–565 (2022).
25. Chothani, S. P. *et al.* A high-resolution map of human RNA translation. *Mol Cell* **82**, 2885-2899.e8 (2022).
26. Zhao, W. *et al.* POSTAR3: an updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Res* **50**, D287–D294 (2022).
27. Van Nostrand, E. L. *et al.* Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. *Genome Biol* **21**, 1–26 (2020).
28. Bahar Halpern, K. *et al.* Nuclear Retention of mRNA in Mammalian Tissues. *Cell Rep* **13**, 2653–2662 (2015).
29. Lai, M.-C., Chang, W.-C., Shieh, S.-Y. & Tarn, W.-Y. DDX3 Regulates Cell Growth through Translational Control of Cyclin E1. *Mol Cell Biol* **30**, 5444–5453 (2010).
30. Clemm von Hohenberg, K. *et al.* Cyclin B/CDK1 and Cyclin A/CDK2 phosphorylate DENR to promote mitotic protein translation and faithful cell division. *Nature Communications 2022 13:1* **13**, 1–14 (2022).
31. Tanenbaum, M. E., Stern-Ginossar, N., Weissman, J. S. & Vale, R. D. Regulation of mRNA translation during mitosis. *Elife* **4**, (2015).

32. Hoye, M. L. & Silver, D. L. Decoding mixed messages in the developing cortex: translational regulation of neural progenitor fate. *Curr Opin Neurobiol* **66**, 93–102 (2021).
33. Harnett, D. *et al.* A critical period of translational control during brain development at codon resolution. *Nature Structural & Molecular Biology 2022 29:12* **29**, 1277–1290 (2022).
34. Pilaz, L. J. *et al.* Prolonged Mitosis of Neural Progenitors Alters Cell Fate in the Developing Brain. *Neuron* **89**, 83–99 (2016).
35. Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M. & Weissman, J. S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc* **7**, 1534–50 (2012).
36. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* (2012) doi:10.1038/nmeth.1923.
37. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
38. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* **12**, 115–121 (2015).
39. Calviello, L., Sydow, D., Harnett, D. & Ohler, U. Ribo-seQC: comprehensive analysis of cytoplasmic and organellar ribosome profiling data. doi:10.1101/601468.
40. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLoS Comput Biol* **9**, e1003118 (2013).
41. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).
42. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
43. Wiener, A. L. and M. Classification and Regression by randomForest. *R News* **2**, 18–22 (2002).
44. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **33**, 1 (2010).