Title

A Ubiquitous GC Content Signature Underlies Multimodal mRNA Regulation by Ddx3x

Author

Jowhar, Ziad Mohamoud

Publication Date

2024

A Ubiquitous GC Content Signature Underlies Multimodal mRNA Regulation by Ddx3x

by
Ziad Mohamoud Jowhar

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biomedical Sciences

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
William Weiss
1EF6E77D125647B...

William Weiss
Chair

DocuSigned by:
LUKE GILBERT
DocuSigned by:
Stephen Floor
3709C3C6AF93461...

LUKE GILBERT

Stephen Floor

Committee Members

Dedications and Acknowledgements

There is an innumerable amount of people who helped me along this long journey that I want to acknowledge, and I want to give sincere thanks to them all. First, I would like to thank Stephen Floor, my thesis advisor. Stephen provided me with tremendous support, encouragement, and mentorship, both scientifically and personally. I am grateful for everyone in the Floor Lab, both past and present. The Floor Lab members created a welcoming space for creativity, hard work, support, and laughter that enhanced our science. I want to shine a light on my graduate student cohort in the lab, Albert Xu, José Liboy Lugo, and Jesslyn Park for always listening to me vent, providing encouragement and advice on experiments, and enjoying my music during long hours in tissue culture. I want to give particular thanks to Albert for sharing his knowledge and being an amazing collaborator and an even better friend. I can confidently say I would not be writing this today without you. Also, thank you to Kevin Wilkins and our junior graduate students Margaret Gadek and Abby Hein for helping foster a supportive environment as well as sharing and providing HEK and HCT116 cells when I was coming back from whatever country I decided to go to. I want to thank Yizuh Lin for answering or providing me with resources after all my endless computational questions. You have taught me so much and have made me a better scientist.

I want to thank my thesis committee, William Weiss and Luke Gilbert, for their support, mentorship, and guidance. They have challenged me and provided me with advice and mentorship that helped me to flourish in my graduate school training. I want to thank my past mentors – Tom Misteli, Anita Corbett, Paul Doetsch, Armin Raznahan, Thomas Reid, and Munira Basrai for not only introducing me to science, but also for training me to be critical and

analytical, providing me with multiple outstanding professional opportunities, and for their continued advocacy. I also want to particularly thank Luke Gilbert and Becky Fu for their mentorship and guidance six years after my lab rotation. Additionally, I want to thank D'anne Duncan, Denise Davis, S. Andrew Josephson, and Talmadge King for their support, mentorship, and continued check-ins, professionally and personally. Last, I want to thank Dennis Chan for providing me with tea that sustained me throughout my training.

Finally, I want to thank my community. To my parents, thank you for all the prayers and motivation to give you an answer to one of your many questions, including when I am graduating because I am still working on the marriage question. To my sister, I love you, a true sensitive soul that taught me to feel, listen, and know you are a central driver for me to continue to be successful. I also want to thank my grandmother, aunts, uncles, and cousins in Atlanta, Oakland, and outside America, particularly Sara Ahmed and Sabrina Hussien, for their prayers, words of advice and encouragement, and at times, free concert tickets. To my friends turned family, Mekdes Tsega, Helina Iyob-Tessema, Yohanna Hailegabriel, and Elijah Nwefo, there are not enough words to describe how much you all mean to me. You all have helped me through my lowest of lows, helped me get into medical school, and have seen me grow into the man I am today, and I am eternally grateful for your love and friendship. To my med school family, Joseph Kidane, India Perez-Urbano, Ivie Tokunboh, Naya Okeke, and Prisca Diala, thank you all for the fun, laughs, and encouragement. You have each been a rock for me in San Francisco. I have learned so much from each of you, and I am a better friend, doctor, and scientist for it. I also want to thank Chris Anigwe, Xavier Cortez, and Joseph Kidane for providing me with a couch to sleep on, a car to use, and friends to joke with during critical experiments that resulted in my

publication. I would also like to thank my brother and nakama, Abdullah Mussa, for being a true friend, a man I look up to, who is always interested in a new adventure, who always has an incredible medical question, and for always pushing me past my limits and challenging me to be better. Thank you, Simisola Alalade, for being a moral guidepost, an outstanding discerner of taste, and a reminder of my blessings. For those left unnamed, please forgive me, and know I appreciate your continued support, friendship, and love. Please charge it to my head, not my heart. I am very lucky and blessed to have so many loving and supportive people around me.

Contributions

This dissertation includes contributions from Albert Xu, Srivats Venkataramanan, Francesco Dossena, Mariah L. Hoye, Debra L. Silver, Stephen N. Floor, Lorenzo Calviello.

Epigraph

“Not everything that is faced can be changed, but nothing can be changed until it is faced”.

— James Baldwin, "As Much Truth As One Can Bear". The New York Times Book Review, January 14, 1962.

A Ubiquitous GC Content Signature Underlies Multimodal mRNA Regulation by Ddx3x

Ziad Mohamoud Jowhar

ABSTRACT

DDX3X encodes for a DEAD-box RNA helicase that promotes the translation of mRNAs that contain long, highly structured 5′ UTRs. DDX3X interacts with translation initiation factors and can remodel mRNA structures or RNA protein complexes in an ATP-dependent manner. Our prior work found that DDX3X depletion leads to changes in translation and RNA levels while missense variants selectively impact translation, implicating DDX3X in regulating mRNA stability. Despite being an essential gene and implicated in several human cancers and developmental diseases, including intellectual disability and autism spectrum disorder, the mechanisms of DDX3X in post-transcriptional gene regulation and mRNA metabolism remain poorly understood.

To address this gap in knowledge in DDX3X function in post-transcriptional gene regulation and mRNA metabolism, this work studied the mechanism of how DDX3X regulates specific mRNA transcripts and confers stability or loss of stability and how this process is altered with loss of DDX3X. DDX3X's function in translation is well characterized, is known to be involved in nuclear export of specific mRNA, and our preliminary data demonstrates that DDX3X specific transcripts have altered stability. However, what features and characteristics of these DDX3X-dependent transcripts are key for DDX3X driven stability remain unknown. This study aimed to define what features of mRNA transcripts are key for how DDX3X regulates RNA metabolism both at the level of translation and transcript stability. This knowledge is critical for advancing our knowledge in post-transcriptional gene regulation and mRNA metabolism.

Table of Contents

List of Figures

List of Tables

INTRODUCTION

This work investigated the mechanism linking *DDX3X* to mRNA stability. *DDX3X* is an X-linked and essential gene, which encodes an RNA-binding protein of the DEAD-box helicase family. While broadly implicated in mRNA metabolism, DDX3 is best characterized as a translational regulator. DDX3 is involved in the translation of mRNAs, modulates cell growth, and affects neuronal development yet we lack a mechanistic understanding of *DDX3X*'s role in mRNA metabolism. Mutations in DDX3X occur in several human diseases including WNT- and SHH-subgroups of medulloblastoma and developmental disorders. Disease-associated mutations typically occur within its two RNA helicase domains and these DDX3 mutations accelerate tumorigenesis *in vivo*. Ultimately, these gaps in knowledge limit the understanding of how DDX3X mutations cause DDX3X Syndrome, medulloblastoma, and a range of other cancers, which are a barrier to therapeutic intervention. My thesis project used unbiased sequencing approaches to understand the role of DDX3X in translation and mRNA stability.

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.

BACKGROUND

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.[12] 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."[13,14]

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.[15] 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.[16] 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.[17]

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[18], where we found near-complete rescue of gene expression changes by DDX3X expression, thus being able to use this tool to monitor DDX3X-dependent changes. 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 2**. 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.

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.[19] INSPEcT uses the proportion of intronic versus exonic reads to identify nascent vs. mature transcripts, and uses a system of ordinary differential equations (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. Briefly, cells were incubated with 4sU to comprehensively label transcribed RNAs, and their abundance was followed after 8h of DDX3X degron activation, using DMSO as control. 4sU treatment induces T>C conversions in the sequenced cDNA molecules, which can be used to monitor mRNA stability changes after a uridine chase, as shown in **Figure 2B**. As expected, we observed a drastic drop in T>C harboring reads after the chase, which reflects mRNA decay rates (**Supplementary Figure 2**). As shown in **Figure 2B**, after a labeling time of 24 hours, the percentage of reads harboring T>C mutations was different

for the regulated categories (**Methods**) after only 8 hours of degron induction, confirming the stabilization of translationally suppressed mRNAs upon DDX3X depletion. While the modest depth and resolution 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. In addition, we profiled RNA abundance via qPCR combining our DDX3X degron system with ActD treatment, to measure RNA stability changes. We selected few target genes: JUND was identified in our data as a stabilized RNA, while EIF2A was identified to be degraded. RACK1, LGALS1, and PFN1 were used as controls to normalize with via RT-PCR with Taq-man probes. JUND RNA was stabilized after 24 hours with knock down of DDX3 and Actinomycin D (ACTd) treatment (**Supplementary Figure 3A**); EIF2A RNA was more degraded after 24 hours with knock down of DDX3 and ACTD (**Supplementary Figure 3B**). These results show an overall good agreement between the qPCR and the sequencing-based assays, despite the difficulty arising from choosing control genes and the modest fold changes observed in the sequencing data.

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 3**). 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.[20] 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.[21,22] 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**). Interestingly, the categories differed in their DDX3X binding pattern (**Supplementary Figure 4**): re-analysis of our previously published PAR-CLIP data showed how stabilized targes (x,-xy groups) have a lower T>C conversion signal in their 5'UTRs, and a higher signal in CDS peaks, with the opposite being true for true translation targets (y group). This analysis suggests that stabilized mRNAs might be regulated differently than "canonical" translationally suppressed targets.

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 (-xy group). 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 5**). 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 similar, and

also identified GC content in the coding sequence as the most relevant feature in predicting TE changes (**Supplementary Figure 6**). GC content in the CDS remained the top predictor when using additional features, such as GC content in different sections of the CDS, or amino acid frequencies (**Supplementary Figure 7**).

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 8**).

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[16] 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**).

We first grouped datasets according to knockdown efficiency, which varied according to knockdown method and cell line (**Supplementary Figure 9**, **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**, **Supplementary Figure 10**). 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[23] focused on mRNA clearance during cell cycle re-entry, where the authors used a FUCCI (fluorescent, ubiquitination-based cell-cycle indicators) 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 (exemplified in the decay curve in **Figure 4E**) 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 5-ethynyl-2'-deoxyuridine (EdU) incorporation followed by FACS analysis (**Methods**, **Supplementary Figure 11**). As shown in **Figure 4F** and **Supplementary Figure 12**, 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.[21] 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**.

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 13**). 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, with moderate albeit significant differences between categories, 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 14**).

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[15] (**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 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 15**), 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 16**).

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, supporting the different analyses reported in this study (**Figure 7**) 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.[24] Despite relatively fast kinetics of DDX3X degradation from our degron system, more work needs to be performed to pinpoint exactly what changes occur right after DDX3X depletion, and to more precisely quantify the lag between translation suppression and mRNA stabilization.

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[25], 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[26], in UTRs, introns etc., alongside codon frequencies[22] 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.[27] 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[28,29], 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[16], 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. Given the presence of GC-related biases in sequencing-based assays, we think that great caution must be taken when observing expression changes driven by GC content features, especially when interpreted as direct effects from single molecular factors. Our degron time course analysis, despite containing dozens of features

pertaining to GC content measures, 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**), qPCR validations (**Supplementary Figure 3**), 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.[20] 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. The extent to which cell cycle changes might depend on direct DDX3X binding and regulation remains to be elucidated. 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[30], 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. Moreover, we observed how binding patterns might different between different regulated classes (**Supplementary Figure 4**). In our previous study we investigated the changes in translation and RNA abundance using a DDX3X

helicase mutant; one of the observations we made pertained to the lack of RNA changes in our data, suggesting a potential function for the helicase activity in orchestrating such changes.

Previous work implicated DDX3X in mediating cell cycle dynamics by a variety of mechanisms[31], including a direct regulation of cyclin E1 translation[32], which however was not among the most regulated mRNAs in our dataset (**Supplementary Table 3**). More work needs to be done to accurately quantify mRNA dynamics and RBP functions in the cell cycle, where translation regulation mechanisms[33,34] 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[35,36], where the equilibrium between proliferation, apoptosis and differentiation[37] shapes the complexity of the developing brain.

MATERIALS AND METHODS

Ribo-seq and RNA seq experimental protocol

HCT116 cells with inducible degradation of DDX3X (as previously described[38]), were plated in 15cm plates at 20% confluency (~3.5×10 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[2] 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[39] using sequences retrieved from UCSC browser and aligning reads were discarded. Remaining reads were mapped to the genome and transcriptome using STAR[40] 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[41] 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[19] 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.*[42] 5' end coverage values were inspected using Bioconductor[41] packages such as *GenomicFeatures*[43]

and *rtracklayer.*[44] *DESeq2*[45] 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 **Figures 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*[46] package with default parameters. Lasso regression was performed on scaled variables using the *glmnet*[47] package.

While the entire feature table is available in **Supplementary Table 3**, 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.[20] 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.[23] 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.*[44] 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 >0.2 and TPM values > 3. Coverage values were 0–1 (min/max) normalized and the first position at value >0.15 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).

TaqMan RT-PCR

DDX3X-mAID tagged HCT116 cells were plated in six-well plates at 30–40% confluency. 24 h post plating 500 μM IAA or DMSO was added to cells with or without 200 nM Actinomycin D (ActD). Total RNA was extracted from cells at 60–70% confluency using Direct-zol kit (Zymo Research) at 0 and 24 h post-ActD and IAA or DMSO treatment. TaqMan probes for JUND, EIF2A, RACK1, LGALS1, and PFN1 were pre-designed and purchased from ThermoFisher Scientific. Probes for degraded (EIF2A) or stabilized mRNAs (JUND) were conjugated with FAM dye while control mRNAs RACK1, LGALS1, and PFN1 were conjugated with VIC dye. For the TaqMan real-time quantitative PCR amplification reactions, we employed an Applied Biosystems QuantStudio 6 Real-Time PCR System instrument. Real-time PCR was conducted using TaqMan Fast Virus 1-Step Master Mix from Applied Biosystems in 384-well plates, following the manufacturer's protocol. Each well contained probes targeting mRNAs subject to degradation (EIF2A) or stabilization (JUND) along with controls (RACK1, LGALS1, or PFN1). All reactions were conducted in triplicate. Thermal cycling conditions adhered to the manufacturer's recommended standard protocol. The quantification of the target input amount was determined using the cycle threshold (CT) value, which corresponds to the point at which the PCR amplification plot crosses the threshold. Expression of degraded and stabilized mRNAs were normalized to each control.

Datasets analyzed

ENCODE accession numbers can be found in **Supplementary Table 4**. *Ddx3x* knockout Ribo-seq and RNA-seq were analyzed from accession number GSE203078.

Data and code availability

The datasets and computer code produced in this study are available in the following databases: Ribo-seq, RNA-seq and SLAM-seq data: Gene Expression Omnibus with accession GSE218433. Code to reproduce all figures and tables, together with processed data, is freely accessible on Github: https://github.com/calviellolab/DDX3X_GC_paper.

FIGURES

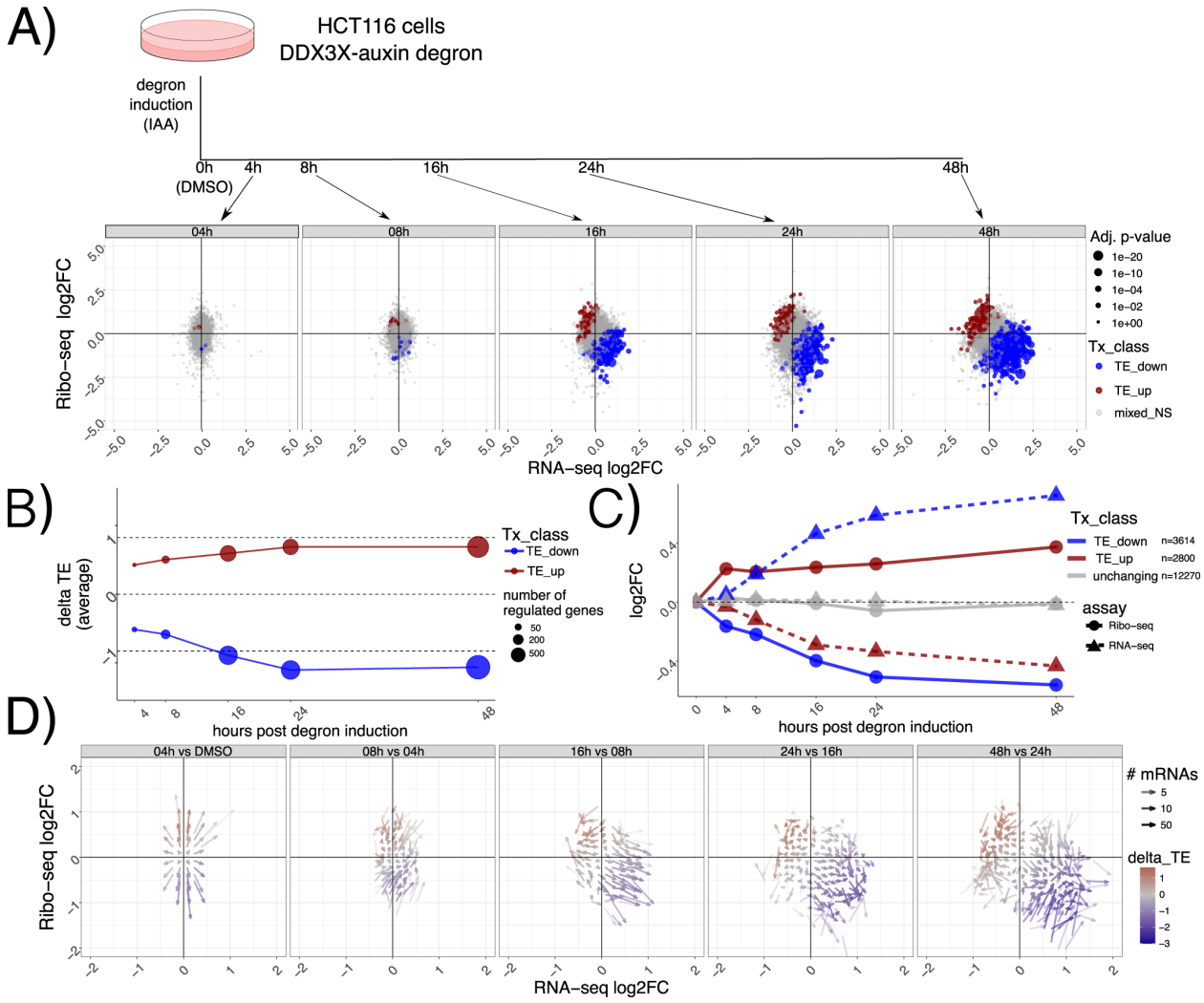


Figure 1: Dynamics of mRNA regulation by DDX3X.

(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 from a differential translation efficiency test using DESeq2 (Methods). NS: not significant.

(B) Average delta TE values (differences in 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 48 h 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).

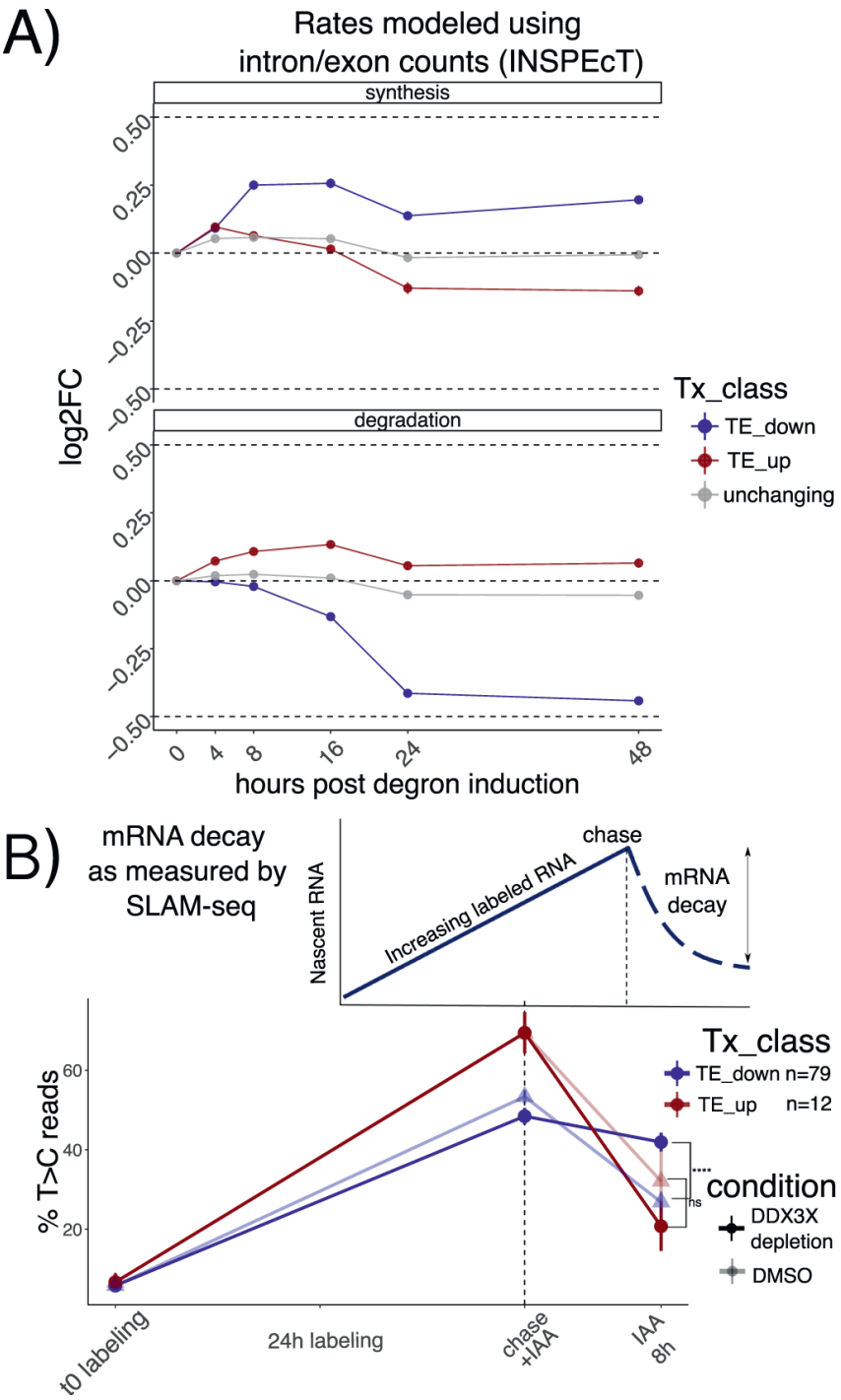


Figure 2: Stabilization of untranslated mRNAs.

(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. n of mRNAs: TE_down = 1809, TE_up = 1401; unchanging = 6136. Error bars represent the standard error of the mean.

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

(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. Significance values from a one-sided Wilcoxon test, using mRNAs significantly changing in both translation (from RNA-seq and Ribo-seq) and stability (from SLAM-seq, Methods), showing the following symbols: “ns” = p > 0.1; “.” = p <= 0.1; “*“ = p <= 0.05; “**“ = p <= 0.01; “***“ = p <= 0.001; “****“ = p <= 0.0001. Error bars represent the standard error of the mean.

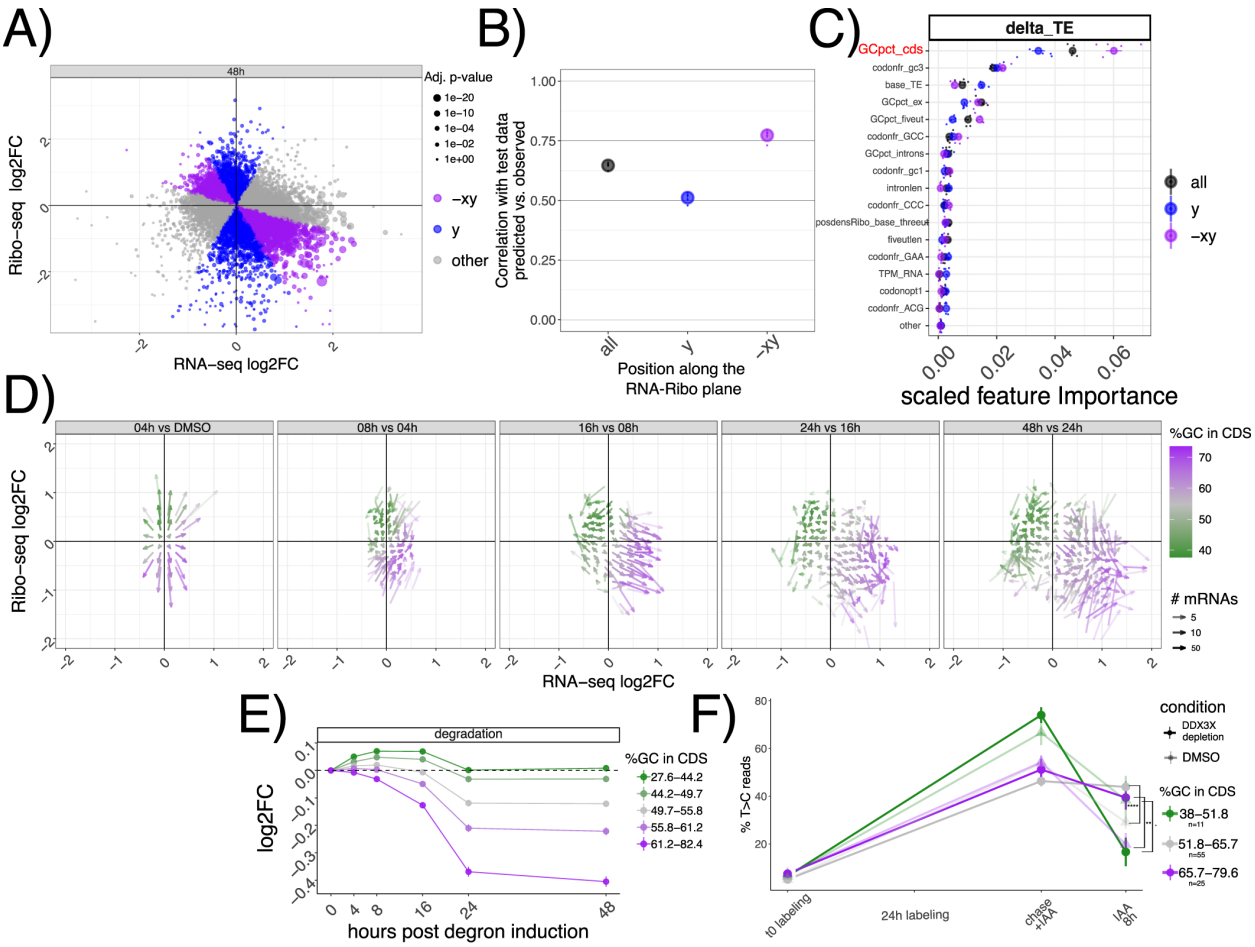


Figure 3: GC content in the coding sequence predicts regulation by DDX3X.

(A) Classification of different mRNAs according to their change in mRNA levels or translation. The size of the dots represents statistical significance from a differential translation efficiency test, as in Fig. 1A.

(B) Model performance (correlation between predicted vs. real values) on unseen test data of the random forest regression model for transcript classes as defined in (A). Performance shown for each round of cross-validation, n = 5; error bars represent the standard error of the mean.

(C) Predictive power of most informative features, with their importance values (Methods) plotted on the x axis. Features pertaining to GC content in different section of transcripts (GCpct*), baseline translation levels (base_TE), codon frequencies (codonfr*), positional read density (posdens*), and length features (intronlen) are displayed. Performance shown for each round of cross-validation, n = 5. Error bars represent the standard error of the mean.

(D) Vector plot as in Fig. 1D, highlighting GCcds values.

(E) Inferred degradation rate for mRNAs partitioned by GCcds values. Number of mRNAs for panel (E): “27.6-44.2” = 1861; “44.2-49.7” = 1892; “49.7-55.8” = 1854; “55.8-61.2” = 1847; “61.2-82.4” = 1892. Error bars represent the standard error of the mean.

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

(F) SLAM-seq profiles for mRNAs partitioned by GCcds values. Significance values in (F) from a one-sided Wilcoxon test, using the same mRNAs from Fig. 2B, showing the following symbols: “ns” = p > 0.1; “.” = p <= 0.1; “*“ = p <= 0.05; “**“ = p <= 0.01; “***“ = p <= 0.001; “****“ = p <= 0.0001. Error bars represent the standard error of the mean.

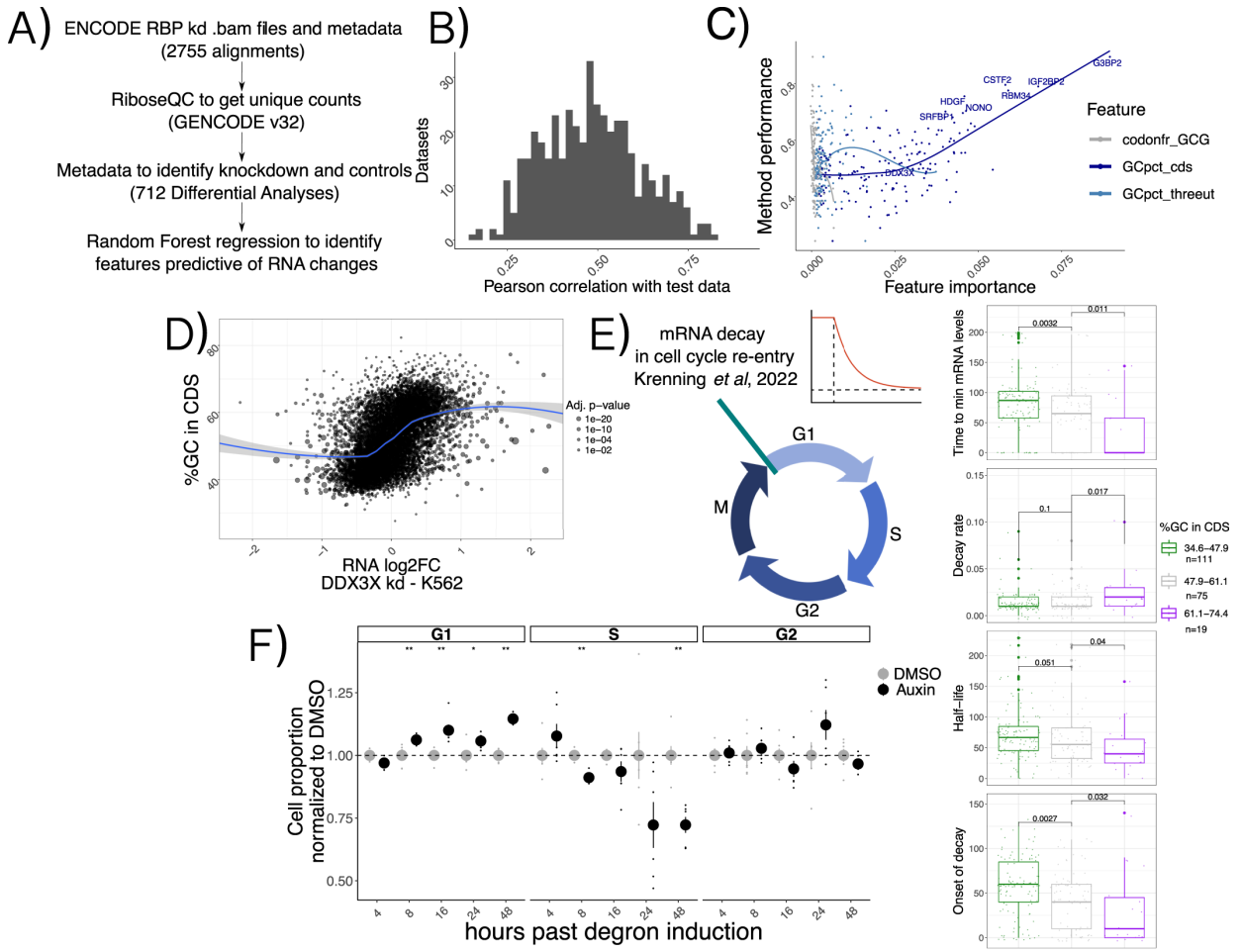


Figure 4: A ubiquitous feature in mRNA regulation.

(A) Schema describing the ENCODE analysis strategy.

(B) Histogram representing overall model performance across datasets.

(C) Model performance (spearman correlation between predicted and real values on unseen test data) on the *y* axis, with importance of 3 example features variables (indicating their predictive value) on the *x* axis. Top knockdown experiments, together with DDX3X, are show with labels. Data shown are from shRNA KD experiments in K562 cells. The linear relationship between GCcds importance and model performance indicates its relevance as the top predictor of RNA changes in dozens of datasets.

(D) mRNA level changes against GCcds values in a DDX3X knockdown experiment in the ENCODE dataset; the size of the dots represents statistical significance of the differential expression test vs. control samples, as defined by DESeq2.

(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. For all box plots, in each box, the central black line indicates the median, and the upper and lower edges denote the 25th and 75th percentiles. Whiskers represent 1.5 times the interquartile range from the box hinge. The number of mRNAs analyzed is shown in the legend box.

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

(F) Normalized cell proportion (obtained by dividing cell percentages between Auxin treatment and DMSO) in different stages of the cell cycle along the degron time course. An increase in G1 and decrease in S phase can be observed at later time points. Significance values calculated from a Wilcoxon two-sided test. Data from two experiments performed in triplicate, $n = 6$. Error bars represent the standard error of the mean.

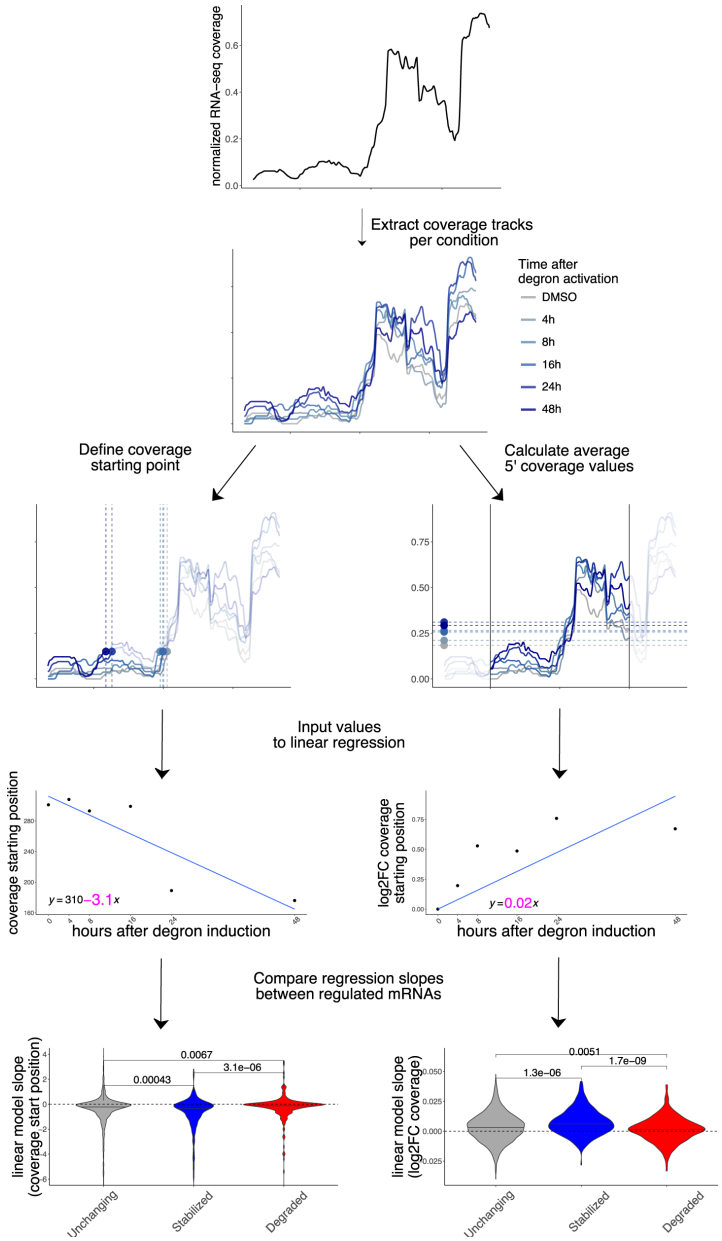


Figure 5: Coverage analysis of regulated mRNAs reveals changes in 5'-3' decay.

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 (in transcript coordinates) and coverage values (log2FC to DMSO) are calculated for each time point and used as input to a linear model. The beta coefficient (shown 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. P-values from one-sided Wilcoxon test.

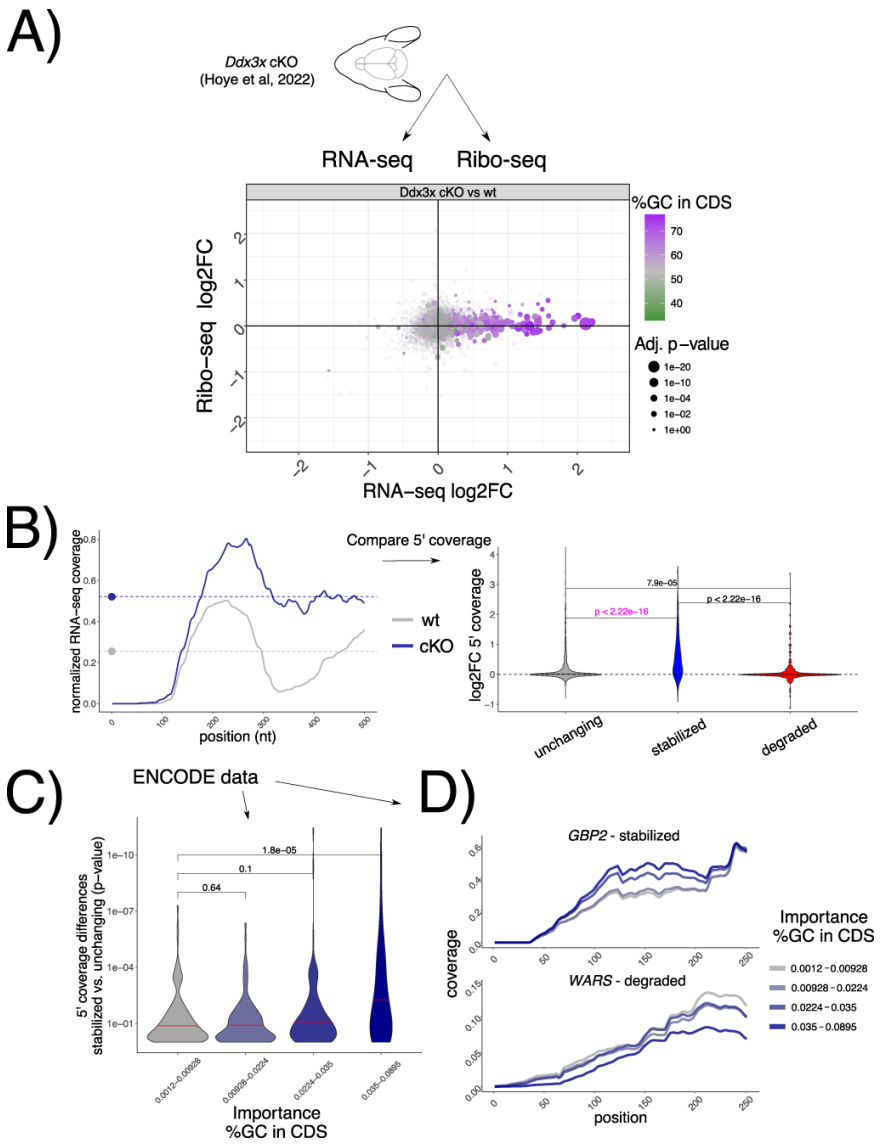


Figure 6: Genes predicted to undergo nuclear degradation (PUNDs) exhibit unique phenotypes related to RNA flow, splicing, and poly(A) tail lengths.

(A) Changes in Ribo-seq and RNA-seq levels in a conditional Ddx3x mouse model, showing GCcds values. The size of the dots represents statistical significance from a differential translation efficiency test, as in Fig. 1A.

(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.

(C) The same strategy as in Fig. 5 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). Significance values from a one-sided Wilcoxon test.

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)

(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. (D) Median poly(A) tail length of PUND genes relative to others across subcellular compartments. The median poly(A) tail length was calculated for each gene covered by >= 10 reads in each sample. Tail lengths were compared between PUND and other genes, and significance was noted as in (A).

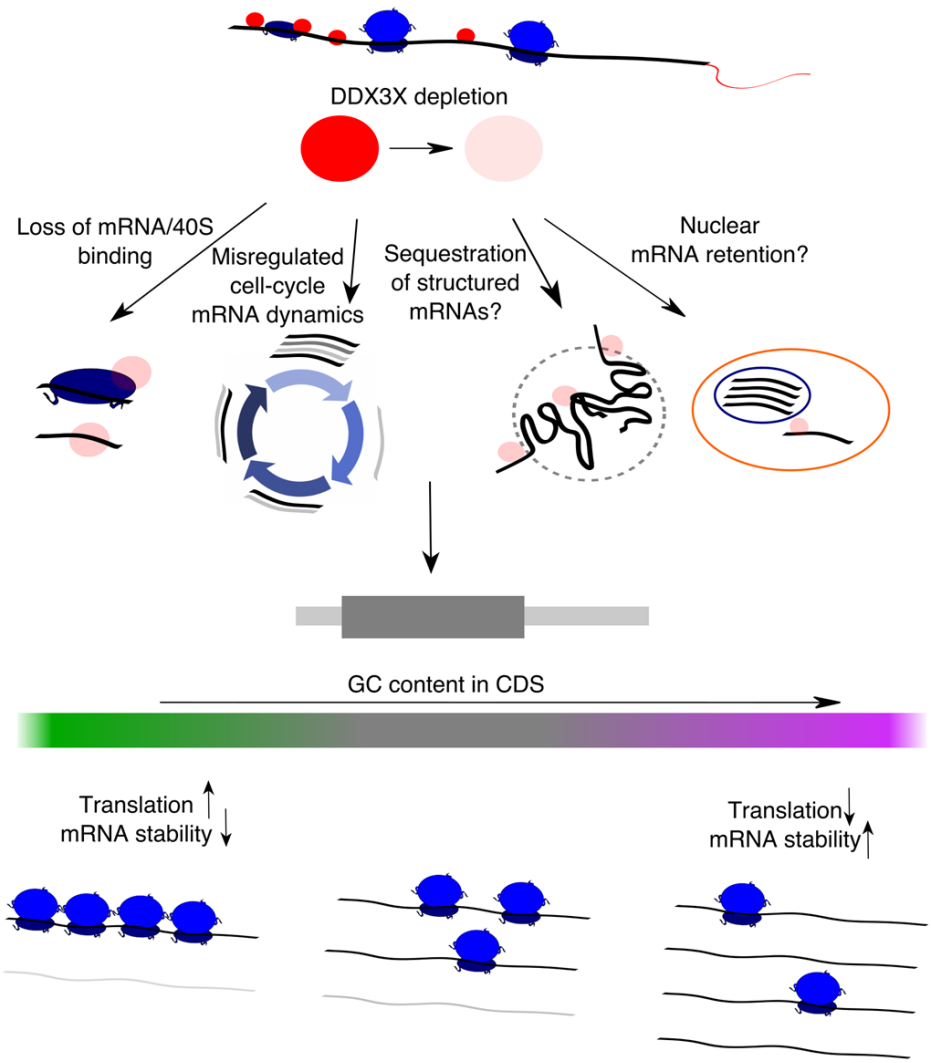
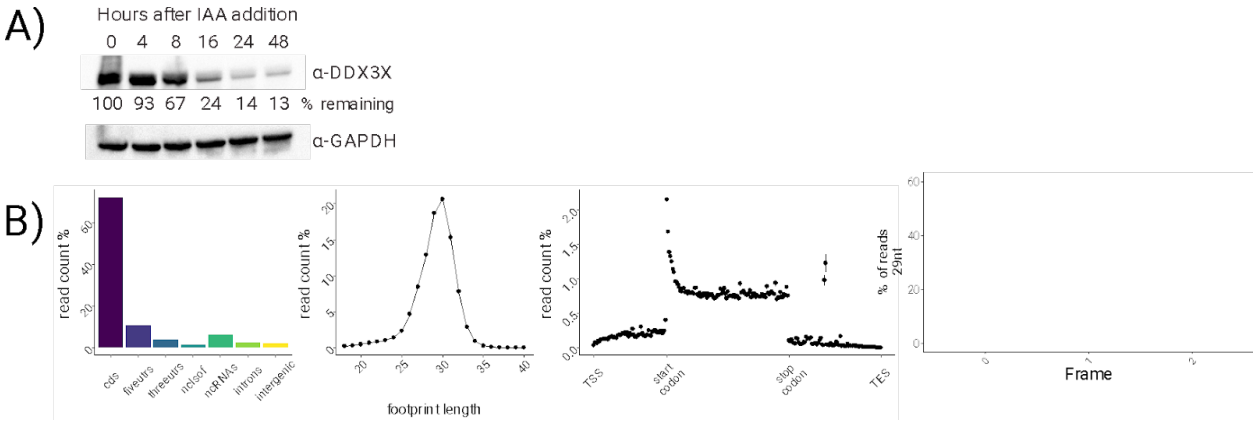


Figure 7: A model for multimodal mRNA regulation by DDX3X.

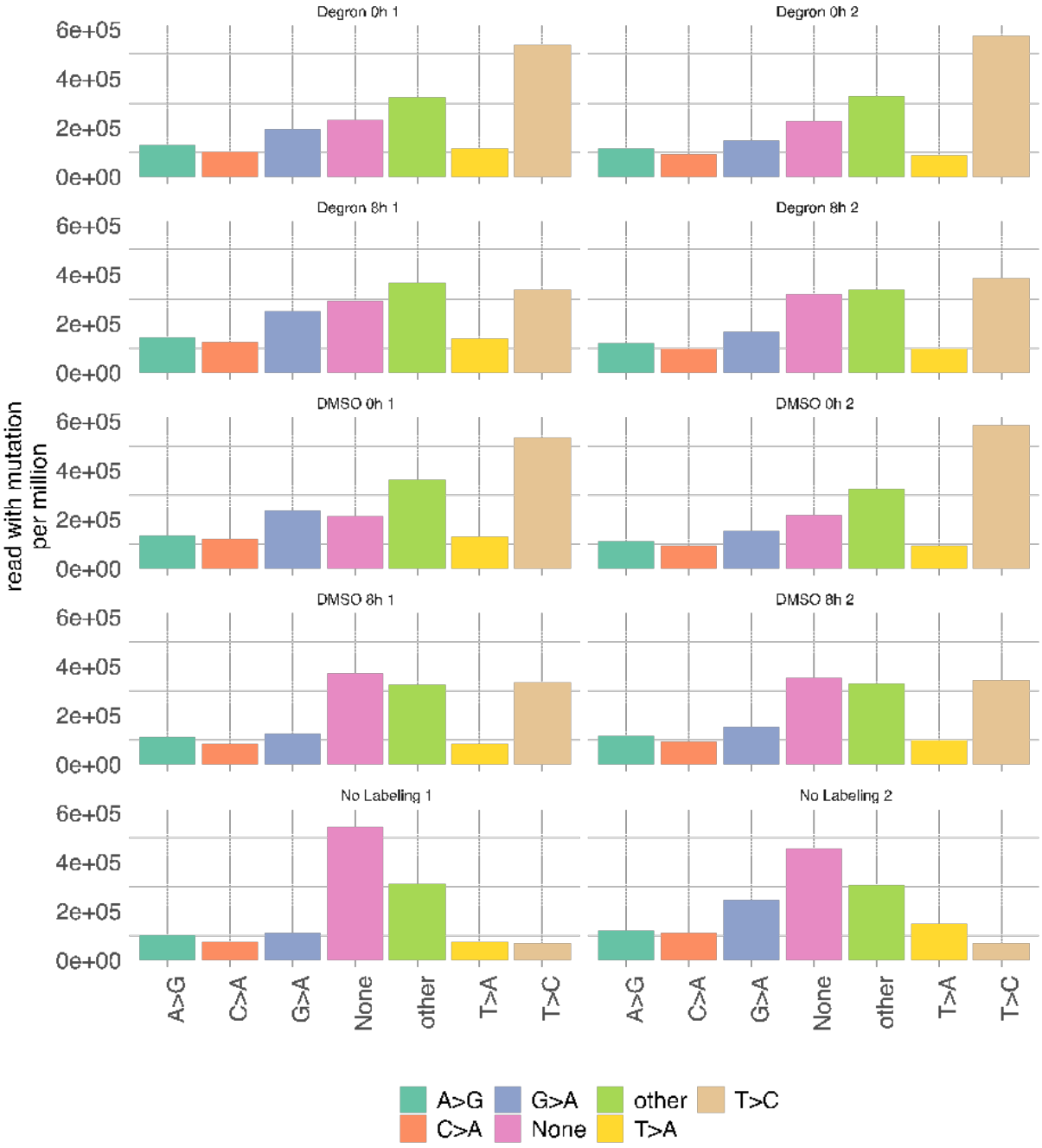
Schematic showing the effects of DDX3X depletion on GC-content related changes in translation and mRNA stability, highlighting potential molecular mechanisms underlying this phenomenon.



Supplemental Figure 1: Degron and Ribo-seq quality control.

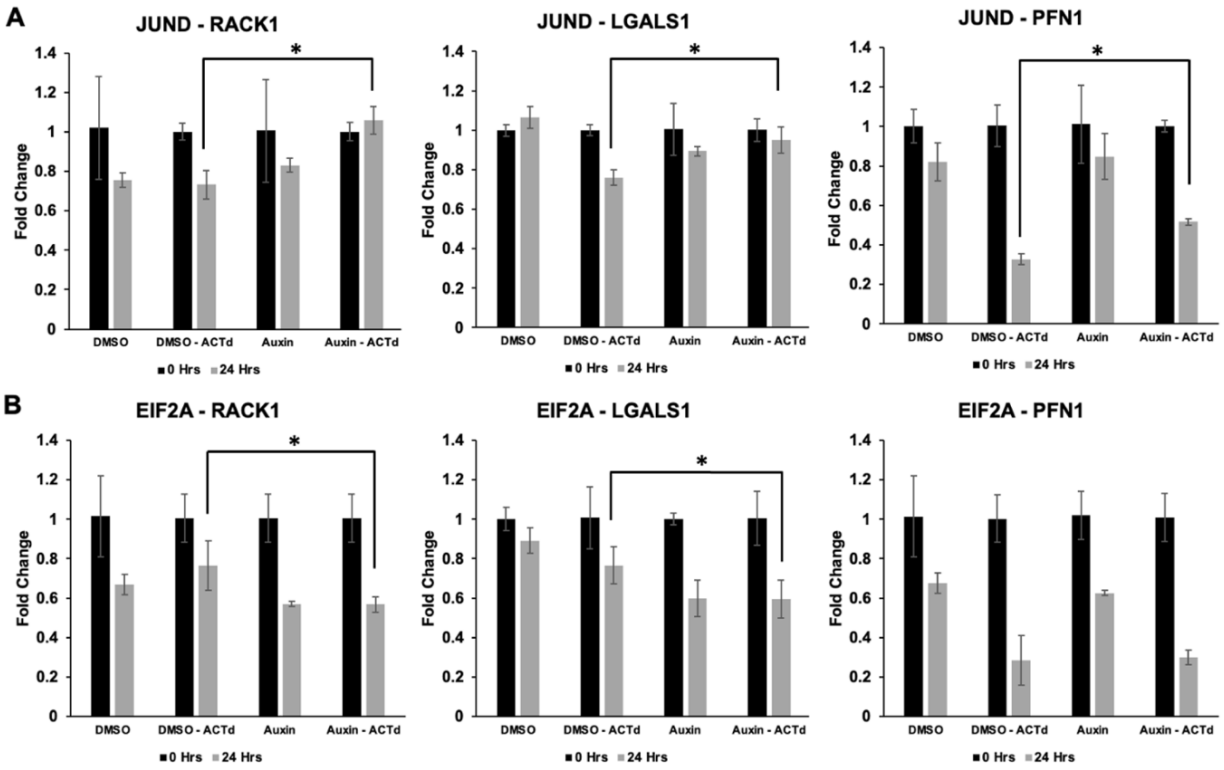
(A) Immunoblot indicating DDX3X and GAPDH levels during the IAA time course. Percent DDX3X is calculated as the DDX3 intensity normalized to GAPDH across two replicates.

(B) Read mapping locations are shown on the far left, followed by footprint lengths distributions, a metatranscript coverage plot, and frame resolution analysis.



Supplemental Figure 2: SLAM-seq quality control.

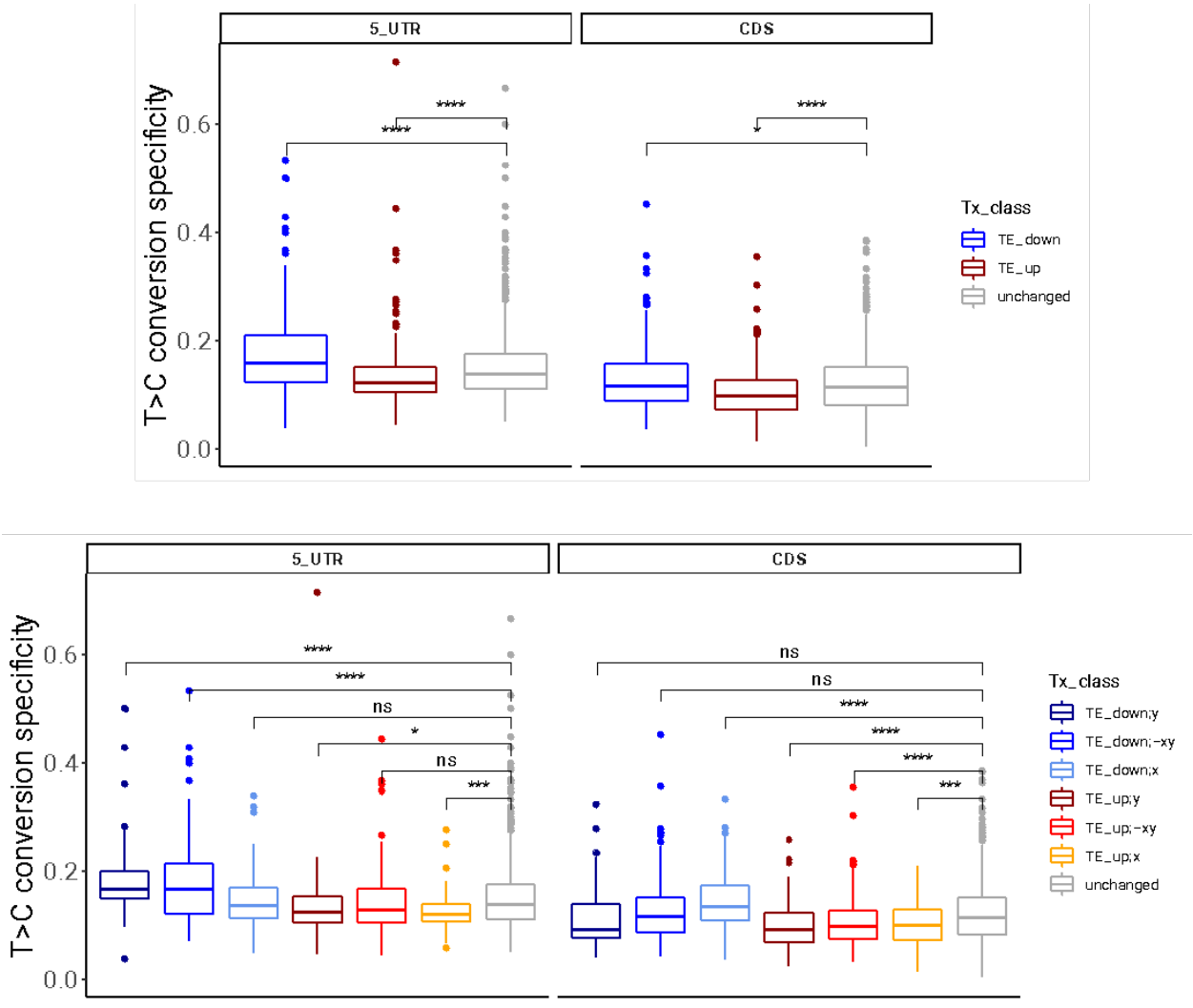
Read counts harboring different mutations are shown across datasets. The SLAM-seq dataset contained background sequencing or RT errors, leading to elevated mutations other than T>C, even without 4sU labeling. However, a decrease in T>C harboring reads can be observed (indicating mRNA decay) for the 8h time points.



Supplemental Figure 3: qPCR validation of mRNA stability changes.

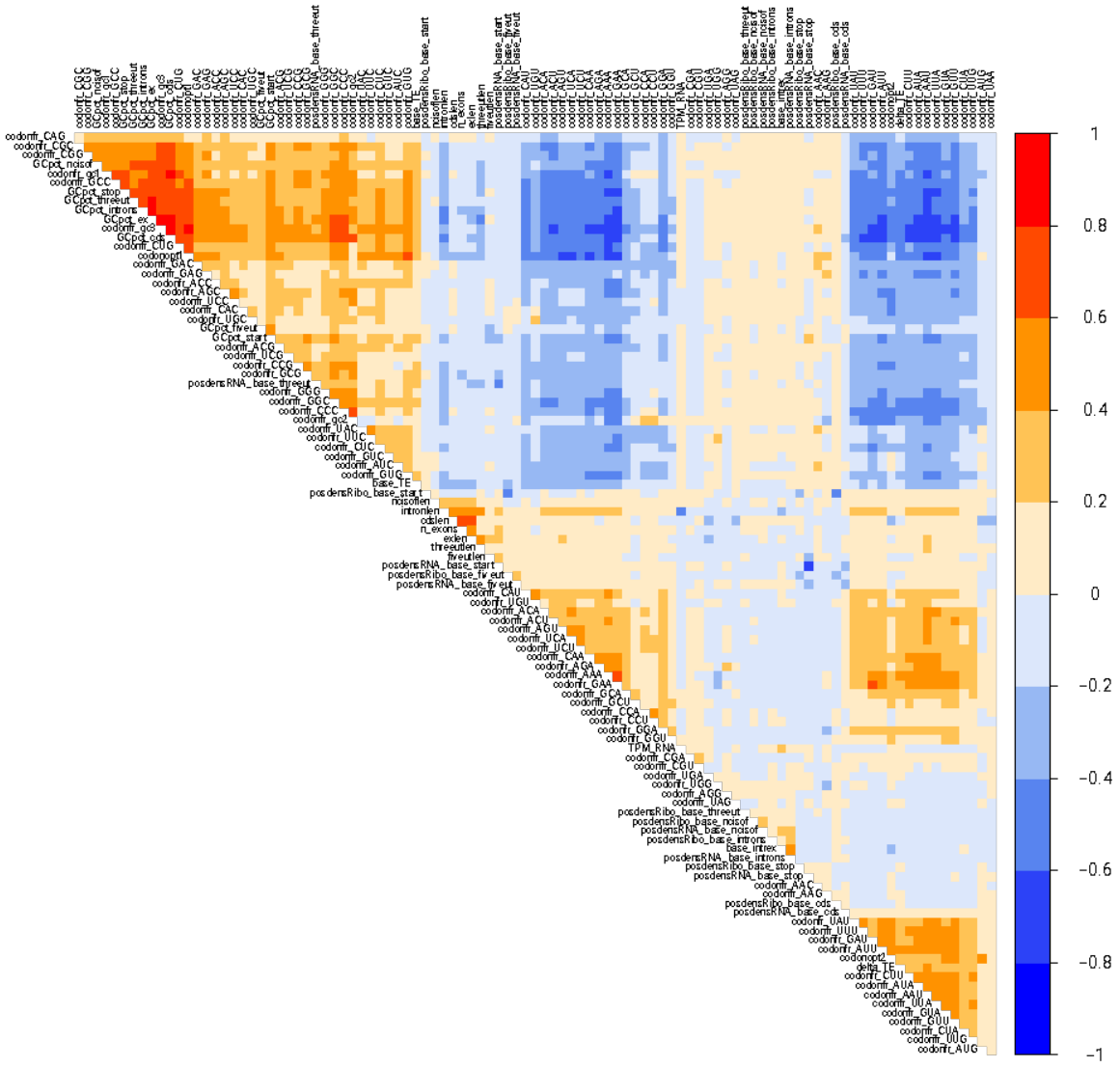
(A) Histogram representing JUND expression by RT-PCR via taqman probes. Fold change was normalized to RACK1, LGALS1, or PFN1. *p-value < 0.05. Performed in triplicate.

(B) Histogram representing EIF2a expression by RT-PCR via taqman probes. Fold change was normalized to RACK1, LGALS1, or PFN1. *p-value < 0.05. Performed in triplicate.



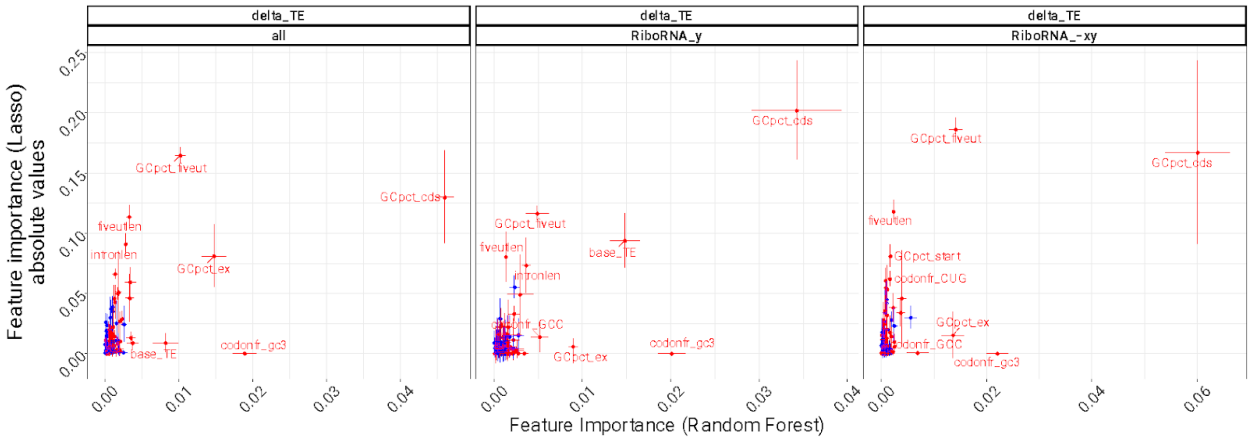
Supplemental Figure 4: mRNA binding pattern on different regulated mRNAs.

T>C conversion specificity on the y-axis (as defined in Calviello and Venkataramanan et al, NAR 2021) is plotted against different regulated transcripts, for both 5'UTR and CDS peaks. Significance values come from two-sided Wilcoxon test against the control gene group.



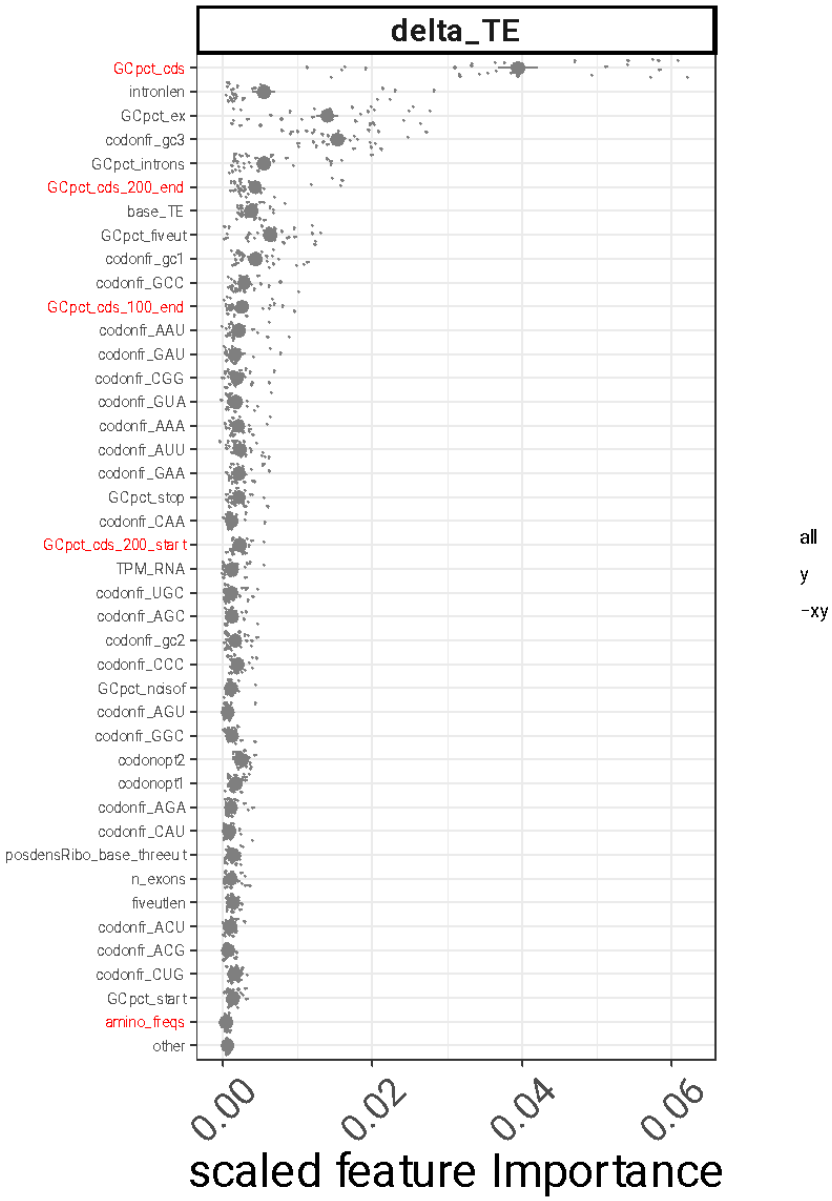
Supplemental Figure 5: Correlation among different mRNA features.

Pearson correlation values across the features used as input for the Random Forest and Lasso regression models. As expected, many features (e.g. GC content along transcripts regions and GC-rich codon frequencies) show high level of correlation.



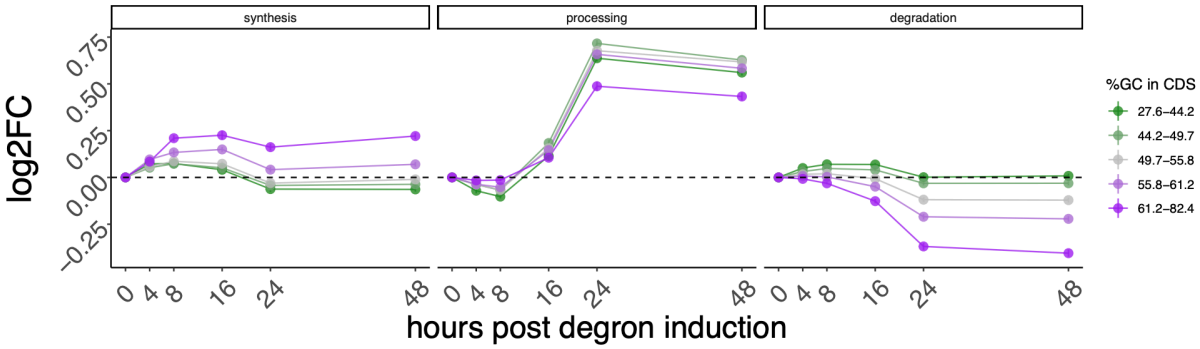
Supplemental Figure 6: Comparison between Lasso and Random Forest feature selection results.

Feature importance according to the Random forest on the x axis, with Lasso coefficients (taking the absolute value) on the y axis. Error bars calculated from 5-fold cross-validation estimates. In red the most relevant features.



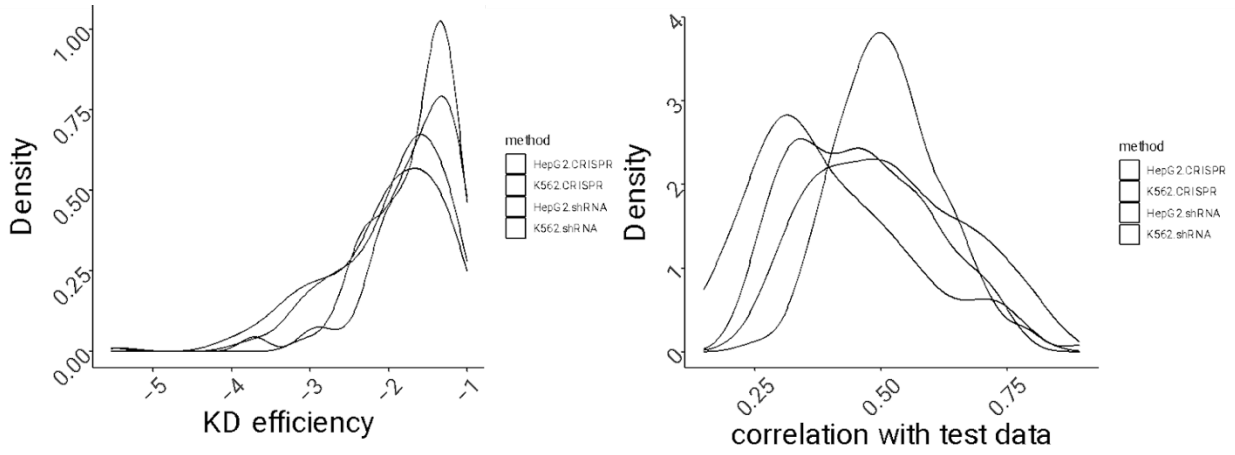
Supplemental Figure 7: Importance plot with additional features.

Predictive power of different features in quantifying translation regulation, with their importance values plotted on the x axis. New variables, alongside GCpct_cds, are colored in red.



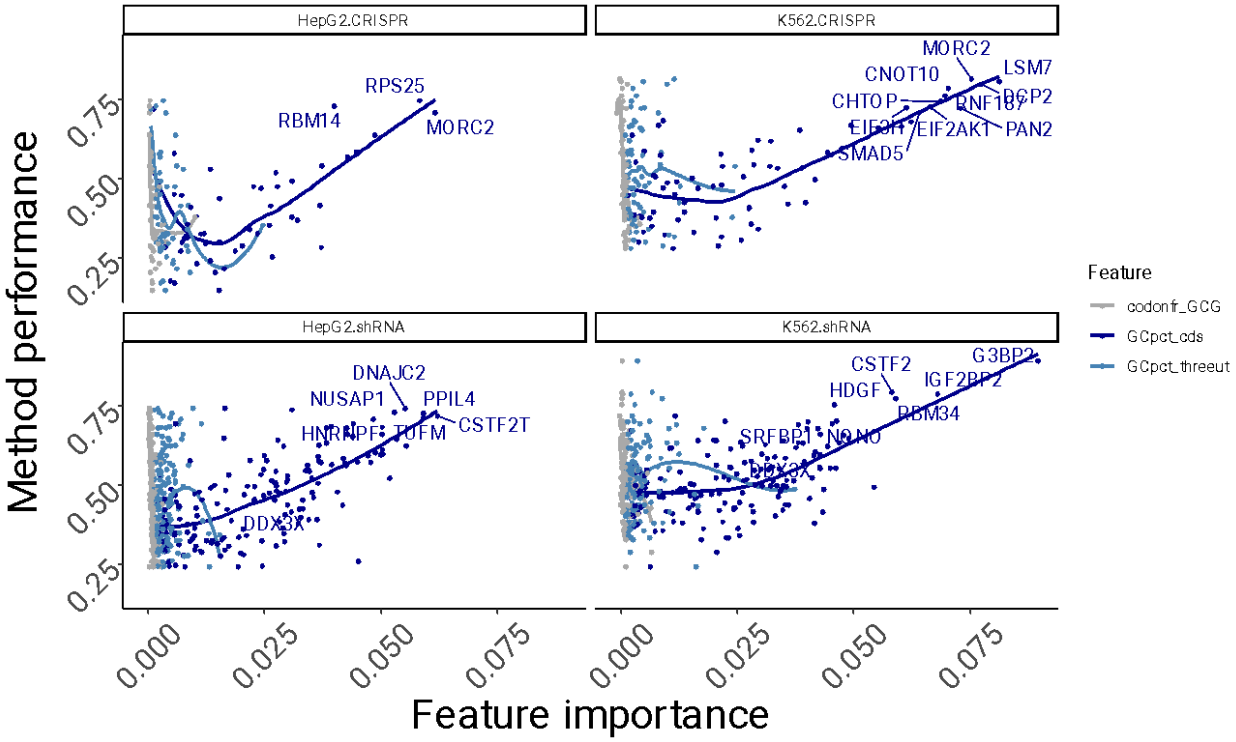
Supplemental Figure 8: mRNA dynamics divided by GCcds values.

Synthesis, processing and degradation values, as inferred by INSPEcT, partitioned by GCcds values.



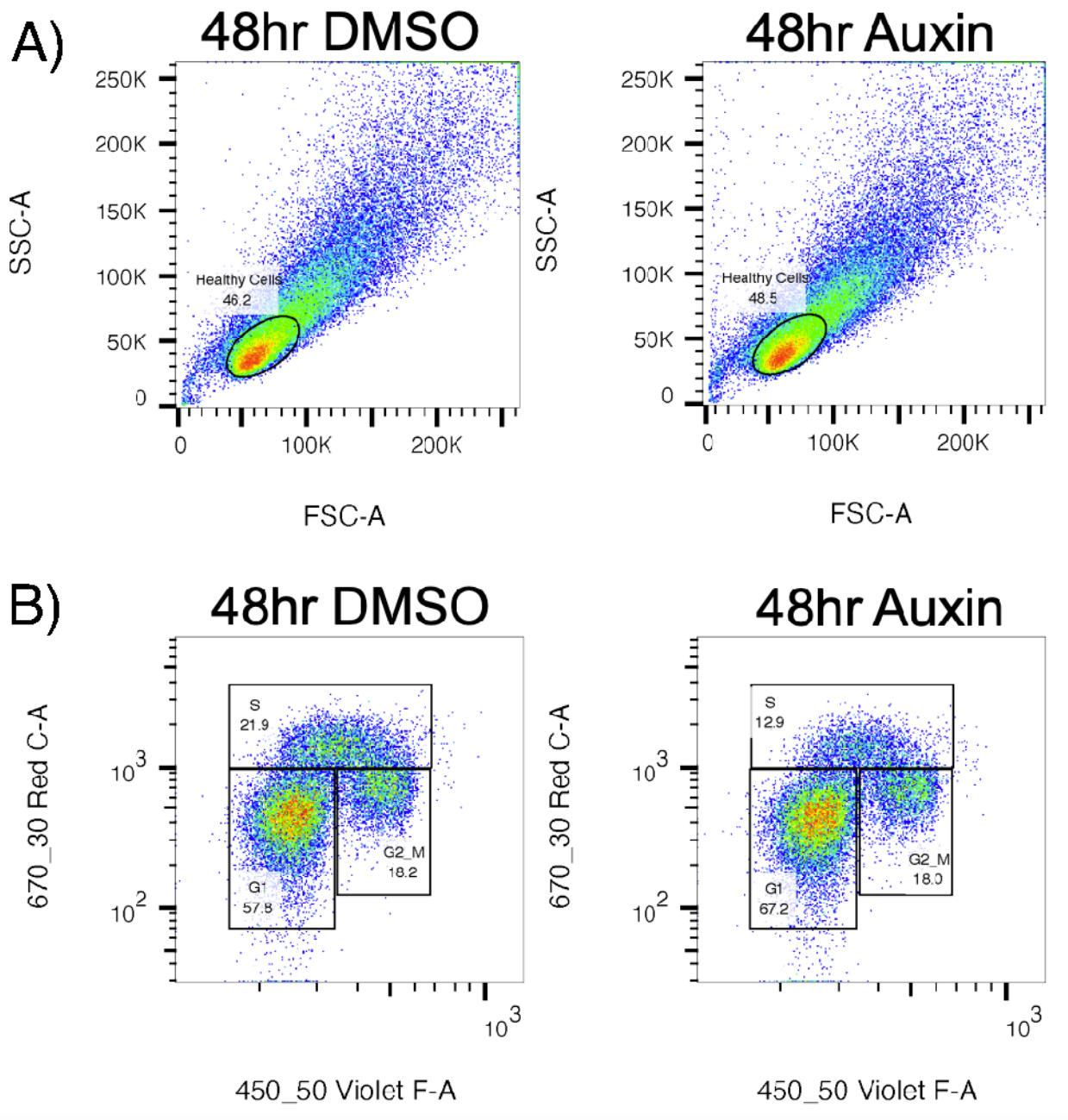
Supplemental Figure 9: ENCODE RBP data analysis overview.

RBP knockdown efficiency (log2FC) across methodologies in the ENCODE dataset (on the left). On the right, performance of the Random Forest model across ENCODE knockdown methods and cell lines.



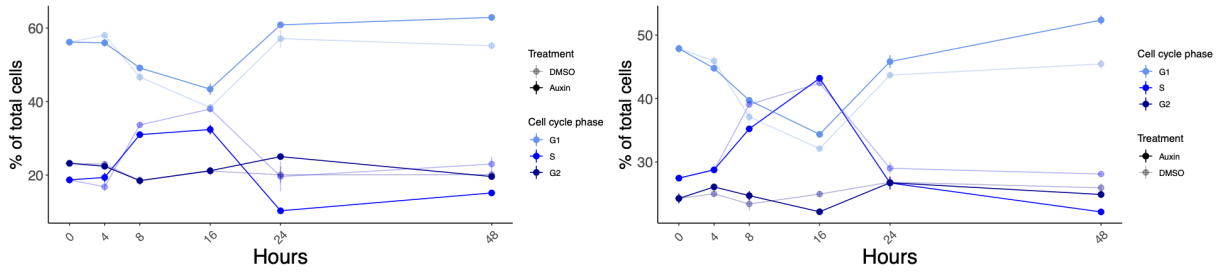
Supplemental Figure 10: GCcds importance across the entire ENCODE dataset.

Model performance (spearman correlation between predicted and real values on unseen test data) on the y axis, with importance of 3 example features variables (indicating their predictive value) on the x axis. Top knockdown experiments, together with DDX3X, are show with labels. Data shown separately for each ENCODE knockdown strategy. The linear relationship between GCcds importance and model performance indicates its relevance as the top predictor of RNA changes in dozens of datasets.



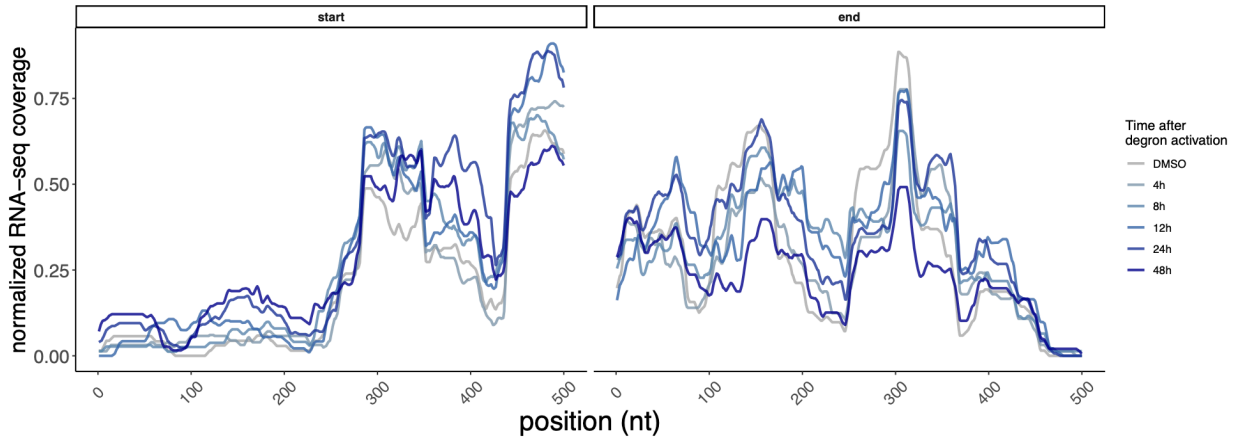
Supplemental Figure 11: Cell cycle staging analysis.

Cell-cycle analysis of HCT116-DDX3-degron cells treated with either DMSO or Auxin (500uM) for 48 hrs. 1 representative replicate of 3 total replicates is shown. A) Single healthy cells were gated on their Forward (FSC-A) and Side Scatter (SSC-A), B) Cells currently undergoing DNA synthesis incorporate EdU (670-30 Red C-A Channel) and FxCycle Violet Stain (450-50 Violet F-A Channel) labels overall DNA content. Cells can then be separated into those in G1, S, and G2-M phases of the cell cycle.



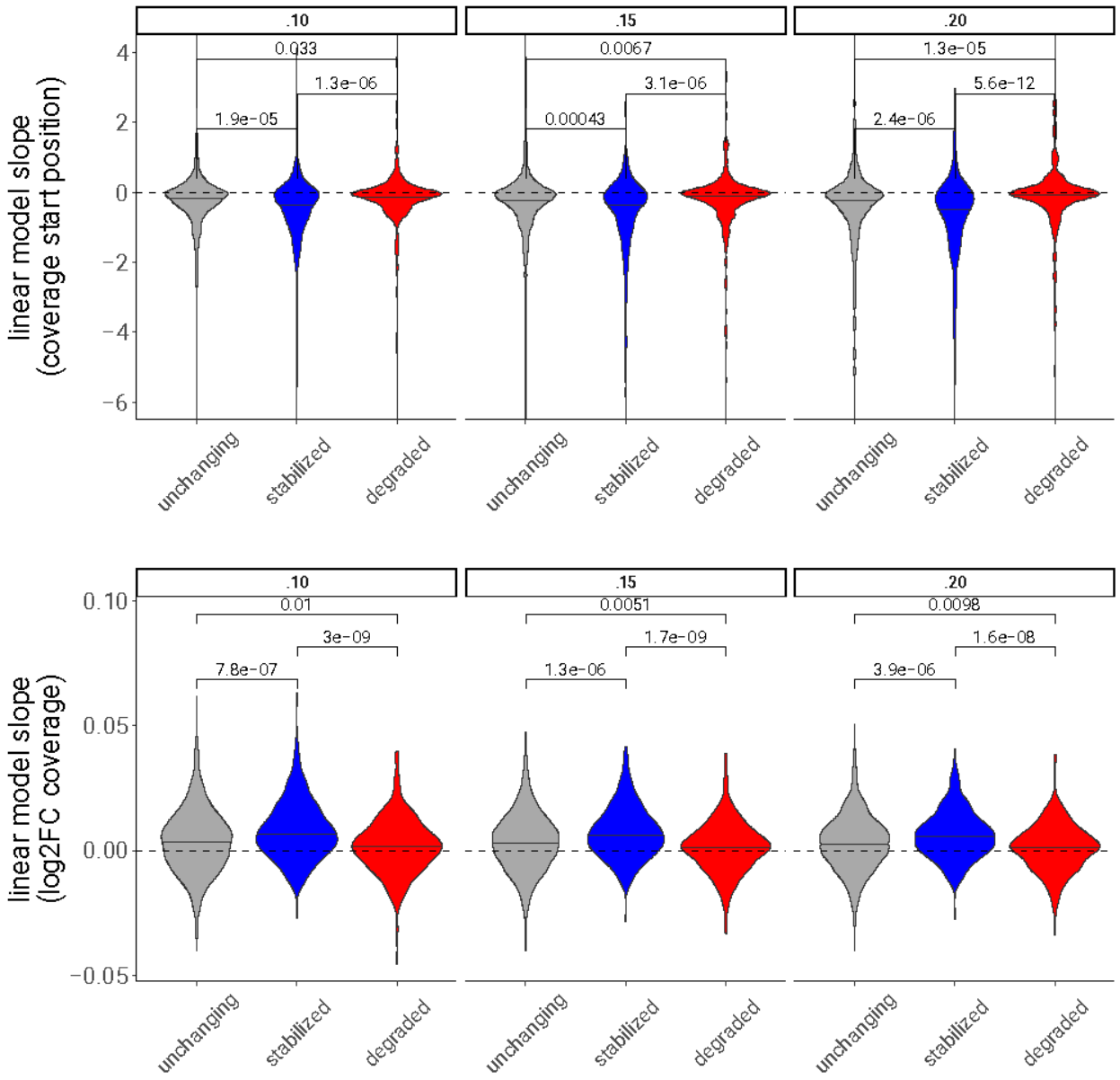
Supplemental Figure 12: Cell cycle dynamics after DDX3X degradation.

Percentage of cells in different stages of the cell cycle along the degron time course. Two independent experiments are shown. Values for degron and control are shown with different transparency values.



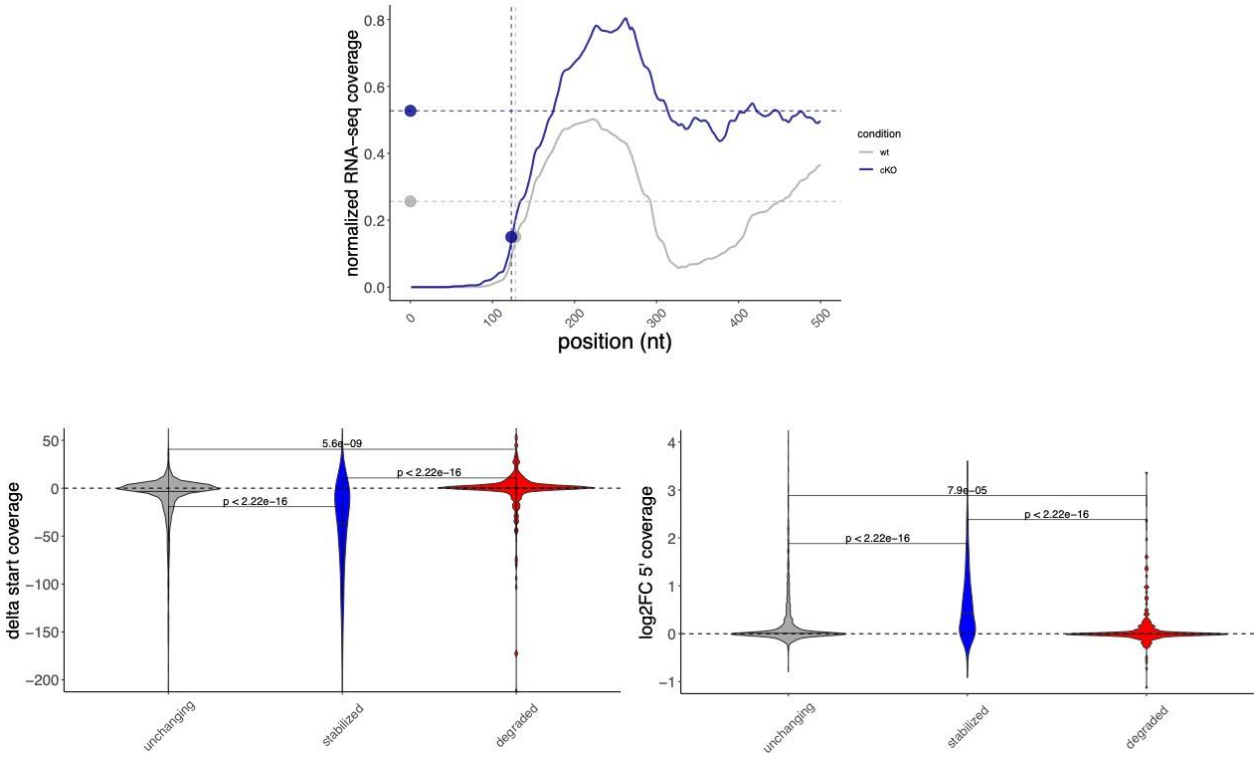
Supplemental Figure 13: Example of RNA-seq coverage changes across the DDX3X degron time course.

RNA-seq coverage tracks around 5' and 3' ends of the stabilized mRNA from the CSRNP2 gene. Coverage values were 0-1 normalized for each dataset.



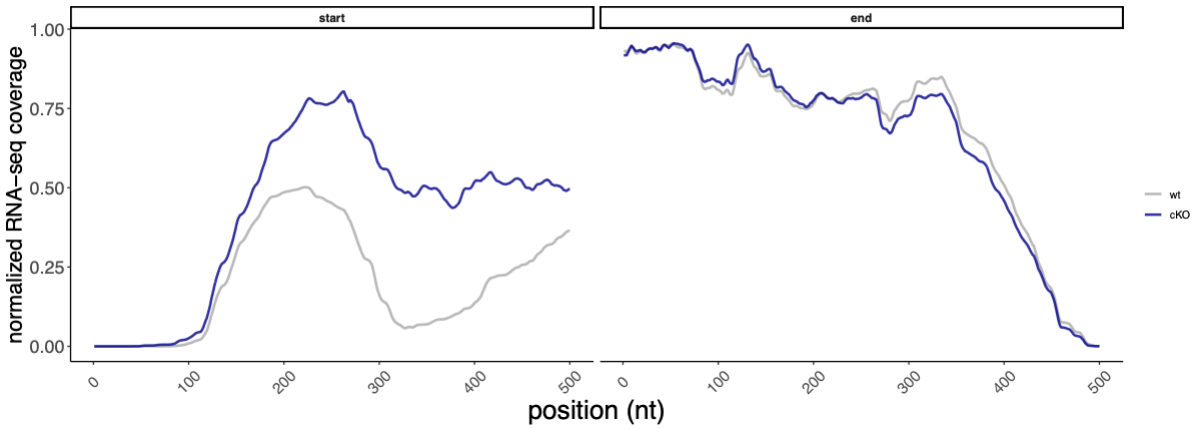
Supplemental Figure 14: Coverage differences between mRNAs are similar using different 5' cutoffs.

Differences between RNA-seq coverage in stabilized, unchanging and degraded mRNAs when using different cutoffs to define coverage starting points. Same strategy as shown in Figure 5A.



Supplemental Figure 15: Changes in RNA-seq coverage values and starting positions in the Ddx3x cKO mouse.

Exemplified strategy using an mRNA from the Ctxn1 gene. Differences in coverage values (log2FC) and starting points are extracted and compared across regulated mRNAs (Methods). Coverage values were 0-1 normalized for each dataset.



Supplemental Figure 16: Example of RNA-seq coverage changes in the Ddx3x cKO mouse.
RNA-seq coverage tracks around 5' and 3' ends of the stabilized mRNA from the Ctxn1 gene. Coverage values were 0-1 normalized for each dataset.

TABLES

Table 1: Reagents and tools.

Reagent/Resource	Reference or Source	Identifier or Catalog Number
Experimental Models		
HCT116 DDX3X-mAID	Venkataramanan et al, 2021	
Antibodies		
anti-GAPDH	Rockland Immunochemicals	Cat# 600-401-A33S
Anti-DDX3	Calviello et al, 2021	custom made by Genemed Synthesis using peptide ENALGLDQQFAGLDLNSSDNQS
Oligonucleotides and other sequence-based reagents		
TaqMan probe RACK1	Thermo Scientific	Chr.5: 181236928 – 181243906 - Hs00272002_m1 -VIC-MGB
TaqMan probe LGALS1	Thermo Scientific	Chr.22: 37675606 – 37679802 - Hs00355202_m1 - VIC-MGB
TaqMan probe PFN1	Thermo Scientific	Chr.17: 4945650 – 4949088 - Hs07291746_gH - VIC-MGB
TaqMan probe JUND	Thermo Scientific	Chr.19: 18279694 – 18281656 - Hs04187679_s1 - FAM-MGB
TaqMan probe EIF2A	Thermo Scientific	Chr.3: 150546678 – 150586016 - Hs00230684_m1 - FAM-MGB
Chemicals, Enzymes and other reagents		
IAA (Indole-3-acetic acid)	Research Products International	Cat# I54000-5.0
RNase I	Ambion	Cat# AM2294

Reagent/Resource	Reference or Source	Identifier or Catalog Number
SUPERase	Ambion	Cat# AM2694
T4 RNA Ligase 2 truncated KQ	NEB	Cat# M0373L
Superscript III	Invitrogen	
CircLigase II	Lucigen	Cat# CL4115K
Software		
FACS DIVA	BD Life Sciences	
FlowJo V10	BD Life Sciences	
bowtie2	Langmead et al, 2009	
STAR	Dobin et al, 2013	
GenomicAlignments	Huber et al, 2015	
GenomicFiles	Huber et al, 2015	
INSPEcT	De Pretis et al, 2015	
RiboseQC	Calviello et al, 2019	
GenomicFeatures	Lawrence et al, 2013	
rtracklayer	Lawrence et al, 2009	
DESeq2	Love et al, 2014	
randomForest	Wiener, 2002	
glmnet	Friedman et al, 2010	
Other		
Illustra Microspin Columns S-400 HR	GE Healthcare	
Direct-zol kit	Zymo Research	

Reagent/Resource	Reference or Source	Identifier or Catalog Number
SLAMseq Kinetics Kit – Catabolic Kinetics Module, 24 preps	Lexogen	SKU 062.24
QuantSeq 3′ mRNA-Seq Library Prep Kit FWD for Illumina	Lexogen	
Click-iT™ Plus EdU Alexa Fluor™ 647 Flow Cytometry Assay Kit	Thermo Fisher	Cat# C10634
FxCycle Violet DNA content stain	Thermo Fisher	Cat# F10347
TruSeq Stranded Total RNA Human/Mouse/Rat kit	Illumina	
TaqMan® real-time PCR	Thermo Fisher	

Table 2: Read mapping statistics for the Ribo-seq RNA-seq DDX3X time course dataset.

Dataset	Total_reads	Clean_reads	Mapped_reads	Unq_mapping_reads
Ribo_04h_rep1	10655141	4665431	4441819	2307930
Ribo_04h_rep2	10247253	4588724	4348254	2159983
Ribo_08h_rep1	12297638	5593107	5355074	2851328
Ribo_08h_rep2	11752754	5222312	4942382	2466369
Ribo_16h_rep1	9057524	3636951	3471776	1650065
Ribo_16h_rep2	9830466	3630795	3446735	1598399
Ribo_24h_rep1	5184826	2399425	2141182	689662
Ribo_24h_rep2	6313788	3164638	2774950	765337
Ribo_48h_rep1	13532132	5169773	4909653	2116900
Ribo_48h_rep2	11220136	4200653	3991891	1588211
Ribo_DMSO_rep1	7766903	3696158	3272945	1379069
Ribo_DMSO_rep2	7443328	2699121	2385696	1048391
RNA_04h_rep1	10866243	10685062	10415938	8845500
RNA_04h_rep2	13433217	13238485	12968820	10898980
RNA_08h_rep1	11016806	10823951	10623083	9091257
RNA_08h_rep2	14590240	14412411	14166210	12172239
RNA_16h_rep1	11320974	11180183	10795246	9359751
RNA_16h_rep2	11184620	11061530	10688412	9204076

Dataset	Total_reads	Clean_reads	Mapped_reads	Unq_mapping_reads
RNA_24h_rep1	12251536	12127901	11849975	9911934
RNA_24h_rep2	13780681	13600543	13434733	11318481
RNA_48h_rep1	11724669	11140330	10985698	9026574
RNA_48h_rep2	10966544	10819673	10412233	8655942
RNA_DMSO_rep1	14673939	14484063	14233622	11889196
RNA_DMSO_rep2	16956673	16729001	16436720	13715774

Table 3: Input to the Random Forest model for the DD3X3 time course dataset.

Due to large file size datasets for can be found at: Jowhar et al. 2024 in Supplementary Material named DATASET EV2.

Table 4: Accession codes for the analyzed ENCODE datasets, with information for each differential analysis. Multiple accession can be technical replicate of a biological replicate.

Due to large file size datasets for can be found at: Jowhar et al. 2024 in Supplementary Material named DATASET EV3.

Table 5: Input to the Random Forest model for the cKO Ddx3x mouse dataset.

Due to large file size datasets for can be found at: Jowhar et al. 2024 in Supplementary Material named DATASET EV4.

REFERENCES

1. Shoemaker CJ, Green R. Translation drives mRNA quality control. *Nat Struct Mol Biol*. 2012;19(6):594-601. doi:10.1038/nsmb.2301

2. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc*. 2012;7(8):1534-1550. doi:10.1038/nprot.2012.086

3. Herzog VA, Reichholf B, Neumann T, et al. Thiol-linked alkylation of RNA to assess expression dynamics. *Nat Methods*. 2017;14(12):1198-1204. doi:10.1038/nmeth.4435

4. Hafner M, Katsantoni M, Köster T, et al. CLIP and complementary methods. *Nature Reviews Methods Primers*. 2021;1(1):20. doi:10.1038/s43586-021-00018-1

5. Sharma D, Jankowsky E. The Ded1/DDX3 subfamily of DEAD-box RNA helicases. *Crit Rev Biochem Mol Biol*. 2014;49(4):343-360. doi:10.3109/10409238.2014.931339

6. Oh S, Flynn RA, Floor SN, et al. Medulloblastoma-associated DDX3 variant selectively alters the translational response to stress. *Oncotarget*. 2016;7(19):28169. doi:10.18632/ONCOTARGET.8612

7. Calviello L, Venkataramanan S, Rogowski KJ, et al. DDX3 depletion represses translation of mRNAs with complex 5′ UTRs. *Nucleic Acids Res*. 2021;49(9):5336-5350. doi:10.1093/NAR/GKAB287

8. Chang YF, Imam JS, Wilkinson MF. The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem*. 2007;76:51-74. doi:10.1146/ANNUREV.BIOCHEM.76.050106.093909

9. D'Orazio Colin Chih Chien; Sinha Niladri K.; Loll-Krippleber Raphaël; Brown Grant W.; Green Rachel KN; W. The endonuclease cue2 cleaves mRNAs at stalled ribosomes during no go decay. *Elife*. 2019;8(NA):NA-NA. doi:10.7554/elife.49117

10. Bazzini AA, Lee MT, Giraldez AJ. Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. *Science*. 2012;336(6078):233-237. doi:10.1126/science.1215704

11. Gadek M, Sherr EH, Floor SN. The variant landscape and function of DDX3X in cancer and neurodevelopmental disorders. *Trends Mol Med*. 2023;29(9):726-739. doi:10.1016/j.molmed.2023.06.003

12. Lennox AL, Hoye ML, Jiang R, et al. Pathogenic DDX3X Mutations Impair RNA Metabolism and Neurogenesis during Fetal Cortical Development. *Neuron*. 2020;106(3):404-420.e8. doi:10.1016/j.neuron.2020.01.042

13. Ingolia NT. Ribosome Footprint Profiling of Translation throughout the Genome. *Cell*. 2016;165(1):22-33. doi:10.1016/j.cell.2016.02.066

14. Lorent J, Kusnadi EP, van Hoef V, et al. Translational offsetting as a mode of estrogen receptor α-dependent regulation of gene expression. *EMBO J*. 2019;38(23). doi:10.15252/EMBJ.2018101323

15. Hoye ML, Calviello L, Poff AJ, et al. Aberrant cortical development is driven by impaired cell cycle and translational control in a DDX3X syndrome model. *Elife*. 2022;11. doi:10.7554/eLife.78203

16. Van Nostrand EL, Freese P, Pratt GA, et al. A large-scale binding and functional map of human RNA-binding proteins. *Nature*. 2020;583(7818):711-719. doi:10.1038/s41586-020-2077-3

17. Hon CC, Carninci P. Expanded ENCODE delivers invaluable genomic encyclopedia. *Nature*. 2020;583(7818):685-686. doi:10.1038/d41586-020-02139-1

18. Venkataramanan S, Gadek M, Calviello L, Wilkins K, Floor SN. DDX3X and DDX3Y are redundant in protein synthesis. *RNA*. 2021;27(12):1577-1588. doi:10.1261/rna.078926.121

19. de Pretis S, Kress T, Morelli MJ, et al. INSPEcT: a computational tool to infer mRNA synthesis, processing and degradation dynamics from RNA- and 4sU-seq time course experiments. *Bioinformatics*. 2015;31(17):2829-2835. doi:10.1093/bioinformatics/btv288

20. Medina-Muñoz SG, Kushawah G, Castellano LA, et al. Crosstalk between codon optimality and cis-regulatory elements dictates mRNA stability. *Genome Biol*. 2021;22(1):14. doi:10.1186/s13059-020-02251-5

21. Courel M, Clément Y, Bossevain C, et al. GC content shapes mRNA storage and decay in human cells. *Elife*. 2019;8. doi:10.7554/eLife.49708

22. Hia F, Yang SF, Shichino Y, et al. Codon bias confers stability to human mRNAs. *EMBO Rep*. 2019;20(11):e48220. doi:10.15252/embr.201948220

23. Krenning L, Sonneveld S, Tanenbaum ME. Time-resolved single-cell sequencing identifies multiple waves of mRNA decay during the mitosis-to-G1 phase transition. *Elife*. 2022;11. doi:10.7554/eLife.71356

24. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. *Nat Rev Genet*. 2014;15(12):829-845. doi:10.1038/nrg3813

25. Avsec Ž, Agarwal V, Visentin D, et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods*. 2021;18(10):1196-1203. doi:10.1038/s41592-021-01252-x

26. Thomas A, Rehfeld F, Zhang H, et al. RBM33 directs the nuclear export of transcripts containing GC-rich elements. *Genes Dev*. 2022;36(9-10):550-565. doi:10.1101/gad.349456.122

27. Chothani SP, Adami E, Widjaja AA, et al. A high-resolution map of human RNA translation. *Mol Cell*. 2022;82(15):2885-2899.e8. doi:10.1016/j.molcel.2022.06.023

28. Van Nostrand EL, Pratt GA, Yee BA, et al. Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. *Genome Biol*. 2020;21(1):90. doi:10.1186/s13059-020-01982-9

29. Zhao W, Zhang S, Zhu Y, et al. POSTAR3: an updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Res*. 2022;50(D1):D287-D294. doi:10.1093/nar/gkab702

30. Bahar Halpern K, Caspi I, Lemze D, et al. Nuclear Retention of mRNA in Mammalian Tissues. *Cell Rep*. 2015;13(12):2653-2662. doi:10.1016/j.celrep.2015.11.036

31. Kotov AA, Olenkina OM, Kibanov M V, Olenina L V. RNA helicase Belle (DDX3) is essential for male germline stem cell maintenance and division in Drosophila. *Biochim Biophys Acta*. 2016;1863(6 Pt A):1093-1105. doi:10.1016/j.bbamcr.2016.02.006

32. Lai MC, Chang WC, Shieh SY, Tarn WY. DDX3 Regulates Cell Growth through Translational Control of Cyclin E1. *Mol Cell Biol*. 2010;30(22):5444-5453. doi:10.1128/MCB.00560-10/SUPPL_FILE/SUPPLEMENTAL_MATERIAL.PDF

33. Tanenbaum ME, Stern-Ginossar N, Weissman JS, Vale RD. Regulation of mRNA translation during mitosis. *Elife*. 2015;4. doi:10.7554/eLife.07957

34. Clemm von Hohenberg K, Müller S, Schleich S, 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*. 2022;13(1):1-14. doi:10.1038/s41467-022-28265-0

35. Harnett D, Ambrozkiewicz MC, Zinnall U, et al. A critical period of translational control during brain development at codon resolution. *Nat Struct Mol Biol*. 2022;29(12):1277-1290. doi:10.1038/s41594-022-00882-9

36. Hoye ML, Silver DL. Decoding mixed messages in the developing cortex: translational regulation of neural progenitor fate. *Curr Opin Neurobiol*. 2021;66:93-102. doi:10.1016/j.conb.2020.10.001

37. Pilaz LJ, McMahon JJ, Miller EE, et al. Prolonged Mitosis of Neural Progenitors Alters Cell Fate in the Developing Brain. *Neuron*. 2016;89(1):83-99. doi:10.1016/j.neuron.2015.12.007

38. Venkataramanan S, Gadek M, Calviello L, Wilkins K, Floor SN. DDX3X and DDX3Y are redundant in protein synthesis. *RNA*. 078926(121).

39. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10(3):R25. doi:10.1186/gb-2009-10-3-r25

40. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635

41. Huber W, Carey VJ, Gentleman R, et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods*. 2015;12(2):115-121. doi:10.1038/nmeth.3252

42. Calviello L, Sydow D, Harnett D, Ohler U. Ribo-seQC: comprehensive analysis of cytoplasmic and organellar ribosome profiling data. doi:10.1101/601468

43. Lawrence M, Huber W, Pagès H, et al. Software for computing and annotating genomic ranges. *PLoS Comput Biol*. 2013;9(8):e1003118. doi:10.1371/journal.pcbi.1003118

44. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics*. 2009;25(14):1841-1842. doi:10.1093/bioinformatics/btp328

45. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8

46. Liaw A, Wiener M. *Classification and Regression by RandomForest*. Vol 2.; 2002. http://www.stat.berkeley.edu/

47. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*. 2010;33(1):1-22.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
DEB822D2A3A54B3...
Author Signature

4/5/2024
Date