


METHODOLOGY ARTICLE

Open Access

# Characterizing RNA stability genome-wide through combined analysis of PRO-seq and RNA-seq data



Amit Blumberg<sup>1†</sup>, Yixin Zhao<sup>1†</sup>, Yi-Fei Huang<sup>1,2</sup>, Noah Dukler<sup>1</sup>, Edward J. Rice<sup>3</sup>, Alexandra G. Chivu<sup>3</sup>, Katie Krumholz<sup>1</sup>, Charles G. Danko<sup>3</sup> and Adam Siepel<sup>1\*</sup> 

## Abstract

**Background:** The concentrations of distinct types of RNA in cells result from a dynamic equilibrium between RNA synthesis and decay. Despite the critical importance of RNA decay rates, current approaches for measuring them are generally labor-intensive, limited in sensitivity, and/or disruptive to normal cellular processes. Here, we introduce a simple method for estimating relative RNA half-lives that is based on two standard and widely available high-throughput assays: Precision Run-On sequencing (PRO-seq) and RNA sequencing (RNA-seq).

**Results:** Our method treats PRO-seq as a measure of transcription rate and RNA-seq as a measure of RNA concentration, and estimates the rate of RNA decay required for a steady-state equilibrium. We show that this approach can be used to assay relative RNA half-lives genome-wide, with good accuracy and sensitivity for both coding and noncoding transcription units. Using a structural equation model (SEM), we test several features of transcription units, nearby DNA sequences, and nearby epigenomic marks for associations with RNA stability after controlling for their effects on transcription. We find that RNA splicing-related features are positively correlated with RNA stability, whereas features related to miRNA binding and DNA methylation are negatively correlated with RNA stability. Furthermore, we find that a measure based on U1 binding and polyadenylation sites distinguishes between unstable noncoding and stable coding transcripts but is not predictive of relative stability within the mRNA or lincRNA classes. We also identify several histone modifications that are associated with RNA stability.

**Conclusion:** We introduce an approach for estimating the relative half-lives of individual RNAs. Together, our estimation method and systematic analysis shed light on the pervasive impacts of RNA stability on cellular RNA concentrations.

**Keywords:** RNA half-life, RNA splicing, Epigenomics, PRO-seq, Structural equation modeling

\* Correspondence: [asiepel@cshl.edu](mailto:asiepel@cshl.edu)

<sup>†</sup>Amit Blumberg and Yixin Zhao contributed equally to this work.

<sup>1</sup>Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

Gene regulation is an exquisitely complex process that operates at all stages of gene expression, ranging from pre-transcriptional chromatin remodeling to post-translational modification of proteins. However, the concentration of RNA molecules in the cell appears to serve as the primary target of many regulatory mechanisms. Many studies of gene regulation focus on the production of RNA, often at the stages of transcriptional pre-initiation, initiation, or release from pausing into productive elongation. RNA concentrations, however, result from a dynamic equilibrium between the production of new RNA molecules and their decay [1–7]. Indeed, bulk differences in RNA concentrations across types of transcription units (TUs) often result from differences in RNA decay rates rather than differences in production rates. For example, protein-coding mRNAs, on average, are relatively stable (having low rates of decay), whereas lincRNAs are less stable, and enhancer RNAs (eRNAs) and other short noncoding RNAs tend to be extremely unstable [3, 6, 8, 9]. Among protein-coding genes, mRNAs associated with housekeeping functions tend to be stable, whereas those associated with regulation of transcription and apoptosis tend to have much shorter half-lives, probably to enable RNA concentrations to change rapidly in response to changing conditions [4, 6, 7, 10, 11]. In some cases, RNA decay is accelerated by condition- or cell type-specific expression of microRNAs or RNA-binding proteins [3, 12].

Over several decades, investigators have developed numerous methods for measuring RNA decay rates or half-lives [13–15]. A classical approach to this problem is to measure the decay in RNA abundance over time following inhibition of transcription, often using actinomycin D [1, 7, 16]. More recently, many studies have employed a strategy that is less disruptive to cellular physiology, based on metabolic labeling of RNA transcripts with modified nucleotides. In this approach, the relative proportions of labeled and unlabeled transcripts are quantified as they change over time, following an initial introduction or removal of labeled nucleotides [6, 15]. Today, metabolic labeling is most commonly accomplished using the nucleotide analog 4-thiouridine (4sU), which is rapidly taken up by animal cells and can be biotinylated for affinity purification [2, 3, 8, 17–19]. Related methods use chemical conversion of 4sU nucleotide analogs to allow identification by sequencing and avoid the need for affinity purification [10, 20]. In most of these assays, sample preparation and sequencing must be performed in a time course, making the protocols labor-intensive and dependent on the availability of abundant and homogeneous sample material (typically

a cell culture). Many of these methods also have limited sensitivity for low-abundance transcripts. Owing to a variety of limitations, estimates of RNA half-lives tend to vary considerably across assays, with median half-lives often differing by factors of 2–3 or more [6, 15]. As yet, there exists no general-purpose assay for RNA half-life that is as robust, sensitive, or versatile as RNA-seq [12, 21, 22] is for measuring cellular RNA concentrations, or PRO-seq [23] and NET-seq [24] are for mapping engaged RNA polymerases.

Recently, it has been shown that changes to RNA half-lives can be identified in a simpler manner, by working directly from high-throughput RNA-seq data [12, 21, 22, 25]. The essential idea behind these methods is to treat RNA-seq read counts obtained from introns as a surrogate for transcription rates, and read counts obtained from exons as a surrogate for RNA abundance. Changes in half-life are then inferred from changes to the ratio of these quantities, under the assumption of a steady-state equilibrium between RNA production and decay. This approach assumes intronic read counts are representative of pre-mRNA abundances, when in fact they may derive from a variety of sources, and it can require a correction for differences in RNA processing rates [21]. Moreover, the dependency on intronic reads limits the method to intron-containing transcription units that are transcribed at relatively high levels. Nevertheless, this simple approach requires no time course, metabolic labeling, transcriptional inhibition, or indeed any experimental innovation beyond standard RNA-seq, making it an inexpensive and effective strategy for identifying genes undergoing cell type- or condition-specific decay [12, 21, 22].

In this article, we show that this same general approach—but using a measure of nascent transcription based on PRO-seq rather than intronic RNA-seq reads—results in improved estimates of relative RNA half-life. Our approach requires only two standard and widely applicable experimental protocols—PRO-seq and RNA-seq. It applies to intron-less as well as intron-containing transcription units, it requires no correction for RNA-processing rates, it makes efficient use of the available sample material and can be extended to tissue samples using ChRO-seq [26], it is relatively nondisruptive to the biological processes under study, and it is sufficiently sensitive to assay TUs expressed at low levels, including many noncoding RNAs (see Additional file 1: Table S1 for a summary of advantages [26, 27]). We show, through a series of analyses, that these combined RNA-seq and PRO-seq measurements are a powerful means for assaying RNA stability that can reveal possible determinants of RNA decay.

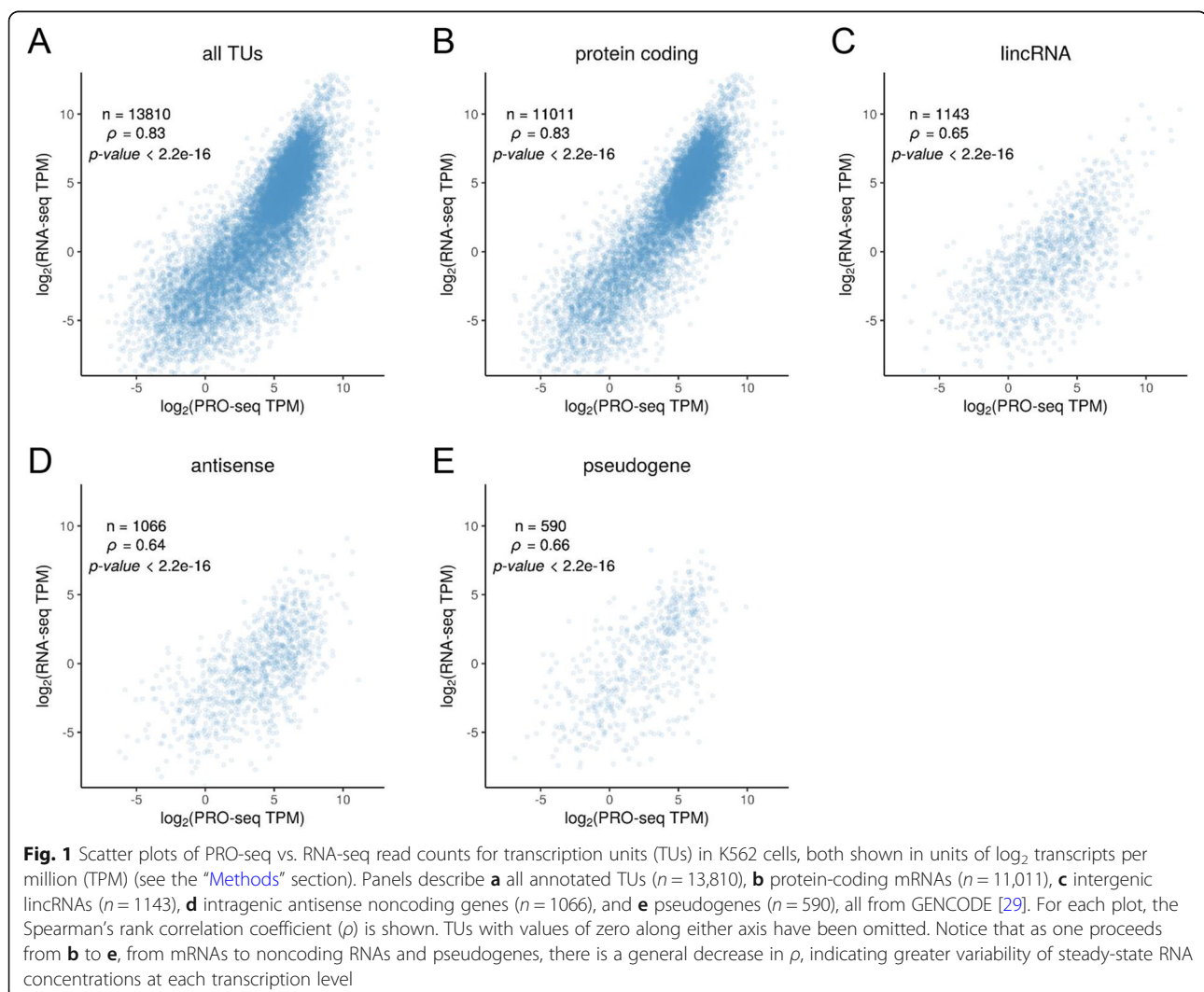
## Results

### Matched PRO-seq and RNA-seq measurements are generally well correlated but suggest reduced stability of noncoding RNAs

We first compared PRO-seq and RNA-seq measurements for various TUs from across the human genome, to assess the degree to which transcriptional activity, as assayed by PRO-seq, is predictive of steady-state RNA concentrations, as assayed by RNA-seq. We obtained previously published PRO-seq [28] and rRNA-depleted poly-A+ RNA-seq data for K562 cells (see the “Methods” section), and pooled the two replicates available for each data type after verifying high concordance between them (Additional file 2: Figure S1). We also collected new PRO-seq and total-RNA RNA-seq data for K562 cells and obtained similar results (see the “Methods” section), but we focus here on the previously published data, which exhibited somewhat reduced technical and biological noise (see the “Discussion” section). When analyzing these data, we considered all annotated

TUs in GENCODE [29], dividing them into mRNA ( $n = 11,011$ ), lincRNA ( $n = 1143$ ), antisense ( $n = 1066$ ), and pseudogene ( $n = 590$ ) classes. We quantified expression by the total number of mapped reads in transcripts per million (TPM), a measure that normalizes by both library size and TU length, and discarded TUs with insufficient read counts from either assay. Notably, we excluded the first 500 bp downstream of the TSS and 500 bp upstream of TES for PRO-seq to avoid a bias from promoter-proximal pausing and polymerase deceleration [23] (see the “Methods” section).

We found that the PRO-seq and RNA-seq measurements were well correlated overall, with Spearman’s  $\rho = 0.83$  (Fig. 1a), suggesting that transcription explains the majority of the variance in mRNA levels. A parallel analysis based on pooled intronic reads from the same RNA-seq libraries showed only a slightly higher correlation, with  $\rho = 0.85$  (Additional file 2: Figure S2). At the same time, there were considerable differences in the



degree of correlation across classes of TUs, ranging from a high of  $\rho = 0.83$  for protein-coding mRNAs to  $\rho = 0.65$  for lincRNAs,  $\rho = 0.64$  for antisense genes, and  $\rho = 0.66$  for pseudogenes (Fig. 1b–e). We observed similar patterns for both intron-containing and intron-less genes (Additional file 2: Figures S3 & S4). Together, these observations suggest that RNA decay has a more pronounced effect on steady-state RNA levels in noncoding RNAs and pseudogenes. These differences remain when TUs are matched by expression level (see the “Methods” section; Additional file 2: Figure S5), when our own K562 data is used (Additional file 2: Figure S6), and when the HeLa cells are evaluated instead (Additional file 2: Figure S7).

Elongation rate is an important potential confounding factor in this analysis, because the PRO-seq density does not directly reflect the synthesis rate of RNA, but rather the synthesis rate divided by the elongation rate, which is known to vary across TUs [30]. However, when we correct for elongation rate using published estimates for K562 cells [31], we find that the correlation with RNA-seq measurements does not improve, and indeed, declines slightly (Additional file 2: Figure S8). Thus, the observed relationships between PRO-seq and RNA-seq measurements do not appear to be driven primarily by differences in elongation rate (see the “Methods” and “Discussion” sections).

#### Relative RNA half-life can be estimated from the RNA-seq/PRO-seq ratio

As noted above, a quantity proportional to RNA half-life can be approximated in a straightforward manner from measurements of transcription rate and steady-state RNA concentration under equilibrium conditions [21, 22, 32]. Briefly, if  $\beta_i$  is the rate of production of new RNAs for each TU  $i$ ,  $\alpha_i$  is the per-RNA-molecule rate of decay, and  $M_i$  is the number of RNA molecules, then, at steady state,  $\beta_i = \alpha_i M_i$ , and the decay rate can be estimated as  $\alpha_i = \beta_i / M_i$  (see Fig. 2a and the “Methods” section). If we assume that  $\beta_i$  is approximately proportional to the normalized PRO-seq read counts for  $i$ , denoted  $P_i$ , and  $M_i$  is proportional to the normalized RNA-seq read counts, denoted  $R_i$ , then the ratio  $P_i / R_i$  is an estimator for a quantity proportional to the decay rate, and its inverse,  $T_{1/2,i}^{PR} = R_i / P_i$ , is an estimator for a quantity proportional to RNA half-life. As noted, the use of PRO-seq, rather than intronic read counts, for the measure of transcription has a number of advantages, including applicability to intron-less TUs and increased sensitivity for TUs expressed at low levels.

Following this approach, we estimated  $T_{1/2}^{PR}$  values for TUs from across the genome using the PRO-seq and RNA-seq data for K562 cells. To validate our estimates, we compared them with estimates of RNA half-life for

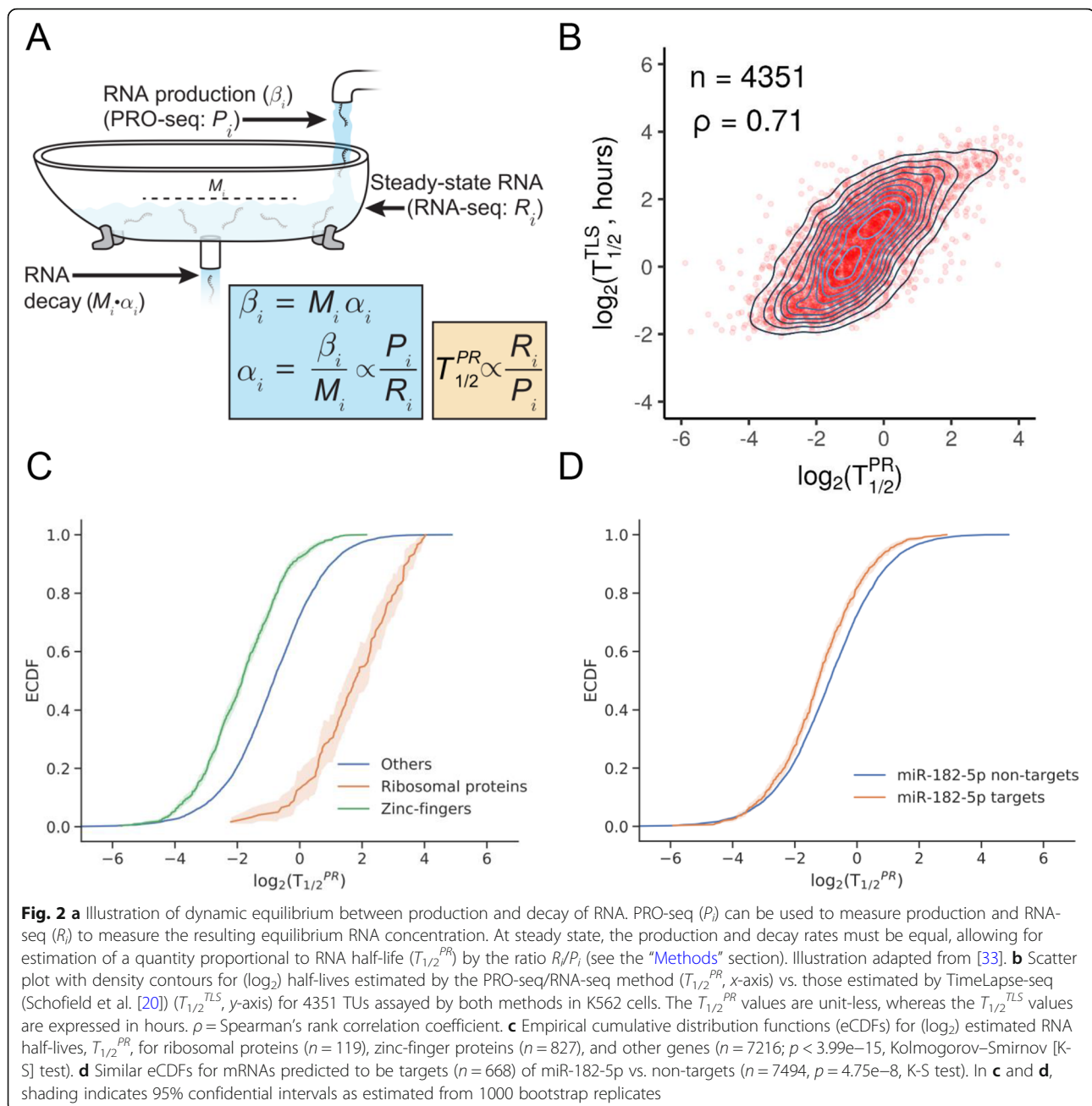
K562 cells from TimeLapse-seq [20], a recently published method based on chemical conversion of 4sU. We compared our estimates of half-life with those from TimeLapse-seq (denoted  $T_{1/2}^{TLS}$ ) at 4351 genes measured by both methods. We found that the two sets of estimates were reasonably well correlated (Spearman’s  $\rho = 0.71$  Fig. 2b), especially considering the substantial differences in experimental protocols and the generally limited concordance of published half-life estimates across experimental methods [6, 15]. By contrast, estimates based on intronic reads showed much poorer agreement with TimeLapse-seq ( $\rho = 0.47$ ; Additional file 2: Figure S9), although it is worth noting that the correction for RNA processing introduced by Alkallas et al. [21] could not be applied in our case, because it requires a comparison of two conditions. We found that our estimated  $T_{1/2}^{PR}$  values were significantly shifted toward lower values for zinc finger proteins (Fig. 2c), many of which play key regulatory roles, and toward higher values for ribosomal proteins, which are representative of “housekeeping” genes. We also found that the predicted targets of numerous miRNAs, including the well-studied miR-182 (Fig. 2d) [34], have significantly reduced stability (see Additional file 2: Figure S10 for additional examples).

As further validation, we extended our comparison to include estimates of RNA half-life for K562 cells based on TT-seq [35], SLAM-seq [36], and the method of Mele et al. [37], focusing on 3449 protein-coding genes for which estimates from all methods are available. In general, all methods show significant but somewhat modest levels of correlation in their half-life estimates, ranging from a high value of Spearman’s  $\rho = 0.80$  for the TimeLapse-seq and Mele et al. [37] methods to a low of  $\rho = 0.51$  for TT-seq and our method (Additional file 2: Figure S11). We attribute these differences in correlation to a variety of both technological and conceptual differences among methods (see the “Discussion” section). Finally, we explicitly adjusted our estimates of relative half-life for elongation rate, and found that the correlation with other methods did not improve (Additional file 2: Figure S12).

#### Properties of transcription units that are predictive of RNA stability

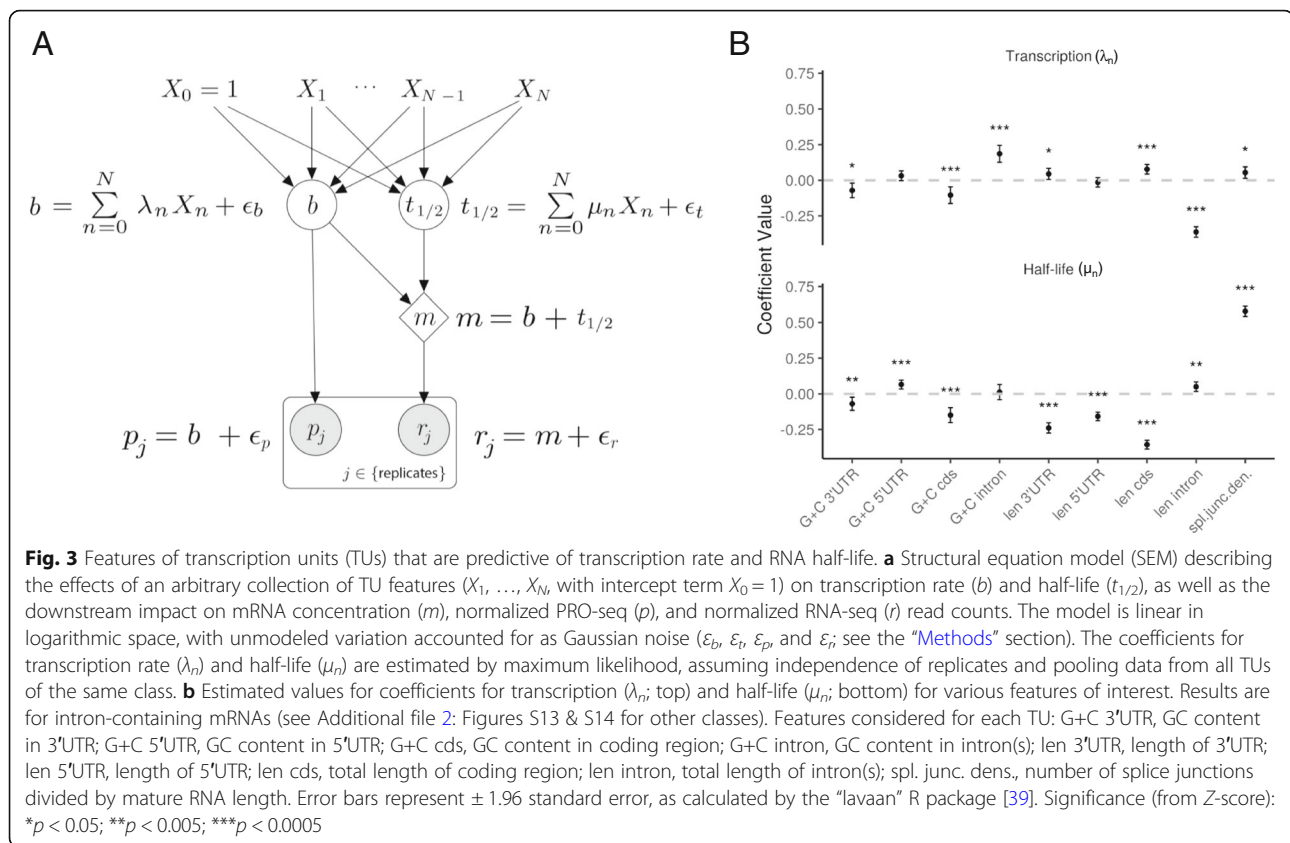
To reveal potential determinants of RNA stability, we sought to identify features of TUs that were predictive of our estimated RNA half-lives. We focused on the mRNA and lincRNA classes, for which we could identify the most informative features. Anticipating an effect from splicing [5, 38], we focused our analysis on intron-containing TUs. We considered nine different features related to splicing patterns, transcript length, and G+C content (Fig. 3 and Additional file 2: Figures S13 & S14).





In previous studies of this kind, investigators have examined the correlation of each feature with half-life, either individually or together in a multiple regression framework. By construction, however,  $T_{1/2}^{PR}$  will tend to be statistically correlated with features predictive of transcription regardless of their true influence on half-life. Therefore, we instead made use of a structural equation model (SEM) [39] that explicitly describes the separate influences of features on transcription and half-life, and the contributions of both to RNA abundance (see the “Methods” section and Fig. 3a).

Our analysis revealed positive correlations with half-life of both splice junction density and total intron length, for intron-containing mRNAs and lincRNAs (Fig. 3b; Additional file 2: Figure S13), although the correlation with splice junction density was not statistically significant in lincRNAs. The observation regarding splice junction density is consistent with previous reports for mRNAs [5, 38, 40, 41] and lincRNAs [42], as well as with the general tendency for intron-containing TUs to be more stable than intron-less TUs (Additional file 2: Figure S15). The correlation with intron length is



intriguing but could be an artifact of increased elongation rates in long introns (see below and the “Discussion” section). We also observed several patterns having to do with G+C content and length that are difficult to interpret owing to the complex correlations of these features with CpGs, transcription, splicing, and RNA half-life. Nevertheless, we found that several features had coefficients of opposite sign for transcription and half-life (e.g., 3’UTR, CDS, and intron length), which could be driven, in part, by stabilizing selection on RNA levels (see the “Discussion” section).

To evaluate the degree to which these findings were influenced by elongation rate, we repeated the SEM analysis for a subset of genes ( $n = 1429$ ) also analyzed by Veloso et al. [31], using an updated estimate of transcription rate that explicitly corrected for the estimated elongation rates of these genes (see the “Methods” section). We found that most of the results above held up under this analysis, with the main exception being the positive correlation between intron length and RNA half-life (Additional file 2: Figure S16). This finding could be an artifact of elongation rate in our uncorrected analysis because there is evidence of increased elongation rate (which would be perceived as reduced

PRO-seq signal, and hence increased RNA-seq/PRO-seq ratio) in long introns [43]. We also observed some differences in the associations with G+C content.

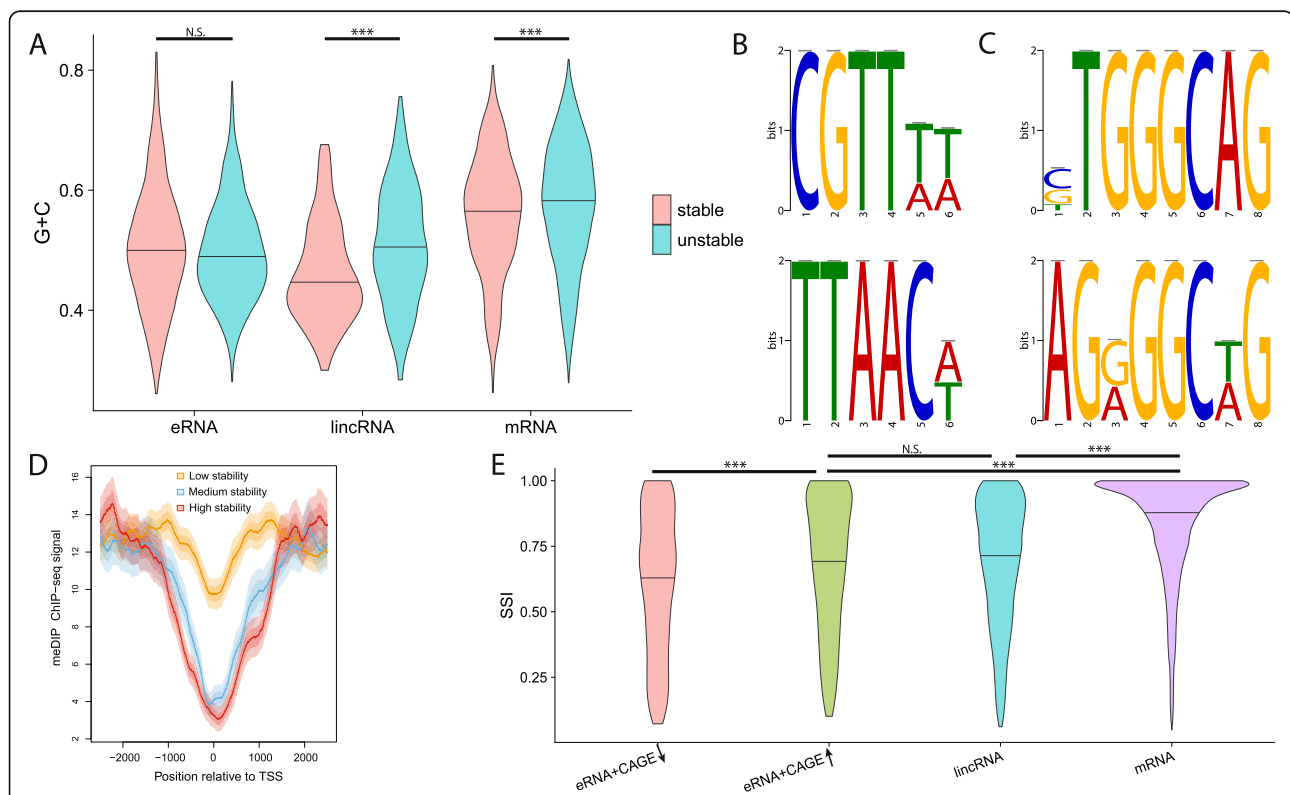
As further validation, we performed a similar analysis using estimates of half-life based on TT-seq [35], SLAM-seq [36], and the study of Mele et al. [37], focusing on 3418 genes for which features and estimates were available from all methods (Additional file 2: Figure S17). In these cases, we did not have separate measures of transcription and steady-state RNA abundance, so in place of the SEM analysis, we performed multiple linear regression (MLR) using the same features as covariates and the estimated half-lives from each of these other studies as outcomes. For comparison, we repeated the analysis of our half-life estimates also by MLR. In general, we found that the observed trends were similar across all methods. The major exception was 3’UTR length, where the other methods found a positive correlation instead of the negative correlation observed with our method. It is possible that this difference might also reflect a bias in our estimates from elongation rate, which has been observed to decrease near the 3’ ends of genes [8]. However, other studies have also noted a negative correlation between 3’UTR length and half-life, possibly related to the presence of miRNA or RBP binding sites [4, 5, 44].

### DNA sequence correlates of RNA stability

Our estimates of RNA half-life for both coding and non-coding TUs provide an opportunity to better characterize DNA sequence correlates of RNA stability near transcription start sites (TSSs) [5, 28, 45, 46]. We tested for associations between half-life and DNA words ( $k$ -mers) of various lengths near the TSS (see the “Methods” section), but we found that the observed trends were predominantly driven by G+C content, with A+T-rich  $k$ -mers being enriched, and G+C-rich  $k$ -mers being depleted, in stable transcripts relative to unstable transcripts (Fig. 4a; Additional file 2: Figures S18–S20). Using the discriminative motif finder DREME [47], we identified several A+T-rich motifs associated with stable transcripts, and several G+C-rich motifs associated with unstable transcripts (Fig. 4b, c). Finally, we expanded our set of TUs to include previously identified eRNAs from K562 cells [28] (see the “Methods” section), and

found, interestingly, that stable eRNAs were slightly enriched, rather than depleted, for G+C-rich sequences close to the TSS (Fig. 4a; Additional file 2: Figure S20). This trend was most strongly associated with CpG dinucleotides within 400 bp of the TSS (Additional file 2: Figure S21).

The atypical patterns around CpG dinucleotides raise the possibility of an association with DNA methylation near the TSS. We therefore compared the methylation patterns of TUs exhibiting low, medium, or high levels of RNA stability, summarizing these patterns with meta-plots of average signal of the methylated DNA immunoprecipitation (MeDIP-seq) assay in K562 cells [48, 49] as a function of distance from the TSS (see the “Methods” section). We found that the medium- and high-stability TUs exhibited similar patterns of methylation, but the low-stability TUs show a clear enrichment (Fig. 4d). A similar trend was evident for lincRNAs (Additional file



**Fig. 4** DNA-sequence, methylation, and RNA-binding-protein correlates of RNA stability near the TSS. **a** Distribution of G+C content (y-axis) for the 20% most (red) and least (blue) stable TUs, according to our estimated half-life ( $T_{1/2}^{PR}$ ), in enhancer RNAs (eRNA, stable:  $n = 510$ ; unstable:  $n = 510$ ), lincRNAs (stable:  $n = 91$ ; unstable:  $n = 198$ ), and mRNAs (stable:  $n = 919$ ; unstable:  $n = 2146$ ). **b**, **c** Two of the most significantly enriched DNA sequence motifs in stable (**b**) and unstable (**c**) mRNAs. **d** Signal for MeDIP-measured DNA methylation for low-, medium-, and high-stability mRNAs (see the “Methods” section) as a function of distance from the TSS. Solid line represents mean signal and lighter shading represents standard error and 95% confidence interval. **e** Distribution of sequence stability index (SSI) based on U1 and polyadenylation sites (see the “Methods” section) for eRNAs ( $n = 1020$ ), lincRNAs ( $n = 989$ ), and mRNAs ( $n = 10,728$ ). Separate plots are shown for eRNAs with low ( $n = 510$ ) and high ( $n = 510$ ) CAGE support, suggesting low and high stability, respectively. Significance of comparisons in **a** and **e** (from Mann–Whitney  $U$  test): \* $p < 0.01$ ; \*\* $p < 0.001$ ; \*\*\* $p < 0.0001$ ; N.S., not significant

2: Figure S22). These observations suggest the possibility of epigenomic as well as DNA sequence differences associated with RNA stability, as we explore further below.

### U1 and polyadenylation sites have limited predictive power for stability

We also directly tested for the possibility that differences in RNA half-life could reflect the presence or absence of either U1 binding sites (5' splice sites) or polyadenylation sites (PAS) downstream of the TSS. Comparisons of (stable) protein-coding TUs and (unstable) upstream antisense RNA (uaRNA) TUs have revealed significant enrichments for proximal PAS in uaRNAs, suggesting that they may lead to early termination that triggers RNA decay. These studies have also found significant enrichments for U1 binding sites in protein-coding TUs, suggesting that splicing may play a role in enhancing RNA stability [45, 46]. In previous work, we showed that these trends generalize to eRNAs as well. In particular, we found that a hidden Markov model (HMM) that distinguished between the occurrence of a PAS prior to a U1 site and the occurrence of a U1 site prior to a PAS could classify TUs as unstable or stable, respectively, with fairly high accuracy [28].

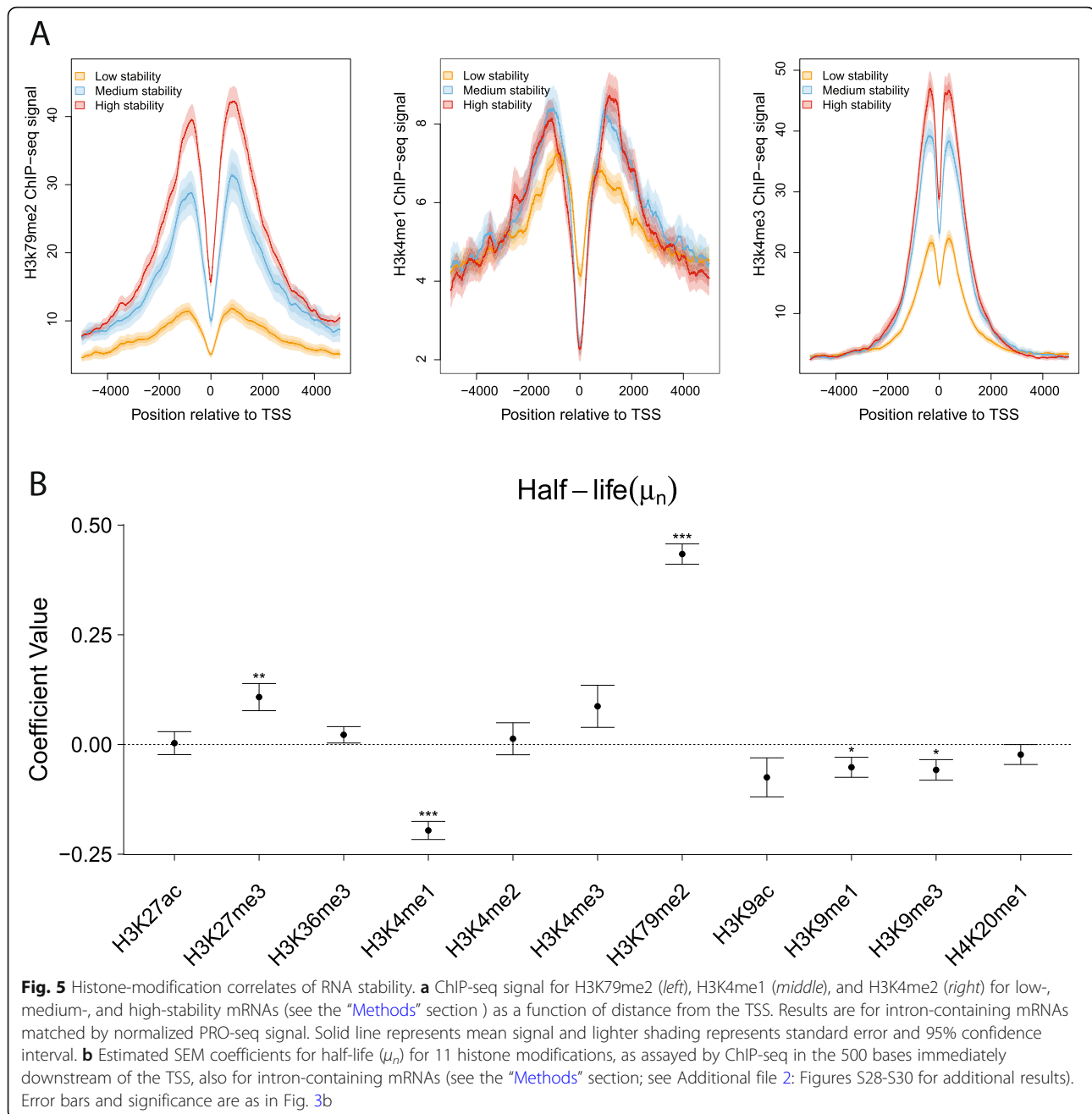
We applied this HMM (see the “Methods” section) to our mRNA and lincRNA TUs and tested whether our DNA sequence-based predictions of stability (as measured by a sequence stability index, or SSI) were predictive of our estimated  $T_{1/2}^{PR}$  values. We also computed the SSI for the eRNAs identified from PRO-seq data and classified as stable or unstable based on CAGE data. We found that the mRNAs had the highest SSI, followed by lincRNAs, and then eRNAs (Fig. 4e), as expected. Interestingly, however, the subset of eRNAs that we find to be stable based on CAGE data also show elevated SSIs, roughly on par with lincRNAs. In addition, intron-containing lincRNAs have significantly higher SSIs than intron-less lincRNAs, although there was little difference in intron-containing and intron-less mRNAs (Additional file 2: Figure S23). Moreover, within each of the mRNA and lincRNA groups, we found that the SSI changed relatively little as a function of  $T_{1/2}^{PR}$ , suggesting that the HMM had almost no predictive power for true RNA stability within these classes (Additional file 2: Figures S24 & S25). These observations suggest that, whereas the U1 and PAS sequence signals do seem to distinguish broad classes of TUs with different levels of stability—namely mRNAs, eRNAs, and uaRNAs—and the same signals are useful in distinguishing stable and unstable eRNAs, other factors likely dominate in determining gradations of stability within the mRNA and lincRNA classes (see the “Discussion” section).

### Additional epigenomic correlates of RNA stability

Finally, we asked whether other epigenomic marks such as histone modifications correlate with RNA stability. Histone modifications are primarily associated with transcriptional activity or repression, but there is increasing evidence that events occurring before or during transcription can be associated with post-transcriptional processes and RNA stability [50–54] (see the “Discussion” section). Similar to the methylation analysis above (Fig. 4d), we produced meta-plots showing the average ChIP-seq signal in K562 cells as a function of distance from the TSS for 11 different common histone modifications [48], separately for low-, medium-, and high-stability classes of expression-matched intron-containing mRNAs (see the “Methods” section). While some of these histone modifications did not differ substantially across stability classes, such as H3K9me1 and H3K9me3, several did show clear relationships with estimated RNA half-life (Additional file 2: Figures S26 & S27). For example, H3K79me2, which is associated with transcriptional activity, gives a substantially higher signal in stable transcripts than in unstable ones, particularly in peaks about 1 kb from the TSS (Fig. 5a). A similar pattern is observed for H3K4me2, H3K4me3, H3K9ac, and H3K27ac. The H3K4me1 mark, which is associated with active enhancers, displays a similar pattern far from the TSS but an inverse pattern close to the TSS (Fig. 5a and Additional file 2: Figure S26).

As an alternative strategy for identifying epigenomic correlates of RNA stability while correcting for transcription, we again applied our SEM framework, this time using the 11 histone marks as covariates for estimated RNA half-life and considering the ChIP-seq signals immediately downstream of each TSS (Fig. 5b and Additional file 2: Figure S28). As expected, the strongest correlations were detected with transcription rate, and these generally had the expected sign, for example, with positive correlations for the activation marks H3K27ac, H3K4me1, and H3K4me3, and negative correlations for the repressive marks H3K9me3 and H3K27me3. These patterns were generally consistent between lincRNAs and mRNAs (Additional file 2: Figures S28 & S29), and they did not change substantially as a function of distance from the TSS (Additional file 2: Figures S30 & S31). However, we did additionally identify several significant correlates of half-life. For mRNAs, these were generally consistent with the ones identified from the ChIP-seq meta-plots, for example, with H3K79me2 showing a positive correlation with RNA half-life, and H3K4me1 showing a negative correlation close to the TSS. In general, the estimated coefficients were similar for mRNAs and lincRNAs, but there were some notable differences: for example, the activity mark H3K36me3 shows a negative correlation with RNA half-life in lincRNAs but a weaker and position-dependent positive correlation with mRNA half-life, and the silencing mark H3K9me3 shows a





position-dependent positive correlation for lincRNA half-life but a negative correlation for mRNA half-life (Additional file 2: Figures S28 & S29). These divergent patterns could possibly reflect differences in the degree to which splicing is co-transcriptional in mRNAs and lincRNAs [55].

## Discussion

In this article, we have introduced a simple method for estimating the RNA half-lives of TUs from across the genome based on matched RNA-seq and PRO-seq data sets. Like previous methods based on intronic reads, our

method assumes equilibrium conditions and produces a relative measure only of half-life. Unlike these methods, however, the use of PRO-seq allows us to interrogate intron-less TUs and TUs that are expressed at low levels. Moreover, even for intron-containing and abundantly expressed genes, the PRO-seq-based measurements appear to be considerably more accurate than those based on intronic reads. Our approach also has a number of advantages in comparison to existing methods for estimating RNA half-lives based on transcriptional inhibition or metabolic labeling. For example, it does not

require collecting data in a time course, which enables efficient use of both time and sample material; it can make use of RNA-seq or PRO-seq data generated for other purposes; it is relatively nondisruptive of the biological processes under study; and it can be extended to tissue samples using ChRO-seq [26] (see Additional file 1: Table S1). We have shown that our measurements of relative half-life are useful in a wide variety of downstream analyses.

Our original design for this study was to generate our own PRO-seq and RNA-seq data from the same source of K562 cells, to ensure the data sets were as closely matched as possible. In addition, we produced total rRNA-depleted RNA-seq libraries, rather than poly-A+ libraries, with the goal of improving our sensitivity for noncoding RNAs. For validation, we compared our results with ones based on previously published PRO-seq data [28] and poly-A+ RNA-seq data from the ENCODE project [48]. As it happened, however, we found that the half-life estimates based on these previously published data sets were less variable, showed better agreement with published estimates, and did not differ substantially in sensitivity from the ones based on our newly collected data. Therefore, we have focused on these estimates with our main analyses. In future work, we hope to more systematically compare the attributes of total RNA and poly-A+ libraries [56, 57]. It may also be informative to compare measurements based on RNA extracted from particular cellular compartments, such as the nucleus or the cytoplasm. In general, it may be possible to begin to disentangle the contributions of distinct RNA decay pathways (e.g., 3'→5' decay, decapping and 5'→3' decay, nonsense-mediated decay), and their differential effects on distinct classes of RNAs, through such comparisons [58]. In addition, it may be worthwhile to examine how RNA stability varies across conditions or cell types, as most studies so far have only measured RNA stability for a particular cell type under a particular set of conditions.

In a comparison of half-life estimates from several methods that have all been applied to K562 cells, including TimeLapse-seq [20], TT-seq [35], SLAM-seq [36], and the method of Mele et al. [37], we found reasonable agreement across methods, but also some differences (Additional file 2: Figure S11). The average pairwise Spearman's correlation coefficient between sets of estimates was relatively modest at  $\rho = 0.64$ . It is difficult at this stage to disentangle the sources of the differences among methods. Most likely, they result both from experimental noise and from a combination of more fundamental differences, including whether the estimates are based on steady-state assumptions or time-course measurements, whether transcriptional inhibition or activation is used, how the rate of transcription is assayed, and whether RNA abundance is based on total RNA or

polyA+ RNA. These differences may make some methods better for certain classes of TUs than others (e.g., coding vs. noncoding RNAs, lowly vs. highly expressed TUs, intron-containing vs. intron-less TUs, or RNAs that are or are not at equilibrium). More work will be required to clarify the relative strengths and weaknesses of the available methods.

Notably, our method has limited sensitivity for highly unstable transcripts. When half-lives are low, the RNA-seq signal tends toward zero, leading to limited ability to identify gradations of stability. For this reason, we have focused our half-life analysis on genes with fairly strong signals from both assays (PRO-seq > 10 TPM and RNA-seq > 1 TPM; see the "Methods" section). At the same time, similar limitations occur with essentially all of the available assays for half-life estimation, and our approach at least has the advantage that PRO-seq is highly sensitive as a measure of transcriptional activity.

Perhaps a more important limitation of our method is that, strictly speaking, PRO-seq is a measure not of the rate of transcription but of the occupancy of engaged RNA polymerases, which reflects both the rate of transcription and the rate of elongation. The PRO-seq signal along a gene body is analogous to the headlight brightness on a highway at night; an increase in signal can reflect either an increased number of cars entering the highway (analogous to an increased rate of transcription), or a back-up in traffic (analogous to a decreased elongation rate). Consequently, variation in  $T_{1/2}^{PR}$  across TUs could in part be driven by variation in elongation rate. We attempted to control for this possibility in several ways. First, we explicitly corrected our estimates of transcription and half-life with previously published estimates of elongation rate for the same cell type [31] (see the "Methods" section). We found that the correction did not improve the correlation of PRO-seq and RNA-seq measurements (Additional file 2: Figure S8), nor did it improve the agreement with independent estimates of half-life (Additional file 2: Figure S12). Second, we repeated our analysis of features predictive of half-life with the corrected estimates and found that it did not substantially alter our results, with one notable exception (Additional file 2: Figure S16; discussed below). Third, we observed that the variation in elongation rate across genes is only about one fifth of the variation in estimated half-lives, indicating that it can account for, at most, a fairly small fraction of the observed variation (Additional file 3: Supplemental Text). We conclude from these analyses that elongation rate does undoubtedly have some impact on our half-life estimates, but overall, the effects appear to be limited. However, more work will be needed to obtain more accurate and more comprehensive estimates of elongation rates, and to fully understand their impact on half-life estimates.

To identify features that are predictive of RNA half-life, we devised a structural equation model (SEM) that explicitly describes the separate effects of each feature on transcription and half-life, as well as the resulting impact on RNA concentrations, PRO-seq, and RNA-seq data. While multivariate regression has been used to identify features associated with RNA stability [5], our analysis is the first, to our knowledge, to attempt to disentangle the separate influences of such features on transcription and RNA stability. It is worth noting that this framework could also be useful for estimators based on intronic reads. The results of the SEM analysis were consistent with previous findings in many respects, particularly regarding the association between RNA splicing and RNA stability. The mechanism underlying this relationship remains unclear, but it is known that the exon junction complex (EJC) remains bound to the mature mRNA after its transport to the cytoplasm and it has been proposed that EJC components may protect the mRNA from decay [5, 41]. In addition to the previously reported positive correlation of splice junction density and RNA half-life, we also observed a positive correlation between intron length and half-life. This observation could potentially indicate that RNA stability is enhanced by recursive splice sites [59] or extended contact with the spliceosome in long introns. However, we could not confirm this finding after our correction for elongation rate using a subset of our full gene set, and it may therefore be an artifact of increased elongation rates in long introns. More work will be needed to confirm or reject this association.

It has recently been reported that U1 binding sites are enriched, and polyadenylation sites are depleted, downstream of the TSS in stable mRNAs relative to unstable upstream antisense RNAs (uaRNAs) and enhancer RNAs (eRNAs), suggesting that RNA stability is determined, in part, by the DNA sequence near the TSS. In this study, we tested not only whether this “U1-PAS axis” could distinguish TUs in stable classes (mRNAs) from those in unstable classes (uaRNAs and eRNAs) but also how predictive it is of half-life within these classes. We confirmed that a U1-PAS-based “sequence stability index” (SSI) is generally elevated for mRNAs, intermediate for lincRNAs, and reduced for eRNAs. Furthermore, this SSI can distinguish between more and less stable eRNAs, as quantified using CAGE (Fig. 4e). Somewhat surprisingly, however, we found that the SSI has essentially no predictive power for relative RNA stability within the generally more stable mRNA and lincRNA classes (Additional file 2: Figures S24 & S25). One possible explanation for this observation is that the U1-PAS axis determines a kind of early “checkpoint” for stable transcripts—for example, by ensuring that premature transcriptional termination is avoided—but that once a

transcript has cleared this checkpoint, these DNA sequence features are no longer relevant in determining RNA stability. Instead, the relative stability of mRNAs and lincRNAs may be predominantly determined by splicing-related processes, binding by miRNAs or RBPs, or other post-transcriptional phenomena. More work will be needed to fully understand the mechanistic basis of these differences in stability.

Some of the associations that we observed with half-life concerned G+C content, but these observations are generally difficult to interpret. Indeed, even the comparatively straightforward question of the relationship between G+C content and transcriptional activity has a long and contradictory literature, with several studies finding correlations between them [60–62], but others claiming that the relationship between G+C and transcription is weak, at best, once confounding factors such as genomic context are properly accounted for [63, 64]. Sharova et al. [5] identified a fairly pronounced negative correlation between RNA stability and the prevalence of CpGs in the 5'UTR, which is not supported by our analysis—although we interrogated only G+C content, not CpGs, in the 5'UTR. These authors raised the intriguing hypothesis this correlation may reflect the activity of splicing-associated methyl CpG-binding proteins [65], but to our knowledge, this idea has not been tested experimentally. In any case, it seems unlikely that the complex relationships among G+C content, CpGs, transcription, RNA stability, and downstream effects such as translational efficiency can be fully disentangled through post hoc statistical analyses. Instead, this effort will require experiments that directly perturb individual features of interest and separately measure the effects on a variety of processes.

There is now substantial evidence for connections between events that occur before or during transcription and a variety of post-transcriptional processes, some of which impact RNA stability. In addition to the apparent enhancement of RNA stability by splicing, there is now evidence that some RNA-binding proteins having roles in RNA export and stability are recruited to the RNA in a promoter-dependent manner [66–69]. Similarly, co-transcriptional processes such as polyadenylation and capping appear are linked to RNA stability [51]. It was also recently shown that disrupting transcription rates could lead to enhanced m6A deposition, shortened polyA tails, and reduced RNA stability [52, 53]. With these observations in mind, we looked for epigenomic correlates of stability. We identified several histone modifications that are significantly associated with increased or decreased half-life, but we cannot rule out the possibility that these correlations reflect indirect relationships with confounding variables not considered here. However, the effect is quite strong for certain marks (such as

H3K79me2 and H3K4me2) and it is apparent both in direct comparisons of PRO-seq-matched TUs (Fig. 5a) and in the SEM setting (Fig. 5b). It therefore seems plausible that it has a direct mechanistic basis, perhaps involving factors that interact both with DNA-bound nucleosomes and the spliceosome. Some divergent patterns for mRNAs and lincRNAs (Additional file 2: Figure S28) suggest the possibility of differences in these splicing-associated processes. Additional work will be needed to test these hypotheses.

One general pattern that emerges from the SEM analysis of histone modifications is that the coefficients for transcription and half-life are often different from zero in opposite directions (Additional file 2: Figures S28–S31). This trend of anti-correlation was less prominent with the TU features, but we did observe it with CDS, intron, and 3'UTR length (Fig. 3b). A possible explanation for this pattern is that it is, at least in part, a reflection of stabilizing selection on gene expression. If selection tends to favor a particular RNA level for each TU, then mutations that increase transcription may tend to be compensated for by mutations that decrease RNA stability, and vice versa. Thus, stabilizing selection might result in a tendency for features that are positively correlated with one measure (transcription or stability) to be negatively correlated with the other. Notably, this type of hypothetical causal interrelationship between transcription and stability is not considered in our SEM, nor in any other statistical model of which we are aware. As a result, it may be difficult to distinguish correlations that have a direct, mechanistic basis (say, relating to transcription) from their indirect “echoes” (say, relating to half-life) resulting from evolutionary constraint. Despite this potential limitation, our framework remains useful for identifying potentially interesting correlations, whose mechanistic underpinnings can then be further investigated through direct experimental perturbation.

## Conclusions

We introduce a novel approach for estimating the relative half-lives of individual RNAs using PRO-seq and RNA-seq. We develop a structural equation model and test multiple features for their associations with RNA stability after controlling for the effects on transcription. Together, our estimation method and systematic analysis shed light on the pervasive impacts of RNA stability on cellular RNA concentrations.

## Methods

### PRO-seq and RNA-seq data preparation and processing

Our main analysis is based on PRO-seq data for K562 [28] and HeLa [70] cell lines as well as RNA-seq data from the ENCODE project [48, 71] (ENCSR000AEM for K562, ENCSR000CPR for HeLa). For comparison, we

also sequenced new PRO-seq ( $n = 2$ ) and RNA-seq ( $n = 4$ ) libraries, generated from cells grown in the same flask under the same conditions. Human K562 cells were cultured using standard cell culture procedures and sterile techniques. The cells were cultured in RPMI-1640 media supplemented with 10% fetal bovine serum (FBS) and 1% penicillin/streptomycin. For PRO-seq, 3' and 5' adapters were ligated as described [26] followed by library preparation as previously published [72]. Sequencing was done by Novogene on a HiSeq instrument with paired-end reads of  $2 \times 150$  bp. For RNA-seq, total RNA was extracted using the Trizol method (see [https://assets.thermofisher.com/TFS-Assets/LSG/manuals/trizol\\_reagent.pdf](https://assets.thermofisher.com/TFS-Assets/LSG/manuals/trizol_reagent.pdf)), followed by rRNA depletion using the Ribozero HMR Gold kit. Libraries were prepared using the NEB kit with TruSeq RNAseq adaptors. Single-end sequencing (length = 75) was performed on a NextSeq500 instrument by the RNA Sequencing Core at the College of Veterinary Medicine, Cornell University. Sequencing data is deposited on GEO under accession GSE153200.

### Read mapping and transcript abundance estimation

Raw data files in fastq format were trimmed using Cutadapt [73] with parameters `-j 0 -e 0.10 --minimum-length=10`. Reads were then aligned using HISAT2 [74] with default parameters (`hisat2 --threads 4 -x {index} -U {input.reads} -S {output} --summary-file {log}`). We used the GRCh38/hg38 reference genome and the associated GENCODE gene annotations. HTSeq [75] was used for read counting for RNA-seq and PRO-seq. For the purposes of read counting with PRO-seq, we omitted the first 500 bases downstream of the TSS and 500 bases upstream of TES to avoid a bias in read counts from promoter-proximal pausing and polymerase deceleration. Finally, we normalized read counts by converting them to transcripts per million (TPM) [76] based on the length of each TU.

### Estimation of RNA half-life from RNA-seq and PRO-seq data

We assume a constant rate of production of new RNAs,  $\beta_i$ ; a constant per-RNA-molecular rate of decay,  $\alpha_i$ ; and a number of RNA molecules,  $M_i$ . At steady state,  $\beta_i = \alpha_i M_i$ ; therefore, the decay rate can be estimated as  $\alpha_i = \beta_i / M_i$ , and the half-life as  $T_{1/2} = \ln(2) / \alpha_i = \ln(2) \times M_i / \beta_i$ . We further assume that the normalized PRO-seq read counts (omitting regions near the TSS and TES) are proportional to the rate of production of new RNAs,  $P_i \propto \beta_i$ , and that the normalized RNA-seq read counts are proportional to the number of RNA molecules,  $R_i \propto M_i$ . Therefore,  $T_{1/2} \propto R_i / P_i$ . We define our unit-less estimator of half-life as  $T_{1/2}^{PR} = R_i / P_i$ , where  $PR$  denotes a PRO-seq/RNA-seq-based estimator. Notice that these unit-less  $T_{1/2}^{PR}$  values



can be compared across experiments only up to a proportionality constant, unless the raw read counts have been appropriately normalized. To compare our PRO-seq-based approach with an approach based on intronic reads, we repeated the estimation using normalized intron reads instead of PRO-seq read counts to represent the transcription rate.

### Correction for elongation rate in PRO-seq vs. RNA-seq correlation and half-life estimations

A potential confounding factor in the comparison of normalized read counts for PRO-seq and RNA-seq is elongation rate. Because PRO-seq read depth reflects a combination of transcription initiation rates and elongation rates [30, 77], some reduction in correlation with RNA-seq could reflect variability across TUs in elongation rate. We investigated this possibility by using published measurements of elongation rate for the same K562 cell type [31], focusing on ~2000 genes that overlap our set. We explicitly adjusted for the estimated elongation rates by multiplying them by the PRO-seq abundance across gene bodies, under the assumption that the PRO-seq density is proportional to the transcription rate divided by the elongation rate. The corrected PRO-seq abundance was then used for comparison with RNA-seq, for half-life estimation, and for the SEM analysis.

### Structural equation model

To separate the effects of TU features on decay from the effects on transcription, we developed a SEM using the “lavaan” R package [39]. Let  $X_n$  be the  $n$ th feature associated with a TU. We assume that the logarithms of this TU's transcription rate and half-life, i.e.,  $b = \log \beta$  and  $t_{1/2} = \log T_{1/2}^{PR}$ , are linear combinations of the  $X_n$ 's and a TU-level random effect:  $b = \sum_{n=0}^N \lambda_n X_n + \epsilon_b$  and  $t_{1/2} = \sum_{n=0}^N \mu_n X_n + \epsilon_t$  where  $\epsilon_b \sim N(0, \sigma_b)$  and  $\epsilon_t \sim N(0, \sigma_t)$  are independent Gaussian random variables explaining all variation not attributable to known features. Assuming a fixed value  $X_0 = 1$  for all genes, the parameters  $\lambda_0$  and  $\mu_0$  can be interpreted as intercepts whereas  $\lambda_{n \neq 0}$  and  $\mu_{n \neq 0}$  are regression coefficients indicating the contributions of feature  $n$  to transcription rate and half-life, respectively.

According to the model derived above, at steady state,  $T_{1/2}^{PR} \propto M/\beta$ , where  $M$  is the number of RNA molecules; therefore,  $m = \log M$  is given by  $m = b + t_{1/2} + C$ , where  $C$  is an arbitrary constant that can be ignored here because it does not affect the estimation of regression coefficients. Denoting  $p_j = \log P_j$  and  $r_j = \log R_j$  as the logarithms of the PRO-seq and RNA-seq measurements in replicate  $j$ , respectively,

we assume  $p_j \sim b + \epsilon_p$  and  $r_j \sim m + \epsilon_r$  where  $\epsilon_p \sim N(0, \sigma_p)$  and  $\epsilon_r \sim N(0, \sigma_r)$  are independent Gaussian random variables describing the noise in PRO-seq and RNA-seq experiments, respectively. Finally, we assume that all observations are independent across TUs. With these assumptions, and pooling information across TUs of the same class, we can estimate separate regression coefficients for transcription rates ( $\lambda_n$ ) and half-life ( $\mu_n$ ) for all features by maximum likelihood.

### Transcription unit features

Transcription unit (TU) sequences were downloaded from BioMart using the R package biomaRt [78, 79]. We considered only one isoform per annotated gene, i.e., the most abundant transcript determined by Salmon [80]. Features based on properties of DNA sequences (e.g., G+C content) were then extracted using Biopython [81]. The intron length was set equal to the transcript length minus the total exon length. The splice junction density was set equal to the intron number divided by the mature RNA length.

### eRNA analysis

We used eRNAs identified from our previous GRO-cap analysis in K562 cells [28] restricting our analysis to putative eRNAs with divergent transcription [27] that fell at least 1 kb away from annotated genes ( $n = 21,816$ ). To measure steady-state RNA levels, we used CAGE in place of RNA-seq owing to its greater sensitivity. We used the Nucleus PolyA and Non-polyA CAGE libraries from ENCODE (GEO accession number GSE344448). To measure transcription rates, we used PRO-seq data from same study [28]. For the stability analysis, we eliminated TUs having no mapped CAGE reads, and then selected the top 10% by CAGE/PRO-seq ratio as “stable” and the bottom 10% as “unstable.” These stable and unstable groups were then matched by PRO-seq signal ( $n = 510$ ).

### DNA word enrichments

We considered all DNA words (all possible combinations of A, C, G, T) of sizes  $k \in \{2, 3, 4\}$ . For each word  $w$ , we counted the total number of appearances in our set of stable TUs (top 20% by  $T_{1/2}^{PR}$ ), denoted  $c_{s,w}$ , and the total number of appearances in unstable TUs (bottom 20% by  $T_{1/2}^{PR}$ ), denoted  $c_{u,w}$ . These counts were collected in 1 kb windows beginning at various distances downstream of the TSS (0, 500, 1000, and 1500 bp). The enrichment score for each word  $w$  and each window position was then computed as  $\log_2(c_{s,w}/c_{u,w})$ . A positive value of this score indicates an enrichment, and a negative score indicates a depletion in stable TUs relative to unstable TUs. For eRNAs, we used a similar procedure but with 400 bp windows at distances of 0, 200, 400, and 600 bp from the TSS.

### Motif discovery

For motif discovery, we used the discriminative motif finder “DREME” [47] with default parameters (core width ranging from 3 to 7). For the stable motifs, we used the top 20% of TUs by  $T_{1/2}^{PR}$  as the primary sequences and the bottom 20% as the control sequences. For the unstable motifs, we reversed the primary and control sequences.

### Sequence stability index

We define the SSI to be the probability that a TU is “stable” based on our previously published U1-PAS hidden Markov model (HMM) [28]. Briefly, the HMM identifies a TU sequence as “stable” if either (1) it has a U1 splicing motif upstream of a PAS motif or (2) it lacks both a PAS motif and a U1 splicing motif, as detailed by Core et al. [28]. We applied the HMM to the first 1 kb of sequence downstream of the annotated TSS and calculated the SSI as 1 minus the probability the TU is unstable, as output by the program. An implementation of the HMM is available at <https://github.com/Danko-Lab/stabilityHMM>.

### Matching by PRO-seq expression

We used the R package “MatchIt” [82, 83] to match groups of TUs by their normalized PRO-seq read counts (method = “nearset”). In cases of multiple groups, one group was selected as the reference and every other group was matched to that reference group.

### MicroRNA targets analysis

We obtained microRNA targets from TargetScanHuman [84], Release 7.2 ([http://www.targetscan.org/vert\\_72/vert\\_72\\_data\\_download/Predicted\\_Targets\\_Info.default\\_predictions.txt.zip](http://www.targetscan.org/vert_72/vert_72_data_download/Predicted_Targets_Info.default_predictions.txt.zip)). We used all default predictions of conserved targets for each conserved miRNA family in the database.

### Gene categories

We obtained lists of genes encoding ribosomal proteins and zinc fingers from the HUGO Gene Nomenclature Committee (<https://www.genenames.org/>).

### Epigenomic analysis

Histone modifications (ChIP-seq; <http://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhHistone/>) and DNA methylation IP (MeDIP; GEO accession number GSM1368906) data were downloaded from the ENCODE consortium [48] as bigwig files annotated to the GRCh37/hg19 reference genome. We partitioned our mRNAs, considering intron-containing TUs only, into five equally sized stability classes based on the estimated  $T_{1/2}^{PR}$  values, and then subsampled from classes 1 (low stability), 3

(medium stability), and 5 (high stability) to obtain distributions matched by PRO-seq signal. We then produced meta-plots for each of these three classes showing the average signal of the histone modifications (ChIP-seq) and methylated DNA immunoprecipitation (MeDIP-seq) assays in K562 cells [48, 49] as a function of distance from the TSS. Meta-plots showing the average values of signals of interest across loci (e.g., Figs. 4d and 5a) were produced using the “plotMeta” function from the “Genomation” [85] R package. The input signal was provided in bigwig format, and the loci were defined in bed format. In all cases, the average signal is plotted as a colored line, with uncertainty indicated by the standard error of the mean (darker shading) and 95% confidence intervals (lighter shading) as specified by the “se” parameter.

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12915-021-00949-x>.

**Additional file 1: Table S1.** Summary of Advantages of Method.

**Additional file 2: Figure S1.** Correlation of PRO-seq and RNA-seq replicates in K562 cells. **Figure S2.** Correlation of exonic reads from RNA-seq with either PRO-seq or intronic reads. **Figures S3-S4.** Scatter plots of PRO-seq vs. RNA-seq for intron-containing (Figure S3) or intron-less (Figure S4) transcription units in K562 cells. **Figure S5.** Scatter plots of PRO-seq vs. RNA-seq for lincRNAs and protein-coding mRNAs in K562 cells after matching by PRO-seq signal. **Figures S6-S7.** Scatter plots of PRO-seq vs. RNA-seq for transcription units in K562 cells (Figure S6) or HeLa cells (Figure S7). **Figure S8.** Scatter plots of PRO-seq vs. RNA-seq for mRNAs and lincRNAs in K562 cells before and after correcting for elongation rate. **Figure S9.** Intronic half-life vs. TimeLapse-seq half-life. **Figure S10.** Empirical cumulative distribution functions (eCDFs) for estimated half-lives of predicted targets of miR-125-5p and miR-19-3p vs. non-targets. **Figure S11.** Correlation of estimated RNA half-lives under various methods. **Figure S12.** Correlation of PRO-seq-based half-lives ( $T_{1/2}^{PR}$ , x-axis) vs. estimates from TimeLapse-seq ( $T_{1/2}^{TL5}$ , y-axis) after correcting for elongation rate. **Figures S13-S14.** SEM results for features of intron-containing (Figure S13) or intron-less (Figure S14) transcription units in K562. **Figure S15.** Estimated half-lives for intron-containing and intron-less transcription units. **Figure S16.** SEM results for features of intron-containing transcripts in K562 cells, with and without a correction for elongation rate. **Figure S17.** Multiple linear regression (MLR) for features of transcription units versus RNA stability in K562 cells. **Figures S18-S19.** DNA word enrichments in stable transcripts for protein-coding mRNAs (Figure S18) or lincRNAs (Figure S19). **Figure S20.** G+C content in intervals downstream of the TSS for various classes of transcription units. **Figure S21.** DNA word enrichments in stable transcripts for eRNAs. **Figure S22.** DNA methylation in lincRNAs of various stability levels. **Figure S23.** Sequence Stability Index of intron-containing versus intron-less genes. **Figures S24-S25.** Sequence Stability Index (SSI) for mRNAs (Figure S24) or lincRNAs (Figure S25) of various stability classes. **Figures S26-S27.** Histone modification signals for protein-coding mRNAs of various stability classes. Half-life estimations are based on published data (Figure S26) or newly collected data for this study (Figure S27). **Figures S28-S29.** Estimated SEM coefficients for transcription ( $\lambda_{ni}$ ) and half-life ( $\mu_{ni}$ ) for 11 histone modifications. It's assayed by ChIP-seq in the 500 bases (Figure S28) or 1000-1500 bases (Figure S29) downstream of the TSS. **Figures S30-S31.** Estimated SEM coefficients for transcription ( $\lambda_{ni}$ , top) and half-life ( $\mu_{ni}$ , bottom) for 11 histone modifications. It's based either on published data (Figure S30) or on the newly collected PRO-seq and RNA-seq data (Figure S31).

**Additional file 3.** Variation in Elongation Rate is Insufficient to Explain Variation in Half-Life.

### Acknowledgements

We thank members of the Siepel and Danko laboratories for stimulating discussions.

### Authors' contributions

A.S., A.B., and C.D. conceived the experiments. A.B., E.R., and A.C. conducted the experiments. Y. Z., A.B., Y.H., and N.D. performed the computational analyses. A.S., A.B., Y.Z., and K.K. contributed to the writing of the manuscript. All authors read and approved the final manuscript.

### Funding

This research was supported, in part, by US National Institutes of Health grants R35-GM127070 (to AS) and R01-HG009309 (to CGD), and by the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health.

### Availability of data and materials

All data generated or analyzed during this study are included in this published article, its supplementary information files and publicly available repositories. PRO-seq and RNA-seq data generated for this study are deposited on GEO under accession GSE153200. Published PRO-seq data for K562 [28] and HeLa [70] cell lines are retrieved from GEO under accession GSE60456 and GSE100742. Published RNA-seq data are retrieved from the ENCODE project [48, 71] (ENCSR000AEM for K562, ENCSR000CPR for HeLa). The source code used for data analysis and visualization is publicly available via Code Ocean at <https://codeocean.com/capsule/7351682/tree/v1>.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

<sup>1</sup>Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. <sup>2</sup>Present Address: Department of Biology and Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA, USA. <sup>3</sup>Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, NY, USA.

Received: 24 November 2020 Accepted: 5 January 2021

Published online: 15 February 2021

### References

- Hao S, Baltimore D. The stability of mRNA influences the temporal order of the induction of genes encoding inflammatory molecules. *Nat Immunol*. 2009;10:281–8.
- Rabani M, Levin JZ, Fan L, Adiconis X, Raychowdhury R, Garber M, et al. Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat Biotechnol*. 2011;29:436–42.
- Rabani M, Raychowdhury R, Jovanovic M, Rooney M, Stumpo DJ, Pauli A, et al. High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. *Cell*. 2014;159:1698–710.
- Schwanhauser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, et al. Global quantification of mammalian gene expression control. *Nature*. 2011;473:337–42.
- Sharova LV, Sharov AA, Nedozovov T, Piao Y, Shaik N, Ko MS. Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. *DNA Res*. 2009;16:45–58.
- Tani H, Mizutani R, Salam KA, Tano K, Ijiri K, Wakamatsu A, et al. Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Res*. 2012;22:947–56.
- Yang E, van Nimwegen E, Zavolan M, Rajewsky N, Schroeder M, Magnasco M, et al. Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res*. 2003;13:1863–72.
- Schwalb B, Michel M, Zacher B, Fruhauf K, Demel C, Tresch A, et al. TT-seq maps the human transient transcriptome. *Science*. 2016;352:1225–8.
- Mukherjee N, Calviello L, Hirsekorn A, de Pretis S, Pelizzola M, Ohler U. Integrative classification of human coding and noncoding genes through RNA metabolism profiles. *Nat Struct Mol Biol*. 2017;24:86–96.
- Herzog VA, Reicholf B, Neumann T, Rescheneder P, Bhat P, Burkard TR, et al. Thiol-linked alkylation of RNA to assess expression dynamics. *Nat Methods*. 2017;14:1198–204.
- Lam LT, Pickeral OK, Peng AC, Rosenwald A, Hurt EM, Giltane JM, et al. Genomic-scale measurement of mRNA turnover and the mechanisms of action of the anti-cancer drug flavopiridol. *Genome Biol*. 2001;2:RESEARCH0041.
- Gosline SJ, Gurtan AM, JnBaptiste CK, Bosson A, Milani P, Dalin S, et al. Elucidating microRNA regulatory networks using transcriptional, post-transcriptional, and histone modification measurements. *Cell Rep*. 2016;14:310–9.
- Hynes NE, Phillips SL. Turnover of polyadenylate-containing ribonucleic acid in *Saccharomyces cerevisiae*. *J Bacteriol*. 1976;125:595–600.
- Kim CH, Warner JR. Mild temperature shock alters the transcription of a discrete class of *Saccharomyces cerevisiae* genes. *Mol Cell Biol*. 1983;3:457–65.
- Wada T, Becskei A. Impact of methods on the measurement of mRNA turnover. *Int J Mol Sci*. 2017;18(12):2723. <https://doi.org/10.3390/ijms18122723>. PMID: 29244760; PMCID: PMC5751324.
- Raghavan A, Ogilvie RL, Reilly C, Abelson ML, Raghavan S, Vasdevani J, et al. Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes. *Nucleic Acids Res*. 2002;30:5529–38.
- Dolken L, Ruzsics Z, Radle B, Friedel CC, Zimmer R, Mages J, et al. High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA*. 2008;14:1959–72.
- Kenzelmann M, Maertens S, Hergenroth M, Kueffer S, Hotz-Wagenblatt A, Li L, et al. Microarray analysis of newly synthesized RNA in cells and animals. *Proc Natl Acad Sci U S A*. 2007;104:6164–9.
- Windhager L, Bonfert T, Burger K, Ruzsics Z, et al. Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. *Genome*. 2012; <http://genome.cshlp.org/content/22/10/2031.short>.
- Schofield JA, Duffy EE, Kiefer L, Sullivan MC, Simon MD. TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding. *Nat Methods*. 2018;15:221–5.
- Alkallas R, Fish L, Goodarzi H, Najafabadi HS. Inference of RNA decay rate from transcriptional profiling highlights the regulatory programs of Alzheimer's disease. *Nat Commun*. 2017;8:909.
- Gaidatzis D, Burger L, Florescu M, Stadler MB. Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. *Nat Biotechnol*. 2015;33:722–9.
- Kwak H, Fuda NJ, Core LJ, Lis JT. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science*. 2013;339:950–3.
- Churchman LS, Weissman JS. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*. 2011;469:368–73.
- Zeisel A, Köstler WJ, Molotski N, Tsai JM, Krauthgamer R, Jacob-Hirsch J, et al. Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. *Mol Syst Biol*. 2011;7:529.
- Chu T, Rice EJ, Booth GT, Salamanca HH, Wang Z, Core LJ, et al. Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme. *Nat Genet*. 2018;50:1553–64.
- Danko CG, Hyland SL, Core LJ, Martins AL, Waters CT, Lee HW, et al. Identification of active transcriptional regulatory elements from GRO-seq data. *Nat Methods*. 2015;12:433–8.
- Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet*. 2014;46:1311–20.
- Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res*. 2019;47:D766–73.
- Danko CG, Hah N, Luo X, Martins AL, Core L, Lis JT, et al. Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. *Mol Cell*. 2013;50:212–22.

31. Veloso A, Kirkconnell KS, Magnuson B, Biewen B, Paulsen MT, Wilson TE, et al. Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Res.* 2014;24:896–905.
32. Baudrimont A, Voegeli S, Vilorio EC, Stritt F, Lenon M, Wada T, et al. Multiplexed gene control reveals rapid mRNA turnover. *Sci Adv.* 2017;3:e1700006.
33. Weingarten-Gabbay S, Segal E. A shared architecture for promoters and enhancers. *Nat Genet.* 2014;46:1253–4.
34. Wei Q, Lei R, Hu G. Roles of miR-182 in sensory organ development and cancer. *Thorac Cancer.* 2015; <https://onlinelibrary.wiley.com/doi/abs/10.1111/1759-7714.12164>.
35. Wachutka L, Caizzi L, Gagneur J, Cramer P. Global donor and acceptor splicing site kinetics in human cells. *eLife.* 2019;8:e45056.
36. Wu Q, Medina SG, Kushawah G, DeVore ML, Castellano LA, Hand JM, et al. Translation affects mRNA stability in a codon-dependent manner in human cells. *eLife.* 2019;8:e45396.
37. Mele M, Mattioli K, Mallard W, Shechner DM, Gerhardinger C, Rinn JL. Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. *Genome Res.* 2017;27:27–37.
38. Hamer DH, Leder P. Splicing and the formation of stable RNA. *Cell.* 1979;18:1299–302.
39. Yves R. Lavaan: an R package for structural equation modeling. *J Stat Softw.* 2012;48:1–36.
40. Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO. Precision and functional specificity in mRNA decay. *Proc Natl Acad Sci U A.* 2002;99:5860–5.
41. Zhao C, Hamilton T. Introns regulate the rate of unstable mRNA decay. *J Biol Chem.* 2007;282:20230–7.
42. Clark MB, Johnston RL, Inostroza-Ponta M, Fox AH, Fortini E, Moscato P, et al. Genome-wide analysis of long noncoding RNA stability. *Genome Res.* 2012;22:885–98.
43. Gressel S, Schwalb B, Decker TM, Qin W, Leonhardt H, Eick D, et al. CDK9-dependent RNA polymerase II pausing controls transcription initiation. *eLife.* 2017;6:e29736.
44. Spies N, Burge CB, Bartel DP. 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res.* 2013;23:2078–90.
45. Almada AE, Wu X, Kriz AJ, Burge CB, Sharp PA. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature.* 2013;499:360–3.
46. Ntini E, Järvelin AI, Bornholdt J, Chen Y, Boyd M, Jørgensen M, et al. Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. *Nat Struct Mol Biol.* 2013;20:923–8.
47. Bailey TL. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics.* 2011;27:1653–9.
48. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012;489:57–74.
49. Vucic EA, Wilson IM, Campbell JM, Lam WL. Methylation analysis by DNA immunoprecipitation (MeDIP). *Methods Mol Biol.* 2009;556:141–53.
50. Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. Regulation of alternative splicing by histone modifications. *Science.* 2010;327:996–1000.
51. Braun KA, Young ET. Coupling mRNA synthesis and decay. *Mol Cell Biol.* 2014;34:4078–87.
52. Slobodin B, Han R, Calderone V, Vrieling JAFO, Loayza-Puch F, Elkon R, et al. Transcription impacts the efficiency of mRNA translation via co-transcriptional N6-adenosine methylation. *Cell.* 2017;169:326–337.e12.
53. Slobodin B, Bahat A, Sehrawat U, Becker-Herman S, Zuckerman B, Weiss AN, et al. Transcription dynamics regulate poly(A) tails and expression of the RNA degradation machinery to balance mRNA levels. *Mol Cell.* 2020;78:434–444.e5.
54. Maekawa S, Imamachi N, Irie T, Tani H, Matsumoto K, Mizutani R, et al. Analysis of RNA decay factor mediated RNA stability contributions on RNA abundance. *BMC Genomics.* 2015;16:154.
55. Tilgner H, Knowles DG, Johnson R, Davis CA, Chakraborty S, Djebali S, et al. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lincRNAs. *Genome Res.* 2012;22:1616–25.
56. Lugowski A, Nicholson B, Rissland OS. Determining mRNA half-lives on a transcriptome-wide scale. *Methods.* 2018;137:90–8.
57. Zhao S, Zhang Y, Gamini R, Zhang B, von Schack D. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci Rep.* 2018;8:4781.
58. Garneau NL, Wilusz J, Wilusz CJ. The highways and byways of mRNA decay. *Nat Rev Mol Cell Biol.* 2007;8:113–26.
59. Sibley CR, Emmett W, Blazquez L, Faro A, Haberman N, Briese M, et al. Recursive splicing in long vertebrate genes. *Nature.* 2015;521:371–5.
60. Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M. High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol.* 2006;4:e180.
61. Urrutia AO, Hurst LD. The signature of selection mediated by expression on human genes. *Genome Res.* 2003;13:2260–4.
62. Versteeg R, van Schaik BDC, van Batenburg MF, Roos M, Monajemi R, Caron H, et al. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res.* 2003;13:1998–2004.
63. Arhondakis S, Clay O, Bernardi G. GC level and expression of human coding sequences. *Biochem Biophys Res Commun.* 2008;367:542–5.
64. Sémon M, Mouchiroud D, Duret L. Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance. *Hum Mol Genet.* 2005;14:421–7.
65. Young JI, Hong EP, Castle JC, Crespo-Barreto J, Bowman AB, Rose MF, et al. Regulation of RNA splicing by the methylation-dependent transcriptional repressor methyl-CpG binding protein 2. *Proc Natl Acad Sci U A.* 2005;102:17551–8.
66. Goler-Baron V, Selitrennik M, Barkai O, Haimovich G, Lotan R, Choder M. Transcription in the nucleus and mRNA decay in the cytoplasm are coupled processes. *Genes Dev.* 2008;22:2022–7.
67. Shalem O, Groisman B, Choder M, Dahan O, Pilpel Y. Transcriptome kinetics is governed by a genome-wide coupling of mRNA production and degradation: a role for RNA pol II. *PLoS Genet.* 2011;7:e1002273.
68. Haimovich G, Medina DA, Causse SZ, Garber M, Millán-Zambrano G, Barkai O, et al. Gene expression is circular: factors for mRNA degradation also foster mRNA synthesis. *Cell.* 2013;153:1000–11.
69. Bregman A, Avraham-Kelbert M, Barkai O, Duek L, Guterman A, Choder M. Promoter elements regulate cytoplasmic mRNA decay. *Cell.* 2011;147:1473–83.
70. Nilson KA, Lawson CK, Mullen NJ, Ball CB, Spector BM, Meier JL, et al. Oxidative stress rapidly stabilizes promoter-proximal paused Pol II across the human genome. *Nucleic Acids Res.* 2017;45:11088–105.
71. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* 2018;46:D794–801.
72. Mahat DB, Kwak H, Booth GT, Jonkers IH, Danko CG, Patel RK, et al. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat Protoc.* 2016;11:1455–76.
73. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal.* 2011;17:10–2.
74. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol.* 2019;37:907–15.
75. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinforma Oxf Engl.* 2015;31:166–9.
76. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* 2012;131:281–5.
77. Jonkers I, Kwak H, Lis JT. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *eLife.* 2014;3:e02407.
78. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics.* 2005;21:3439–40.
79. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc.* 2009;4:1184–91.
80. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. *Nat Methods.* 2017;14:417–9.
81. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics.* 2009;25:1422–3.



82. Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit Anal.* 2007;15:199–236.
83. Ho DE, Imai K, King G, Stuart EA. MatchIt: nonparametric preprocessing for parametric causal inference. *J Stat Softw.* 2011;42. <https://doi.org/10.18637/jss.v042.i08>.
84. Agarwal V, Bell GW, Nam J-W, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. *Elife.* 2015;4. <https://doi.org/10.7554/eLife.05005>.
85. Akalin A, Franke V, Vlahoviček K, Mason CE, Schübeler D. Genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics.* 2015;31:1127–9.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

