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

Amit Blumberg[1], Yixin Zhao[1], Yi-Fei Huang[1,*], Noah Dukler[1], Edward J. Rice[2], Katie Krumholz[1], Charles G. Danko[2], and Adam Siepel[1,#]

[1]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
[2]Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, NY, USA
[*]Current address: Department of Biology and Huck Institutes of the Life Sciences,
Pennsylvania State University, University Park, PA, USA
[#]To whom correspondence should be addressed: asiepel@cshl.edu

Abstract

The rate at which RNA molecules are degraded is a key determinant of cellular RNA concentrations, yet current approaches for measuring RNA half-lives 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 and sequencing (PRO-seq) and RNA sequencing (RNA-seq). 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 degradation required for a steady-state equilibrium. We show that this approach can be used to assay relative RNA half-lives genome-wide, with reasonable accuracy and good 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, including intron length, are positively correlated with RNA stability, whereas features related to miRNA binding, DNA methylation, and G+C-richness are negatively correlated with RNA stability. Furthermore, we find that a measure of predicted stability based on U1 binding sites 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 after controlling for their correlations with transcription. Together, our estimation method and systematic analysis shed light on the pervasive impacts of RNA stability on cellular RNA concentrations.

Introduction

Gene regulation is an exquisitely complex and multifaceted process that operates at all stages of gene expression, ranging from pre-transcriptional chromatin remodeling to post-translational modification of proteins. Nevertheless, the concentration of RNA molecules in the cell is a key intermediate quantity in this process and serves 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 degradation, and therefore rates of RNA degradation are potentially as important as rates of production in determining concentrations of RNA molecules[1–7]. Indeed, there is evidence that bulk differences in RNA concentrations result in large part from substantial differences in RNA degradation rates across different kinds of transcription units (TUs). For example, protein-coding mRNAs, on average, are relatively stable, whereas lincRNAs are somewhat less stable, and enhancer RNAs (eRNAs) and other short noncoding RNAs tend to be extremely unstable[7–9]. Among protein-coding genes, mRNAs associated with housekeeping functions tend to be fairly 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[1,5,6,10,11]. In some cases, RNA degradation is accelerated by condition- or cell-type-specific expression of micro-RNAs or RNA-binding proteins[7,12].

Over a period of more than four decades, investigators have developed numerous methods for measuring RNA decay rates or half-lives[13–15]. A classical approach to this problem, still in use today, is to measure the decay in RNA abundance over time following inhibition of transcription, often using the antibiotic actinomycin D[1,3,16]. More recently, many studies have employed a strategy that is conceptually similar but considerably 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 due to RNA turnover, 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, also known as s^4U), which is rapidly taken up by animal cells and can be biotinylated for affinity purification[4,7,9,17–19]. Related methods use chemical conversion of 4sU nucleotide analogs to allow direct identification by sequencing and avoid the need for affinity purification[11,20]. 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 various limitations of the assays, differences in the populations of accessible TUs, and potential disruptions to cellular physiology, 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, convenient, or versatile as RNA-seq is for measuring cellular RNA concentrations, or PRO-seq[21] and NET-seq[22] are for measuring nascent transcription.

Recently, it has been shown that condition- or cell-type-dependent changes to RNA half-lives can be identified in a simpler manner, by working directly from high-throughput RNA-seq data[12,23–25]. 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 concentrations. Changes in half-life can then be estimated from changes to the ratio of these quantities, under the assumption of a steady-state equilibrium between RNA production and degradation. This approach is crude in several respects. For example, it 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 condition-specific differences in RNA processing rates[25]. Moreover, the dependency on intronic reads limits the method to spliced transcripts that are transcribed at relatively high levels. Finally, this method produces only relative, rather than absolute, estimates of RNA half-lives. Nevertheless, this simple approach has the important advantage of requiring no time course, metabolic labeling, transcriptional inhibition, chemical conversion, RNA pull-down, or indeed, any experimental innovation beyond standard RNA-seq. As a result, it can be an inexpensive and effective strategy for identifying genes undergoing cell-type- or condition-specific degradation[12,24,25].

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 estimates of relative RNA half-life that have improved accuracy and are broadly useful for downstream analysis. This approach requires only the application of two standard and widely applicable experimental protocols—PRO-seq and RNA-seq—to matched cells. Importantly, it applies to unspliced as well as spliced transcripts, it requires no correction for RNA-processing rates, and it is sufficiently sensitive to assay TUs expressed at low levels, including many noncoding RNAs. We validate estimates of RNA half-life for K562 cells by comparing them with estimates based on TimeLapse-seq[20], and by showing that several subclasses of TUs have expected patterns of relative RNA stability. We then perform a systematic analysis to identify features of TUs, DNA sequences, and epigenomic marks that are specifically associated with RNA half-life, after controlling for their associations with transcription. Together, these analyses establish combined RNA-seq and PRO-seq measurements as a simple but powerful means for assaying RNA stability, and shed light on several possible determinants of RNA degradation.

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. To reduce technical noise we collected new data of each type in multiple replicates (two for PRO-seq, four for RNA-seq), all from the same source of K562 cells. After verifying high concordance between replicates (**Supplemental Fig. 1**), we combined replicates to create a single PRO-seq and a single RNA-seq data set. When analyzing these data, we considered all annotated TUs in GENCODE[26], dividing them into mRNA (*n*=16,338), lincRNA (*n*=2,880), antisense (*n*=2,636), and pseudogene (*n*=2,653) classes. For each data type and each TU, we estimated expression by the total number of mapped reads in transcripts per million (TPM), a measure that normalizes by both library size and TU length. We excluded the first 250 bp downstream of the TU for PRO-seq to avoid a bias from promoter-proximal pausing[27], and discarded TUs with insufficient read counts from either assay (see **Methods**).

We found that the PRO-seq and RNA-seq measurements were well correlated overall, with Spearman's ρ=0.83 (**Fig. 1**), 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 ρ=0.90 (**Supplemental Fig. 2**). At the same time, there were considerable differences in the degree of correlation across classes of TUs, ranging from a high of ρ=0.85 for mRNAs to ρ=0.72 for lincRNAs, ρ=0.71 for antisense genes, and only ρ=0.57 for pseudogenes (**Fig. 1**). Similarly, the slopes of the lines of best fit on the log/log scatter plots decreased substantially (by roughly 50%) when proceeding from mRNAs to the noncoding RNAs and pseudogenes. We observed similar patterns for spliced and unspliced genes, but reduced values of ρ and slopes overall in unspliced genes (**Supplemental Figs. 3 & 4**). Together, these observations suggest that RNA degradation rates have a more pronounced effect on steady-state RNA levels in noncoding RNAs and pseudogenes. To ensure that these differences were not simply artifacts of differences in expression level across TU classes, we repeated the comparison with sets of genes matched by expression level (**Methods**; **Supplemental Fig. 5**) and observed similar results. We also repeated the analysis in GM12878 cells with comparable results (**Supplemental Fig. 6**).

A potential confounding factor in this comparison is elongation rate. Because PRO-seq read depth reflects a combination of transcription initiation rates and elongation rates[28,29], some reduction in correlation with RNA-seq could reflect variability across TUs in elongation rate. However, when we examined a subset of ~2000 genes for which elongation rates have been estimated for the same cell type[30] and explicitly adjusted for the estimated rates, we observed no improvement (indeed, a slight decline) in the correlation of PRO-seq and RNA-seq measurements (**Supplemental Fig. 7**). Thus, elongation rate does not appear to be a dominant factor in the analysis (but see **Discussion**).

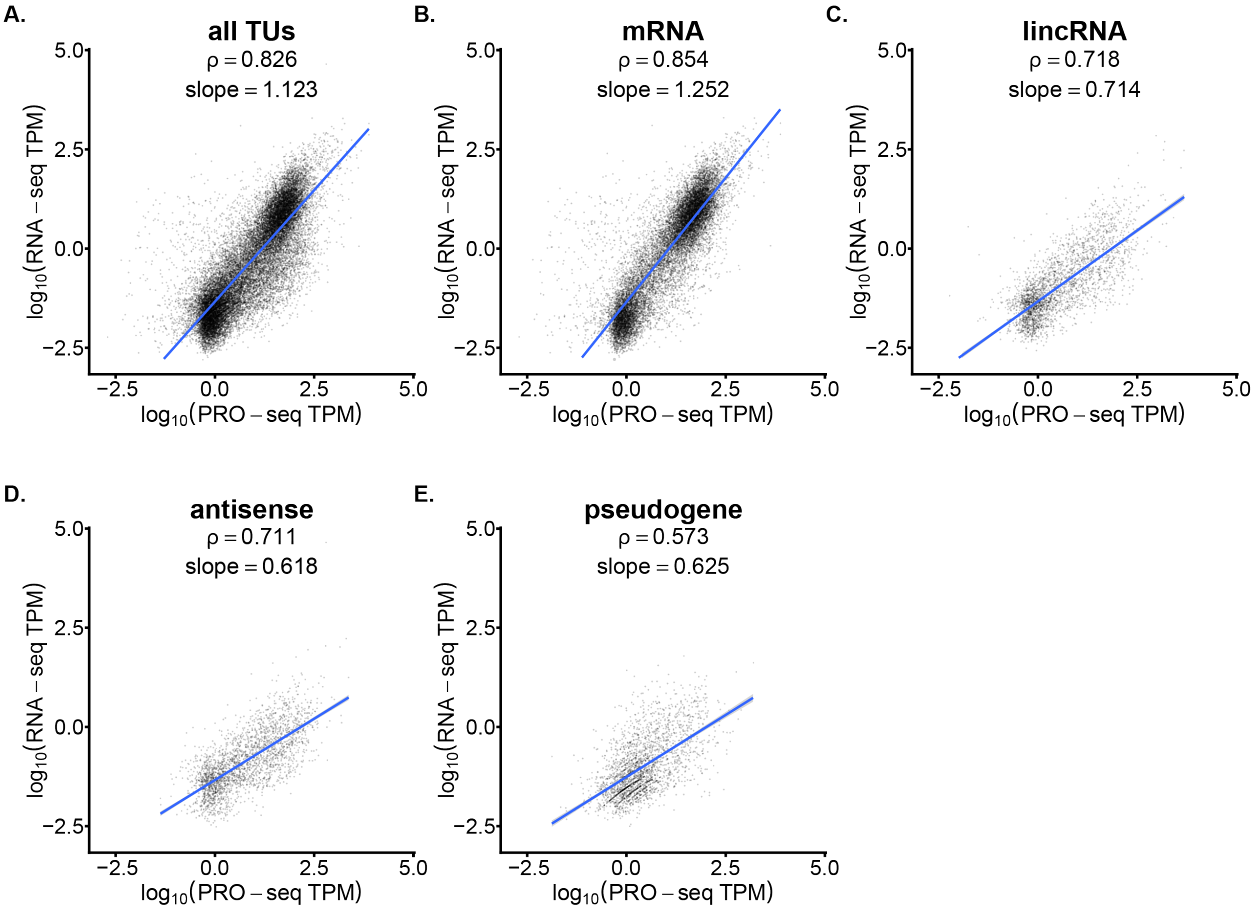

Figure 1. Scatter plots of PRO-seq vs. RNA-seq read counts for transcription units (TUs) in K562 cells, both shown in units of $\log_{10}$ transcripts per million (TPM) (see **Methods**). Panels describe (**A**) all annotated TUs (*n*=24,966), (**B**) mRNAs (*n*=16,338), (**C**) intergenic lincRNAs (*n*=2,880), (**D**) intragenic antisense non-coding genes (*n*=2,636), and (**E**) pseudogenes (*n*=2,653), all from GENCODE[26]. For each plot, the linear regression line is shown together with Spearman's rank-order correlation coefficient (ρ) and the slope of the regression line. Notice that as one proceeds from panel **B** to panel **E**, from mRNAs to noncoding RNAs and pseudogenes, there is a general decrease in both ρ, indicating greater variability of steady-state RNA concentrations at each transcription level, and the slope, indicating reduced average RNA concentrations for highly transcribed TUs.

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[24,25]. Briefly, if β_i is the rate of production of new RNAs for each TU i, α_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** & **Methods**). If we assume that β_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 (where "*PR*" denotes a PRO-seq/RNA-seq-based estimator). 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 unspliced TUs and increased sensitivity for TUs expressed at low levels. 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.

Following this approach, we estimated $T_{1/2}^{PR}$ values for TUs from across the genome using our 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 5,112 genes measured by both methods. We found that the two sets of estimates were reasonably well correlated (Spearman's ρ=0.54; **Fig. 2B**), especially considering the substantial differences in experimental protocols and the generally poor concordance of published half-life estimates across experimental methods[6,15]. Moreover, if we remove the 50% of genes expressed at the lowest levels (as measured by PRO-seq), for which the noise contribution will tend to be largest, the correlation improves to ρ=0.61.

To compare our PRO-seq-based approach with an approach based on intronic reads, we repeated the estimation for about 21,000 TUs in our data set for which we could retrieve adequate numbers of RNA-seq reads mapped to introns (see **Methods**). We found that these intron-based estimates of half-life, $T_{1/2}^{intr}$, also correlated with the estimates from TimeLapse-seq but the correlation was substantially poorer than for the PRO-seq-based estimates (only ρ=0.20 for 5,039 TUs accessible to both methods; **Supplemental Fig. 8**), suggesting that the PRO-seq-based approach provides less noisy estimates of transcription, and hence, reduced variance in estimates of half-life.

As additional validation, we considered TUs for two classes of genes that have consistently been shown to exhibit unusually low, or high, levels of RNA stability: zinc finger proteins and ribosomal proteins, respectively. We found, as expected, that the estimated $T_{1/2}^{PR}$ values were significantly shifted toward lower values—indicating higher turnover rates—for zinc finger proteins (**Fig. 2C**), many of which play key regulatory roles. By contrast, the estimated $T_{1/2}^{PR}$ values were significantly shifted toward higher values—indicating lower turnover rates—for ribosomal proteins, which tend to have fairly stable levels of expression across cell types and conditions and are considered "housekeeping" genes. Similarly, we tested TUs having experimentally verified target sites for several miRNAs that are expressed in K562 cells. In this case, we found that the targets of numerous miRNAs, including the well-studied[31] miR-182 (**Fig. 2D**), have significantly reduced stability (the reduction in half-life was significant for the predicted targets of 55 out of 217 miRNAs; see **Supplemental Fig. 9** for additional examples). We also found that the estimated half-lives of the same TUs in two different cell lines (K562 and GM12878) were fairly consistent (ρ=0.660).

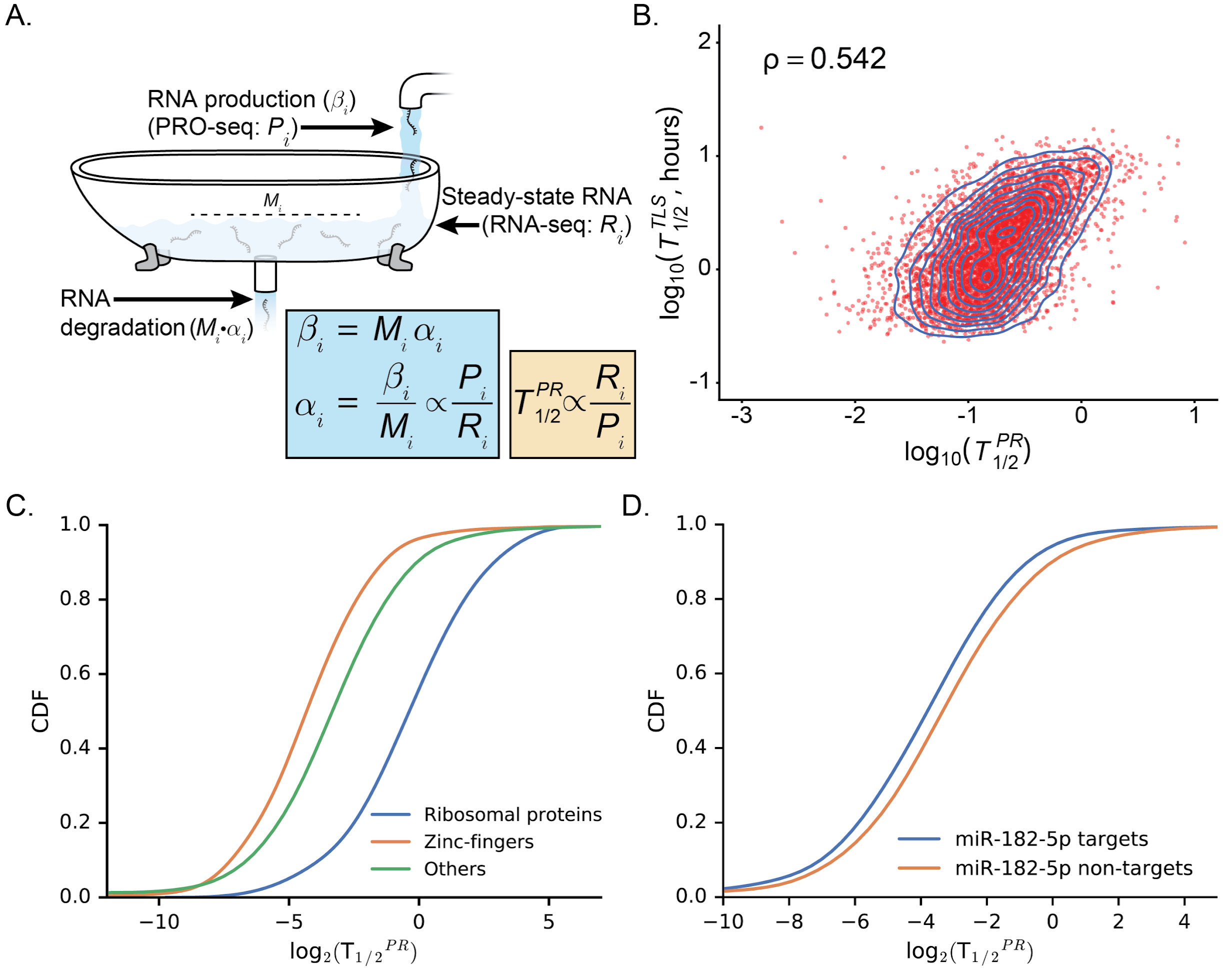

Figure 2. (A) Illustration of dynamic equilibrium between production and degradation of RNA. PRO-seq (P_i) can be used to measure production and RNA-seq (R_i) to measure the resulting equilibrium RNA concentration. At steady-state, the production and degradation rates must be equal, allowing for estimation of a quantity proportional to RNA half-life ($T_{1/2}{}^{PR}$) by the ratio P_i / R_i (see **Methods**). Illustration adapted from ref. [32]. **(B)** Scatter plot with density contours for ($\log_{10}$) half-lives estimated by the PRO-seq/RNA-seq method ($T_{1/2}{}^{PR}$, *x*-axis) vs. those estimated by TimeLapse-seq[20] ($T_{1/2}{}^{TLS}$, *y*-axis) for 5,112 TUs assayed by both methods in K562 cells. The $T_{1/2}{}^{PR}$ values are unit-less, whereas the $T_{1/2}{}^{TLS}$ values are expressed in hours. **(C)** Cumulative distribution functions (CDF) for ($\log_2$) estimated RNA half-lives, $T_{1/2}{}^{PR}$, for ribosomal proteins, zinc-finger proteins, and other genes. **(D)** Similar CDFs for mRNAs predicted to be targets of miR-182-5p vs. non-targets.

Properties of transcription units that are predictive of RNA stability

We sought to shed light on determinants of RNA stability by identifying 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[2,33], we focused our analysis on spliced TUs. For the features of mRNAs, we considered the splice junction density (number of exons divided by the total length of all exons), the G+C content and length of the annotated coding sequence (CDS), the G+C content and total length of all introns, and the G+C content and length of each of the 5'UTR and the 3'UTR. The features for lincRNAs were similar but, naturally, did not distinguish between UTRs and CDSs. We carried out a parallel analysis of unspliced TUs and found qualitatively similar results (not shown).

A typical approach to this analysis would be to measure the correlation of each feature with half-life, either individually or together in a multiple regression framework. Because $T_{1/2}^{PR}$ is estimated from the RNA-seq/PRO-seq ratio, however, it will tend to be statistically correlated with features predictive of transcription regardless of their true influence on half-life. To identify features of TUs that are predictive of RNA abundance after accounting for their correlation with transcription, we instead made use of a Structural Equation Model (SEM)[34] that explicitly describes the separate influences of features on transcription and half-life, and the contributions of both to RNA abundance (see **Methods** & **Fig. 3A**). To our knowledge, this is the first attempt to identify features associated with RNA half-life that disentangles these separate influences (see **Discussion**).

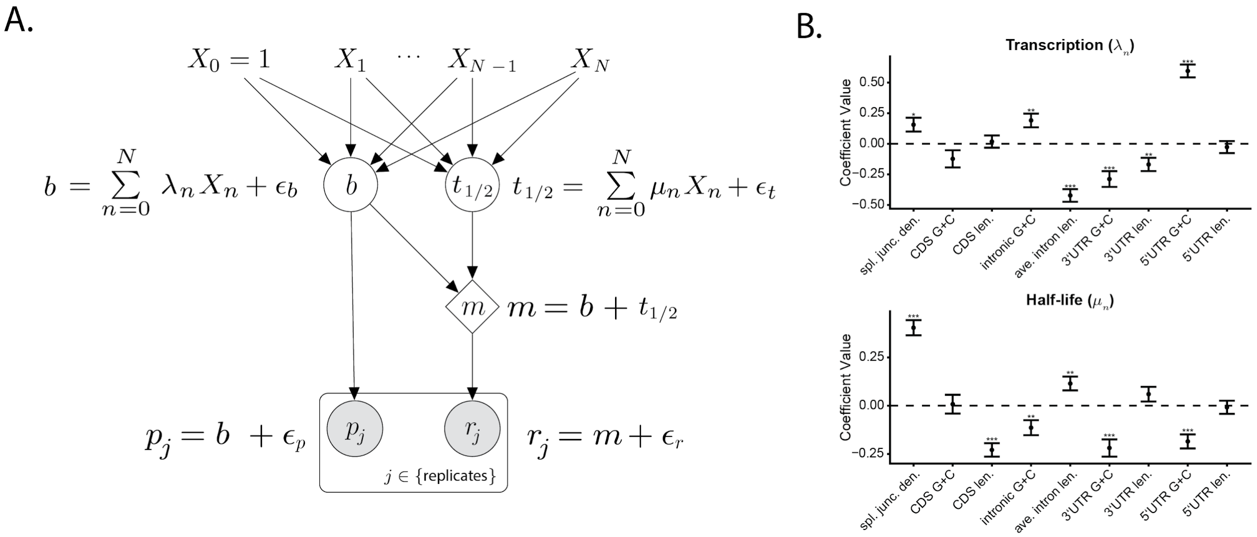

Figure 3. Features of transcription units (TUs) that are predictive of transcription rate and RNA half-life. **(A)** Structural Equation Model (SEM) describing the effects of an arbitrary collection of TU features ($X_1,\ldots,X_N$, with intercept term X_0=1) on transcription rate (b) and half-life ($t_{1/2}$), as well as the downstream impact on mRNA concentration (m), normalized PRO-seq (p), and normalized RNA-seq (r) read counts. The model is linear in logarithmic space, with unmodeled variation accounted for as Gaussian noise (ε_b, ε_t, ε_p, and ε_r; see **Methods**). The coefficients for transcription rate (λ_n) and half-life (μ_n) are estimated by maximum likelihood, assuming independence of replicates and pooling data from all TUs of the same class. **(B)** Estimated values for coefficients for transcription (λ_n; top) and half-life (μ_n; bottom) for various features of interest. Results are for spliced mRNAs (see **Supplemental Figs. 10 & 11** for other classes). Features considered for each TU: spl. junc. dens. – number of splice junctions divided by mature RNA length. CDS G+C – GC content in coding region. CDS len. – total length of coding region. Intronic G+C – GC content introns. Ave. intron len. – average of intron length. 3'UTR G+C – GC content in 3`UTR. 3' UTR len. - length of 3'UTR. 5'UTR G+C – GC content in 5'UTR. 5' UTR len. - length of 5'UTR. Error bars represent ±1.96 standard error, as calculated by the 'lavvan' R package[35]**.** Significance (from Z-score): * p<0.05; ** p<0.005; *** p<0.0005.

Our analysis revealed significant positive correlations with half-life of both splice junction density and average intron length, for spliced mRNAs and lincRNAs (**Fig. 3B; Supplemental Figs. 10 & 11**). The observation regarding splice junction density is consistent with previous reports for mRNAs[2,33,36,37] and lincRNAs[38], as well as with the general tendency for spliced TUs to be more stable than unspliced TUs (**Supplemental Fig. 12**). However, the correlation with intron length is new, to our knowledge, and may indicate that RNA stability is enhanced by recursive splice sites[39] or extended contact with the spliceosome in long introns (see **Discussion**). We also found that CDS length is negatively correlated with half-life, but this observation may reflect a positive correlation between CDS length and number of introns, and/or a confounding effect from elongation rate (see **Discussion**).

Several additional significant associations concerned G+C content, at the levels of both transcription and half-life. The most prominent of these is a positive correlation between G+C content in the 5'UTR and transcription. However, we also observe a fairly pronounced negative correlation between G+C content in the 5'UTR and RNA half-life, which is consistent with reports of an association between degradation rate and the number of CpG dinucleotides in the 5'UTR, hypothesized to reflect the action of methyl CpG-binding proteins that regulate splicing[2] (see **Discussion**). We observed a similar negative correlation with intronic G+C content, which could potentially also be related to methyl-CpG-mediated splicing. This negative correlation is also apparent with G+C content in the 3'UTR. Notably, a trend in the opposite direction would be expected if it were driven by driven by AU-rich elements (ARE) and PUF-binding sites, whose presence in the 3'UTR has been reported to be associated with reduced stability[2]. These G+C-related correlations held in lincRNA introns, but not in lincRNA exons (**Supplemental Fig. 10**). In general, the observed G+C-related patterns are somewhat difficult to interpret, because of the complex correlations among G+C content, CpGs, transcription, splicing, and RNA half-life (see **Discussion**).

Notably, several features had coefficients of opposite sign for transcription and half-life (e.g., intronic and 5'UTR G+C; CDS, intron, and 3'UTR length), despite our attempt to explicitly account for separate effects on both processes using the SEM. This pattern of anti-correlation could be driven, at least in part, by stabilizing selection on RNA levels, which could lead to evolutionary compensation at the transcription and degradation steps (see **Discussion**).

DNA sequence correlates of RNA stability

Our estimates of RNA half-life for both coding and noncoding TUs provide an opportunity to better characterize DNA sequence correlates of RNA stability near transcription start sites (TSSs)[2,40–42]. Toward this end, we first compared the sequences from 0 to 2500 bp downstream of the TSS for the 20% least and most stable TUs, according to the estimated $T_{1/2}^{PR}$, testing for global enrichments of all possible nucleotide *k*-mers in either class. We considered DNA word sizes of $k \in \{2, 3, 4\}$ and separately tested for enrichments in 1000 bp windows at various distances from the TSS (**Supplemental Figs. 13 & 14**). As above, we examined mRNAs and lincRNAs separately. In all cases, we matched the stable and unstable transcripts by their PRO-seq abundance estimates (see Methods) to avoid a bias from differences in transcription rates.

These tests identified a number of *k*-mers that were either enriched or depleted in stable transcripts, but these trends were almost completely explained 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**). Notice that, while this observation is generally consistent with the results of our SEM analysis, it specifically relates to G+C content near the TSS. Interestingly, these trends were largely independent of window position near the TSS, although they were slightly less pronounced in the first 500bp than in downstream windows **(Supplemental Fig. 15**). The patterns were similar for mRNAs and lincRNAs. An exception to the overall depletion of G+C-rich sequences occurred with CpG dinucleotides, which were slightly enriched in stable mRNAs (but not lincRNAs) near the TSS (**Supplemental Fig. 13**), a pattern that has been previously observed[2]. Using the discriminative motif finder DREME[43], we identified several A+T-rich motifs associated with stable transcripts, and several G+C-rich motifs associated with unstable

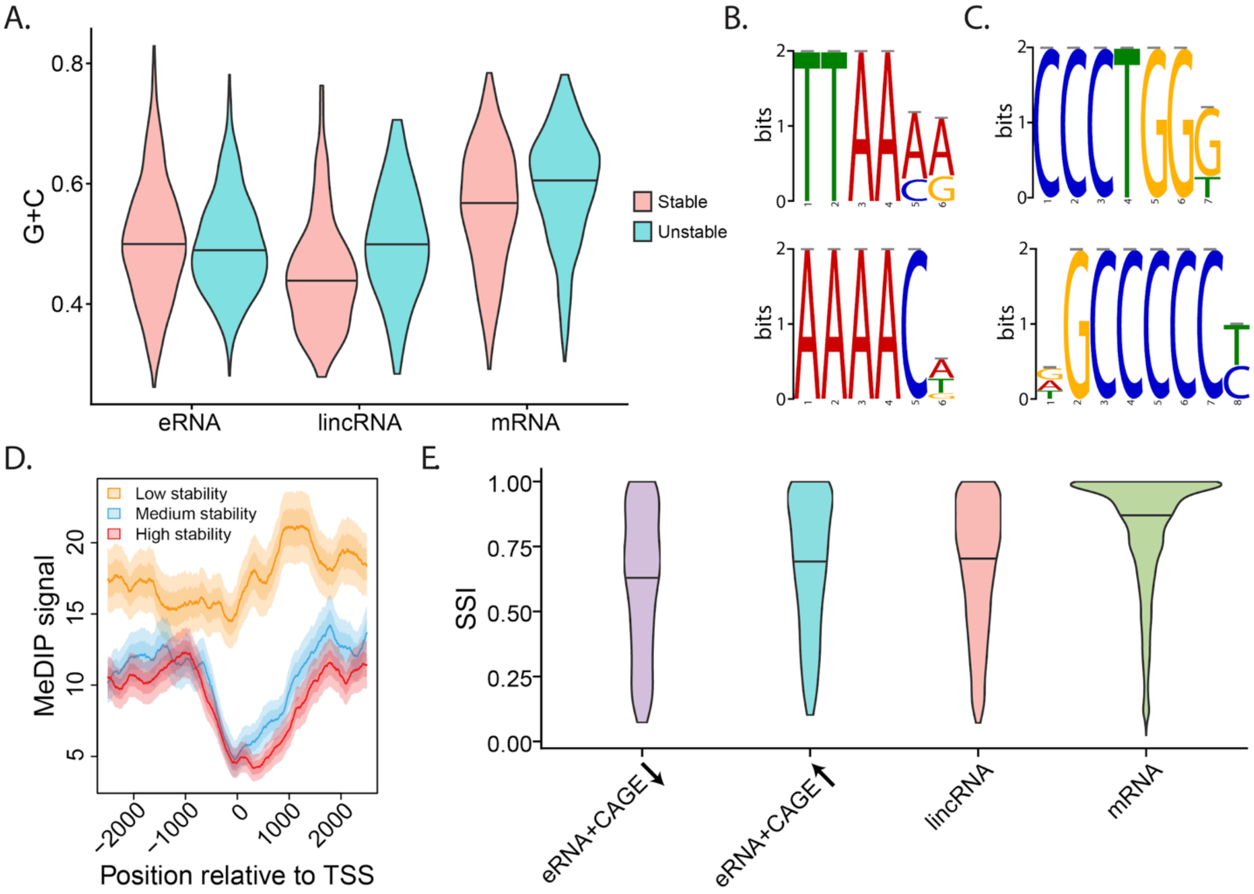

Figure 4. DNA-sequence, methylation, and RNA-binding-protein correlates of RNA stability near the TSS. **(A)** Distribution of G+C content (*y*-axis) for the 20% most (red) and least (blue) stable TUs, according to our estimated half-life ($T_{1/2}^{PR}$), in enhancer RNAs (eRNA), lincRNAs and mRNAs (*x*-axis). **(B&C)** Two 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 **Methods**) 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 **Methods**) for eRNAs, lincRNAs, and mRNAs. Separate plots are shown for eRNAs with low and high CAGE support, suggesting low and high stability, respectively.

transcripts (**Fig. 4B&C**), but we were unable to establish any clear biological significance for these motifs.

To shed further light on the local DNA sequence determinants of RNA stability, we expanded our set of TUs to include about 22,000 eRNAs from K562 cells, identified by GRO-cap in a previous study[42]. We obtained a relative measure of stability for these eRNAs based on the ratio of CAGE reads to PRO-seq reads within each TU (see **Methods**), using CAGE in this case because it was more sensitive than our RNA-seq data for eRNAs. We excluded eRNAs with no mapped CAGE reads. We then considered the 10% least stable and the 10% most stable of the remaining eRNAs (n=510 in each set), matching the two sets by normalized PRO-seq read counts downstream of the pause site. Interestingly, in this case, we found that stable eRNAs were enriched, rather than depleted, for G+C-rich sequences (**Fig. 4A**). This trend was strongly evident only for the first 400bp downstream of the TSS. It was especially pronounced for CpG dinucleotides (**Supplemental Fig. 16**). In contrast, AT (and to a lesser extent, TA) dinucleotides showed a fairly pronounced depletion for stable eRNAs.

The atypical patterns around CpG dinucleotides raise the possibility of an association with DNA methylation near the TSS. To address this question, we compared the average methylation levels of TUs exhibiting low, medium, or high levels of RNA stability. Specifically, we partitioned our mRNAs, considering spliced 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 (see **Methods**). We then produced meta-plots for each of these three classes showing the average signal of the methylated DNA immunoprecipitation (MeDIP-seq) assay in K562 cells[44,45] as a function of distance from the TSS. We found that the medium- and high-stability TUs exhibited similar patterns of methylation, with intermediate levels 1-2kb upstream and downstream of the TSS, and a pronounced dip at the TSS which extended to about 1kb downstream (**Fig. 4D**). The low-stability TUs, by contrast, show a clearly distinct pattern, with elevated methylation levels across the whole region, no pronounced dip at the TSS, and a peak about 1kb downstream. These observations suggest the possibility of epigenomic as well as DNA sequence differences associated with RNA stability, as we explore further below. Interestingly, these differences in methylation were not nearly as pronounced for lincRNAs (**Supplemental Fig. 17**).

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 degradation. 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[40,41]. 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 both coding and noncoding TUs as unstable or stable, respectively, with fairly high accuracy[42].

We applied this HMM (see **Methods**) 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. In all cases, we applied the HMM to the 1kb immediately downstream of the TSS, on the sense strand. 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, spliced lincRNAs have significantly higher SSIs than unspliced lincRNAs, although the difference for spliced and unspliced mRNAs was not as pronounced (**Supplemental Fig. 18**). Moreover, within each of the mRNA and lincRNA groups, we found that the SSI changed 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 (**Supplemental Figs. 19 & 20**). 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 **Discussion**).

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 they are also known to interact with the process of splicing[46], and for this reason or others they could influence RNA stability. 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[45], separately for low-, medium-, and high-stability classes of expression-matched spliced mRNAs (see **Methods**). 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 (**Supplemental Fig. 21**). For example, H3k79me2, which is associated with transcriptional activity, gives a substantially higher signal in stable transcripts than in unstable ones, particularly in a peak about 1kb downstream from the TSS (**Fig. 5A)**. A similar pattern is observed for H3K4me2. However, an inverse relationship is observed with H3K4me1, which is associated with active enhancers, with an elevated signal in unstable transcripts relative to stable ones.

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, Supplemental Fig. 22**). 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, H3K4me2, and H3K4me3, and negative correlations for the repressive marks H3K9me3 and H3K27me3. All of these patterns were consistent between lincRNAs and mRNAs (**Supplemental Fig. 22 & 23**), and they did not change substantially as a function of distance from the TSS, within 2kb (**Supplemental Fig. 24**). However, we did additionally identify several significant correlates of half-life. For mRNAs

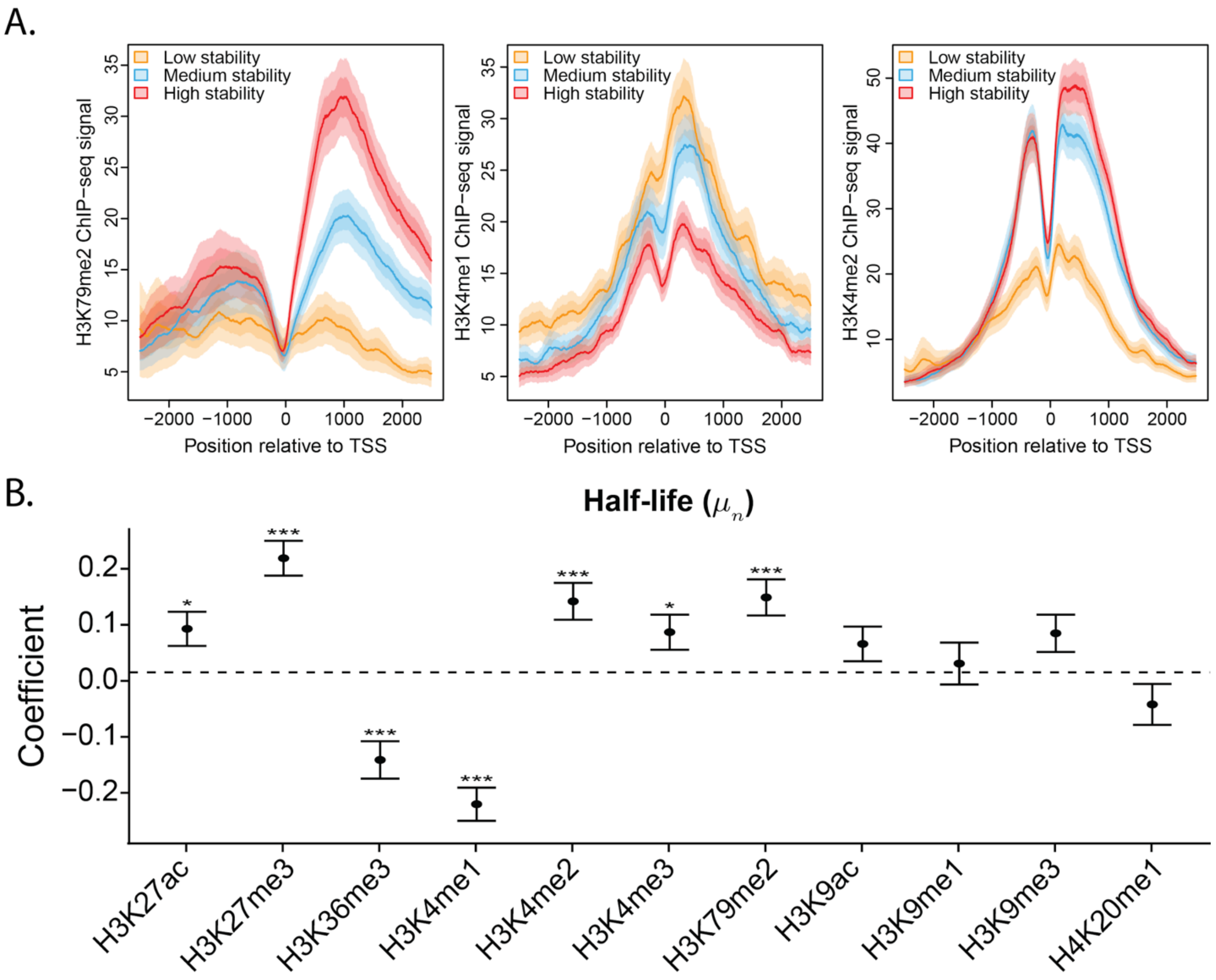

Figure 5. Histone-modification correlates of RNA stability. **(A)** ChIP-seq signal for H3K79me2 (*left*), H3K4me1 (*middle*), and H3K4me2 (*right*) for low-, medium-, and high-stability mRNAs (see **Methods**) as a function of distance from the TSS. Results are for spliced mRNAs matched by normalized PRO-seq signal. Solid line represents mean signal and lighter shading represents standard error and 95% confidence interval. **(B)** Estimated SEM coefficients for half-life (μ_n) for 11 histone modifications, as assayed by ChIP-seq in the 500 bases immediately downstream of the TSS, also for spliced mRNAs that were matched by transcription rate (**Methods**; see **Supplemental Figs. 22-24** for additional results). Error bars and significance are as in **Fig. 3B**.

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. 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 strong negative correlation with RNA half-life in lincRNAs but a weaker and position-dependent positive or negative correlation with mRNA half-life; and the silencing marks H3K9me1 and H3K9me3 show positive correlations for lincRNA half-life but negative or near-zero correlations for mRNA half-life (**Supplemental Fig. 24**). These divergent patterns could possibly reflect differences in the degree to which splicing is co-transcriptional in mRNAs and lincRNAs[47] .

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 of half-life only, based on standard high-throughput transcriptomic data. The use of PRO-seq data in place of intronic reads, however, brings with it several benefits, including the ability to interrogate unspliced TUs and TUs that are expressed at low levels. Moreover, even for spliced and abundantly expressed genes, the PRO-seq-based measurements appear to be considerably more accurate than those based on intronic reads, according to our comparisons with a third method, TimeLapse-seq. We have shown that our measurements of relative half-life are useful in a wide variety of downstream analyses, including the identification of various features that are predictive of RNA half-life.

To identify such features, 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[2], 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 degradation[2,37]. In addition to the previously reported positive correlation of splice junction density and RNA half-life, we found that intron length is also positively correlated with half-life. The causal basis of this correlation is also unknown, but intriguingly, long introns have also been reported to be enriched for recursive splice sites[39], suggesting that these “hidden” splice sites could contribute to increased RNA stability. The SEM also revealed some more difficult-to-interpret associations with G+C content and epigenomic marks (discussed further below).

An important finding in the recent literature on RNA stability is the observation 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). This observation suggests that RNA stability is determined, at least in part, by the DNA sequence near the TSS. In addition to its mechanistic implications, this “U1-PAS axis”[40] for the determination of stability has evolutionary implications; for example, it provides a potential mutational mechanism for the emergence of new genes[48]. In this study, we examined the U1-PAS axis from a slightly different perspective, testing not only whether it could distinguish TUs belonging to relatively stable classes (mRNAs) from those in unstable classes (uaRNAs and eRNAs) but also how predictive it is of relative RNA half-life within these classes. Using our previously developed HMM[42], we confirmed that a U1-PAS-based “sequence stability index” (SSI) is generally elevated for mRNAs, intermediate for lincRNAs, and reduced for eRNAs. Furthermore, the SSI can distinguish, to a degree, 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 (**Supplementary Figs. 19 & 20**). 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 determined by splicing-related processes, binding by miRNAs or RBPs, or other posttranscriptional phenomena. More work will be needed to fully understand the mechanistic basis of these differences in stability.

One important technical limitation of our method is that PRO-seq does not measure transcription directly, but rather 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 "signal" on a highway at night; an increase in signal can reflect either an increased number of cars passing by per unit time (analogous to an increased rate of transcription), or a back-up in traffic (analogous to a decreased elongation rate). As a consequence, variation in $T_{1/2}^{PR}$ across TUs could in part be driven by variation in elongation rate, with slower elongation rates leading to over-estimation of the transcription rate and therefore underestimation of $T_{1/2}^{PR}$, and faster elongation rates leading to under-estimation of transcription rate and over-estimation of $T_{1/2}^{PR}$. We attempted to control for a confounding effect from elongation rate by explicitly adjusting our PRO-seq abundance estimates based on published elongation rates for ~2000 genes, and found that it had little effect on the relationship to RNA-seq signal. However, more work will be needed to obtain more accurate and more comprehensive estimates of elongation rates, and to examine their impact on half-life estimates.

Some of the most pronounced associations that we observed with half-life, both near the TSS and across TUs, concerned G+C content. However, these observations are difficult to interpret owing to the complex patterns of correlation between G+C content and a wide variety of genomic and epigenomic features. 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[49–51], 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[52,53]. Sharova et al.[2] identified a fairly pronounced negative correlation between RNA stability and the prevalence of CpGs in the 5'UTR, which is consistent with—but not identical to—our observation of a negative correlation with G+C content in the 5'UTR. These authors raised the intriguing hypothesis this correlation may reflect the activity of splicing-associated methyl CpG-binding proteins[54], but, to our knowledge, this idea has not been tested experimentally. Sharova et al. did not examine G+C content in introns or the 3'UTR, where we also found negative correlations. 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 carefully designed experiments that directly perturb individual features of interest and separately measure the effects on a variety of transcriptional and post-transcriptional processes.

Our observations of epigenomic correlates of transcription and stability are similarly challenging to interpret. The observed negative correlation between DNA methylation and RNA stability near the TSS (**Fig. 4D**) is broadly consistent with Sharova et al.'s[2] observations regarding CpG content in the 5'UTR (and ours regarding G+C content), and with the hypothesis of a splicing association. However, we also identified several histone modifications that are significantly associated with increased or decreased half-life, using two complementary approaches—direct comparison of low-, medium-, and high-stability TUs after matching by PRO-seq signal (**Fig. 5A**) and joint consideration of all features using our SEM (**Fig. 5B**). At face value, an association between histone modifications and RNA stability would seem surprising, and 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 during co-transcriptional splicing. Divergent patterns in these histone marks for mRNAs and lincRNAs (**Supplemental Fig. 24**) 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 significantly different from zero in opposite directions (**Supplementary Figs. 22-24**). This trend of anti-correlation was less prominent with the TU features, but we did observe it with intronic and 5'UTR G+C content, and CDS, intron, and 3'UTR length (**Fig. 3B**). A possible explanation for this pattern is that it is 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. A related idea is that there is a fundamental evolutionary tradeoff in achieving a desired RNA concentration between low-transcription/high-stability solutions, which are energetically favorable, and high-transcription/low-stability solutions, which allow for rapid responses to stimuli[5]. Different TUs will fall at different points along the continuum between these solutions. Either of these evolutionary forces would create 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 interrelationship between transcription and stability is one that our SEM cannot fully disentangle—these evolutionary forces would effectively induce a correlation between transcription and half-life that is not considered in the model. As a result, it is 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. Nevertheless, our framework remains useful for identifying potentially interesting correlations, whose mechanistic underpinnings can then be further investigated through direct experimental perturbation.

Materials and Methods

PRO-seq and RNA-seq data preparation and processing

To minimize technical differences, we 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[55], followed by library preparation as previously published[56]. Sequencing was done by Novogene on a HiSeq instrument with paired-end reads of 2×150bp. For RNA-seq, RNA was extracted using the Trizol method (see 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.

Read mapping and transcript abundance estimation

Raw data files in fastq format were trimmed using TrimGalore with default parameters (`trim_galore fastqfile.fq.gz`). Reads were then aligned using 'STAR' with default parameters (`STAR --runMode alignReads --readFilesCommand zcat --quantMode GeneCounts --outSAMtype BAM`). We used the GRCh37/hg19 reference genome and the associated GENCODE gene annotations. For RNA-seq, the index for STAR was derived from the GENCODE GTF standard annotation file. For PRO-seq, we modified this GTF annotation file so that, for each gene, the entire transcript length (from transcription start site to termination site) was represented as a single "exon" in the STAR index. Importantly, however, for the purposes of read counting with PRO-seq we omitted the first 250 bases downstream of the TSS to avoid a bias in read counts from promoter proximal pausing. For RNA-seq, read counts per TU were obtained using 'kallisto' with default parameters (`kallisto quant -i index.idx --single -l 200 -s 30 -o`). Finally, we normalized read counts by converting them to transcripts per million (TPM)[58] 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, β_i, a constant per-RNA-molecular rate of decay, α_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 halflife 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 the pause peak) 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 $T_{1/2}^{PR} = R_i / P_i$.

Structural equation model (SEM)

To separate the effects of TU features on decay from the effects on transcription, we developed an SEM using the 'lavvan' R package[35]. 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: $b = \sum_{n=0}^{N} \lambda_n X_n + \varepsilon_b$ and $t_{1/2} = \sum_{n=0}^{N} \mu_n X_n + \varepsilon_t$ 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 λ_0 and μ_0 can be interpreted as intercepts whereas $\lambda_{n\neq0}$ and $\mu_{n\neq0}$ 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 $p_j \sim b + \varepsilon_p$ and $r_j \sim m + \varepsilon_r$ where $\epsilon_r \sim N(0, \sigma_r)$ 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 (λ_n) and half-life (μ_n) for all features by maximum likelihood.

Transcription unit features

Transcription unit (TU) sequences were downloaded from BioMart[61,62] (http://grch37.ensembl.org) using the R package biomaRt. We considered only one isoform per annotated gene, selecting the one with the highest estimated TPM by kallisto (above) and including only TUs having one clearly dominant isoform (i.e., where the isoform accounted for at least 75% of the TPM for the gene). Features based on properties of DNA sequences (e.g., G+C content) were then extracted using Biopython[63]. The intron length was set equal to the transcript length minus the total exon length. The splice junction density was set equal to twice the intron number divided by the coding sequence length.

eRNA analysis

We used eRNAs identified from our previous GRO-cap analysis in K562 cells[42] restricting our analysis to putative eRNAs with divergent transcription[57] that fell at least 1kb 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. To measure transcription rates, we used PRO-seq data from same study[42]. 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 1kb 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 600bp from the TSS.

Motif discovery

For motif discovery, we used the discriminative motif finder 'DREME'[43] with default parameters (core width ranging from 3-7) . 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 (SSI)

We define the SSI to be the probability that a TU is "stable" based on our previously published U1-PAS hidden Markov model (HMM)[42]. 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.[42]. We applied the HMM to the first 1kb 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'[59,60] 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.

Metaplots

Metaplots showing the average values of signals of interest across loci (e.g., **Figs. 4D & 5A**) were produced using the 'plotMeta' function from the 'Genomation'[64] 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 lined, with uncertainty indicated by the standard error of the mean (darker shading) and 95% confidence intervals (lighter shading) as specified by the “se” parameter.

MicroRNA targets analysis

We obtained microRNA targets from TargetScanHuman[65], Release 7.2 (http://www.targetscan.org/vert_72/vert_72_data_download/Predicted_Target_Locations.default_predictions.hg19.bed.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 Resources

Histone modifications, DNA methylation IP (MeDIP) and eCLIP data were downloaded from the ENCODE consortium[45] as bigwig files annotated to the GRCh37/hg19 reference genome (https://www.encodeproject.org/).

Acknowledgements

This research was supported, in part, by US National Institutes of Health grants R35-GM127070 (to AS) and R01-HG009309 (to CGD). The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health.

References

1. Yang, E. *et al.* Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res.* **13**, 1863–1872 (2003).
2. Sharova, L. V. *et al.* Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. *DNA Res.* **16**, 45–58 (2009).
3. Hao, S. & Baltimore, D. The stability of mRNA influences the temporal order of the induction of genes encoding inflammatory molecules. *Nat. Immunol.* **10**, 281–288 (2009).
4. Rabani, M. *et al.* Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat. Biotechnol.* **29**, 436–442 (2011).
5. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
6. Tani, H. & Akimitsu, N. Genome-wide technology for determining RNA stability in mammalian cells: historical perspective and recent advantages based on modified nucleotide labeling. *RNA Biol.* **9**, 1233–1238 (2012).
7. Rabani, M. *et al.* High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. *Cell* **159**, 1698–1710 (2014).
8. Tani, H. *et al.* Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Res.* **22**, 947–956 (2012).
9. Schwalb, B. *et al.* TT-seq maps the human transient transcriptome. *Science* **352**, 1225–1228 (2016).
10. Lam, L. T. *et al.* Genomic-scale measurement of mRNA turnover and the mechanisms of action of the anti-cancer drug flavopiridol. *Genome Biol.* **2**, RESEARCH0041 (2001).
11. Herzog, V. A. *et al.* Thiol-linked alkylation of RNA to assess expression dynamics. *Nat. Methods* **14**, 1198–1204 (2017).

12. Gosline, S. J. C. *et al.* Elucidating MicroRNA Regulatory Networks Using Transcriptional, Post-transcriptional, and Histone Modification Measurements. *Cell Rep.* **14**, 310–319 (2016).
13. Hynes, N. E. & Phillips, S. L. Turnover of polyadenylate-containing ribonucleic acid in Saccharomyces cerevisiae. *J. Bacteriol.* **125**, 595–600 (1976).
14. Kim, C. H. & Warner, J. R. Mild temperature shock alters the transcription of a discrete class of Saccharomyces cerevisiae genes. *Mol. Cell. Biol.* **3**, 457–465 (1983).
15. Wada, T. & Becskei, A. Impact of Methods on the Measurement of mRNA Turnover. *Int. J. Mol. Sci.* **18**, (2017).
16. Raghavan, A. *et al.* Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes. *Nucleic Acids Res.* **30**, 5529–5538 (2002).
17. Kenzelmann, M. *et al.* Microarray analysis of newly synthesized RNA in cells and animals. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 6164–6169 (2007).
18. Dolken, L. *et al.* High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA* **14**, 1959–1972 (2008).
19. Windhager, L., Bonfert, T., Burger, K. & Ruzsics, Z. Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. *Genome* (2012).
20. Schofield, J. A., Duffy, E. E., Kiefer, L., Sullivan, M. C. & Simon, M. D. TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding. *Nat. Methods* **15**, 221–225 (2018).
21. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**, 950–953 (2013).
22. Churchman, L. S. & Weissman, J. S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469**, 368–373 (2011).
23. Zeisel, A. *et al.* Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. *Mol. Syst. Biol.* **7**, 529 (2011).

24. Gaidatzis, D., Burger, L., Florescu, M. & Stadler, M. B. Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. *Nat. Biotechnol.* **33**, 722–729 (2015).

25. Alkallas, R., Fish, L., Goodarzi, H. & Najafabadi, H. S. Inference of RNA decay rate from transcriptional profiling highlights the regulatory programs of Alzheimer's disease. *Nat. Commun.* **8**, 909 (2017).

26. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

27. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).

28. Danko, C. G. *et al.* Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. *Mol. Cell* **50**, 212–222 (2013).

29. Jonkers, I., Kwak, H. & Lis, J. T. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *Elife* **3**, e02407 (2014).

30. Veloso, A. *et al.* Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Res.* **24**, 896–905 (2014).

31. Wei, Q., Lei, R. & Hu, G. Roles of miR-182 in sensory organ development and cancer. *Thoracic cancer* (2015).

32. Weingarten-Gabbay, S. & Segal, E. A shared architecture for promoters and enhancers. *Nat. Genet.* **46**, 1253–1254 (2014).

33. Hamer, D. H. & Leder, P. Splicing and the formation of stable RNA. *Cell* **18**, 1299–1302 (1979).

34. Kaplan, D. *Structural Equation Modeling: Foundations and Extensions.* (SAGE Publications, 2008).

35. Yves, R. Lavaan: An R package for structural equation modeling. *J. Stat. Softw.* **48**, 1–36 (2012).

36. Wang, H.-F., Feng, L. & Niu, D.-K. Relationship between mRNA stability and intron presence. *Biochem. Biophys. Res. Commun.* **354**, 203–208 (2007).

37. Zhao, C. & Hamilton, T. Introns regulate the rate of unstable mRNA decay. *J. Biol. Chem.* **282**, 20230–20237 (2007).

38. Clark, M. B. *et al.* Genome-wide analysis of long noncoding RNA stability. *Genome Res.* **22**, 885–898 (2012).

39. Sibley, C. R. *et al.* Recursive splicing in long vertebrate genes. *Nature* **521**, 371–375 (2015).

40. Almada, A. E., Wu, X., Kriz, A. J., Burge, C. B. & Sharp, P. A. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**, 360–363 (2013).

41. Ntini, E. *et al.* Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. *Nat. Struct. Mol. Biol.* **20**, 923–928 (2013).

42. Core, L. J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46**, 1311–1320 (2014).

43. Bailey, T. L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**, 1653–1659 (2011).

44. Vucic, E. A., Wilson, I. M., Campbell, J. M. & Lam, W. L. Methylation analysis by DNA immunoprecipitation (MeDIP). *Methods Mol. Biol.* **556**, 141–153 (2009).

45. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

46. Luco, R. F. *et al.* Regulation of alternative splicing by histone modifications. *Science* **327**, 996–1000 (2010).

47. Tilgner, H. *et al.* Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* **22**, 1616–1625 (2012).

48. Wu, X. & Sharp, P. A. Divergent transcription: a driving force for new gene origination? *Cell*

155, 990–996 (2013).

49. Urrutia, A. O. & Hurst, L. D. The signature of selection mediated by expression on human genes. *Genome Res.* **13**, 2260–2264 (2003).

50. Versteeg, R. *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.* **13**, 1998–2004 (2003).

51. Kudla, G., Lipinski, L., Caffin, F., Helwak, A. & Zylicz, M. High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol.* **4**, e180 (2006).

52. Sémon, M., Mouchiroud, D. & Duret, L. Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance. *Hum. Mol. Genet.* **14**, 421–427 (2005).

53. Arhondakis, S., Clay, O. & Bernardi, G. GC level and expression of human coding sequences. *Biochem. Biophys. Res. Commun.* **367**, 542–545 (2008).

54. Young, J. I. *et al.* Regulation of RNA splicing by the methylation-dependent transcriptional repressor methyl-CpG binding protein 2. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 17551–17558 (2005).

55. Chu, T. *et al.* Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme. *Nat. Genet.* **50**, 1553–1564 (2018).

56. Mahat, D. B. *et al.* Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.* **11**, 1455–1476 (2016).

57. Danko, C. G. *et al.* Identification of active transcriptional regulatory elements from GRO-seq data. *Nat. Methods* **12**, 433–438 (2015).

58. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281–285 (2012).

59. Ho, D. E., Imai, K., King, G. & Stuart, E. A. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis* **15**, 199–

236 (2007).

60. Ho, D. E., Imai, K., King, G. & Stuart, E. A. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software* **42**, (2011).

61. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).

62. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).

63. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

64. Akalin, A., Franke, V., Vlahoviček, K., Mason, C. E. & Schübeler, D. Genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics* **31**, 1127–1129 (2015).

65. Agarwal, V., Bell, G. W., Nam, J.-W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **4**, (2015).