6414–6423 *Nucleic Acids Research, 2012, Vol. 40, No. 14* *Published online 10 April 2012*
doi:10.1093/nar/gks304

Computational inference of mRNA stability from histone modification and transcriptome profiles

Chengyang Wang[1], Rui Tian[1], Qian Zhao[1], Han Xu[2], Clifford A. Meyer[2], Cheng Li[2], Yong Zhang[1] and X. Shirley Liu[2,*]

[1]Department of Bioinformatics, School of Life Science and Technology, Tongji University, 1239 Siping Road, Shanghai 20092, China and [2]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard school of Public Health, 450 Brookline Ave, Boston, MA 02215, USA

Received February 8, 2012; Revised March 20, 2012; Accepted March 22, 2012

ABSTRACT

Histone modifications play important roles in regulating eukaryotic gene expression and have been used to model expression levels. Here, we present a regression model to systematically infer mRNA stability by comparing transcriptome profiles with ChIP-seq of H3K4me3, H3K27me3 and H3K36me3. The results from multiple human and mouse cell lines show that the inferred unstable mRNAs have significantly longer 3′Untranslated Regions (UTRs) and more microRNA binding sites within 3′UTR than the inferred stable mRNAs. Regression residuals derived from RNA-seq, but not from GRO-seq, are highly correlated with the half-lives measured by pulse-labeling experiments, supporting the rationale of our inference. Whereas, the functions enriched in the inferred stable and unstable mRNAs are consistent with those from pulse-labeling experiments, we found the unstable mRNAs have higher cell-type specificity under functional constraint. We conclude that the systematical use of histone modifications can differentiate non-expressed mRNAs from unstable mRNAs, and distinguish stable mRNAs from highly expressed ones. In summary, we represent the first computational model of mRNA stability inference that compares transcriptome and epigenome profiles, and provides an alternative strategy for directing experimental measurements.

INTRODUCTION

In eukaryotic cells, DNA winds around histone octamers to form the basic chromatin structure. Combinatorial modifications of the histone N-terminals, such as acetylation, methylation and phosphorylation, are related to distinct chromatin states (1–4). Histone modifications are also implicated in a wide range of biological processes, especially with transcriptional gene regulation (2,5,6). Transcription starts with the pre-initiation complex formation, then proceeds to a dynamic cycle of Pol II initiation, elongation and termination (5,7,8). All these events are closely associated with different histone modifications. For example, H3K4me3 is associated in transcription initiation (9), H3K36me3 with transcription elongation (10,11) and H3K27me3 with RNA polymerase pausing and elongation repression (12–14). Although it is unclear whether histone modifications are the cause or consequence of transcription (15), they do play a key role in regulating gene expression (1,16–19).

A number of computational methods have been proposed to model histone modification profiles. These methods have been used to identify differential histone modification sites (20), find key transcription factors (21,22), predict tissue-specific gene regulation (23) and infer relationships among different histone modifications (24). Most notably, quantitative models have been developed to show that histone modifications are predictive of gene expression levels (25–27). Karlic *et al.* (27) proposed a linear regression model that successfully demonstrated the predictive power of histone marks with respect to gene expression. Cheng *et al.* (25) constructed support vector regression models to integrate histone modifications and reported that histone modifications and transcription factors are statistically redundant for predicting gene expression levels (26). These studies focused on the predictive power of their models but largely ignored the deviations between the model predictions and the mRNA levels.

Steady-state mRNA levels, as assessed by microarrays or RNA-seq, represent the dynamic balance between transcript production and degradation at specific cell conditions. We assume that, in a non-stimulus environment, statistical models of histone modification levels

*To whom correspondence should be addressed. Tel: +1 617 632 3012/+1 617 632 3498; Fax: +1 617 632 2444; Email: xsliu@jimmy.harvard.edu

could represent a stable transcription rate, and that the mRNA degradation rate is independent of the transcription rate. Under this assumption, the difference between the mRNA levels predicted from histone modifications and the mRNA levels measured from microarrays or RNA-seq could be used to infer RNA stability. The mRNAs that are more abundant than predicted from histone marks are inferred to be stable mRNAs with low degradation rates whereas those that are less abundant than predicted are inferred to be unstable mRNAs with high degradation rates.

In this study, we proposed, for the first time, that the computational models systematically comparing epigenome and transcriptome profiles could be used to infer mRNA stability. For the epigenome, we used ChIP-seq data of three widely profiled histone methylations, H3K4me3, H3K27me3 and H3K36me3, all of which have well-defined roles in transcription. We selected specific features to optimize the linear models of the histone marks to fit the transcriptome profiles in multiple human (ENCODE: K562, GM12878, HepG2, HSMM, Huvec, NHEK) and mouse (MEF and NPC) cell lines. Using residuals between the mRNA levels predicted from histone marks and the levels measured by microarrays or RNA-seq, we inferred the mRNA stability in each cell line. We investigated the sequence features, functional characteristics and cell-type specificity of the stable and unstable mRNAs. Our computational inference yielded very consistent results with the mRNA half-life measurements from pulse-labeling experiments, supporting this method as a cost-effective alternative strategy.

MATERIALS AND METHODS

Data sources

The RefSeq gene annotations for hg19 and mm9 were downloaded from the UCSC genome browser (http://genome.ucsc.edu). The gene expression and ChIP-seq data were downloaded from the NCBI Gene Expression Omnibus database. The expression data include microarray data for the ENCODE cell lines K562, GM12878, HepG2, Huvec, HSMM and NHEK [Accession designation: GSE26312 (28)], RNA-seq and GRO-seq data for MEFs [GSE27843 (29), GSE27037 (30)] and RNA-seq data for NPCs [GSE20851 (31)]. The binding data for the histone modifications include ChIP-seq data for the ENCODE cells [GSE26320 (32)], mouse embryonic fibroblasts (MEFs) [GSE12241 (12)] and Neural progenitor cell (NPCs) [GSE16256 (33)]. The averaged half-life data of 5029 transcripts in NIH3T3 cells are from the Supplementary data of *Global quantification of mammalian gene expression control* (34).

Data pre-processing

The microarray data were processed by the limma package in the R program (35). The fastq files for the RNA-seq data were mapped back to the corresponding genome using hg19 or mm9 with tophat v1.1.4 (36). The FPKMs (fragments per kilobase of exon per million mappable fragments) were calculated as the expression level by cufflinks (37). The fastq files for the ChIP-seq data were mapped to the human (hg19) or mouse (mm9) genome by bowtie (38). A software to generate genome-wide mapped reads intensity and call peaks (MACS) was used to generate wiggle files (39).

Defining the histone modification features for linear regression

We defined 15 regions that cover 5 kb upstream of the transcription start sites (TSSs) to 1 kb downstream of the transcription termination sites (TTSs) for each gene, including up to 1000, up to 2000, up to 5000 (1000/2000/5000-bp regions upstream of the TSSs), TSS 1000, 2000, 2500 (2000, 4000 or 5000 bp centered at the TSSs), exon, gene-body (downstream of the TSS to upstream of the TTS), TTS region 30%, TTS region 50%, TTS region 70% (30, 50 and 70% of the length of the gene body upstream of the TTS), TTS 30%, TTS 50%, TTS 70% (0.3, 0.5, 0.7 % of the exons closest to the TTS), and TTS1000 (2000-bp regions centered at the TTSs).

For each transcript, the read coverage of each histone modification in the 15 regions (read count per bp) was calculated and normalized according to the sequencing library sizes. To reflect the relative contribution of each histone modification in our model, each feature was centered and scaled to have a mean of 0 and a standard deviation of 1 in multiple transcripts. We take log2 (FPKM+1) as the expression index measured by RNA-seq so that the microarray and RNA-seq data are comparable. In total, 20 421 transcripts from ENCODE cell lines and 26 400 transcripts from MEFs and NPCs were used to build the linear models.

Studentized residuals

In our regression model:

$$mRNA\ level\ \sim b_0+b_1N_{H3K4me3}+b_2N_{H3K27me3}+b_3N_{H3K36me3}+e$$

where the errors e are assumed to independently follow a normal distribution $N(0, \sigma^2)$. The residuals ê, unlike the errors, are the deviations between the model predictions and the mRNA levels. Under the assumption of the distribution of errors, the residuals may have different variances. To reasonably compare the residuals among multiple mRNAs, the residuals should be normalized to studentized residuals.

Prediction of microRNA-binding sites

The predicted microRNA binding sites were downloaded from TargetScanHuman (http://www.targetscan.org/vert_61/) and TargetScanMouse (http://www.targetscan.org/mmu_61/) (40). The TargetScan database includes both conserved and non-conserved targeting sites within the 3′UTRs.

Gene ontology analysis

The DAVID functional annotation tool (http://david.abcc.ncifcrf.gov/) was used to analyze the gene function enrichment (41). The *P*-value cutoff was set as 0.01.

RESULTS

Histone modifications are predictive of mRNA levels in multiple cell lines

We first investigated the quantitative relationship between histone modification and expression profiles in eight different human and mouse cell lines with consistent data format. The expression data are measures of exon arrays for six human ENCODE cell lines and RNA-seq for two mouse cell lines. The histone mark profiles include ChIP-seq of H3K4me3, H3K36me3 and H3K27me3, which are the most frequently profiled in published studies. To assign histone modification signals to genes, we used ChIP-seq reads from 5 kb upstream of TSSs to 1 kb downstream of TTSs. To identify the distinct histone mark features that are predictive of mRNA levels, we defined 15 features across each gene for each histone mark, such as read coverage in promoters, exons or tails of gene bodies (details in section 'Materials and Methods'). We fitted the mRNA level as a linear combination of individual histone mark features:

$$mRNA\ level\ \sim b_0+b_1 N_{H3K4me3}+N_{H3K27me3}+N_{H3K36me3}+e$$

where N represents the normalized read coverage and e indicates error, which is assumed to independently follow a normal distribution.

We tested all of the possible triple combinations of the three histone mark features and used the Bayesian information criterion (BIC) to evaluate the performance of each model. In the K562 cell line, the model with H3K4me3 reads in TSS1000 (defined as the 2000 bps centered at the TSSs), H3K36me3 in gene bodies and H3K27me3 in exons yielded the lowest and optimal BIC score. Our computational model with only three histone mark features fits very well with the mRNA levels in the K562 cell line, as measured by Pearson correlation (Figure 1A) and regression P-value (<2.2 e-16). The regression coefficients in the model (Figure 1B) indicate that H3K4me3 and H3K36me3 contribute most to the prediction model, consistent with its role in transcriptional initiation and elongation. We found that including non-linear terms in the model did not significantly increase the R-square fitting or change any of the downstream conclusions, so we kept the simple linear model for efficiency. The same method was applied to the other cell lines, and each produced a high correlation between the expression levels predicted from the histone marks and the mRNA levels (Supplementary Figure S1). The relative contributions of the individual histone marks to the regression model are also similar in different cells (Supplementary Figure S2).

Deviations between model predictions and mRNA levels can infer mRNA stability

To investigate the fitness of our model, we examined the studentized residuals to identify mRNAs whose expression levels significantly deviate from the model predictions (details in section 'Materials and Methods', Figure 2A). For example, DNAJC24 and RERE exhibited similar profiles for all the three histone marks, yet their expression levels measured by microarrays differed by two orders of

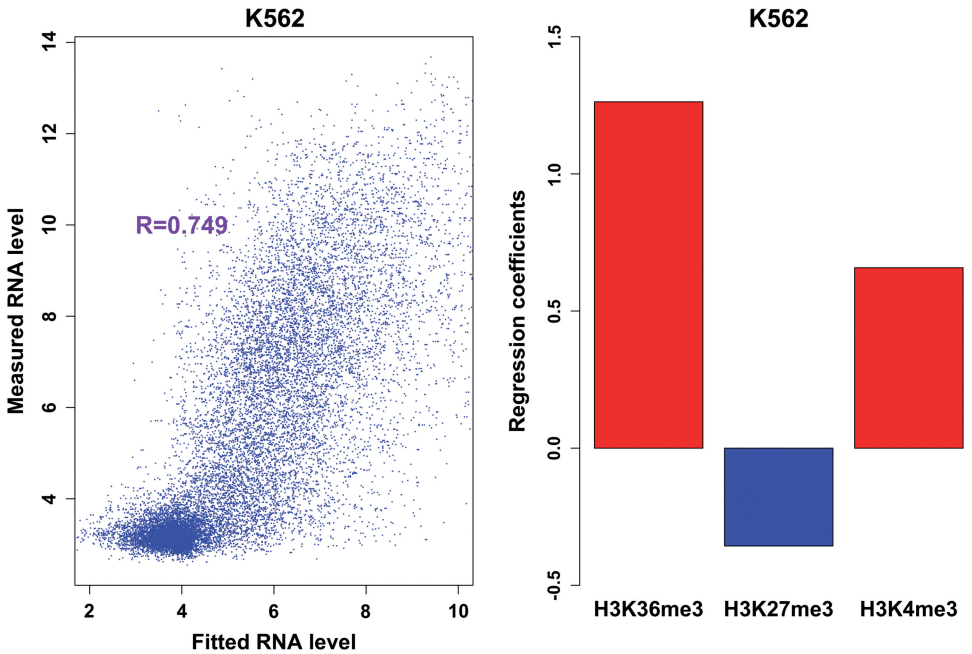


Figure 1. Fitting level and regression coefficients. (**A**) The scatter plot of the measured mRNA levels against the fitted RNA levels of the optimal model with the least BIC score in K562 cells. Each dot represents a transcript involved in our model. (**B**) The regression coefficients for H3K36me3, H3K27me3 and H3K4me3. The corresponding histone modification features are gene-body, exon and TSS1000 (1000 bp centered on the TSS), respectively.

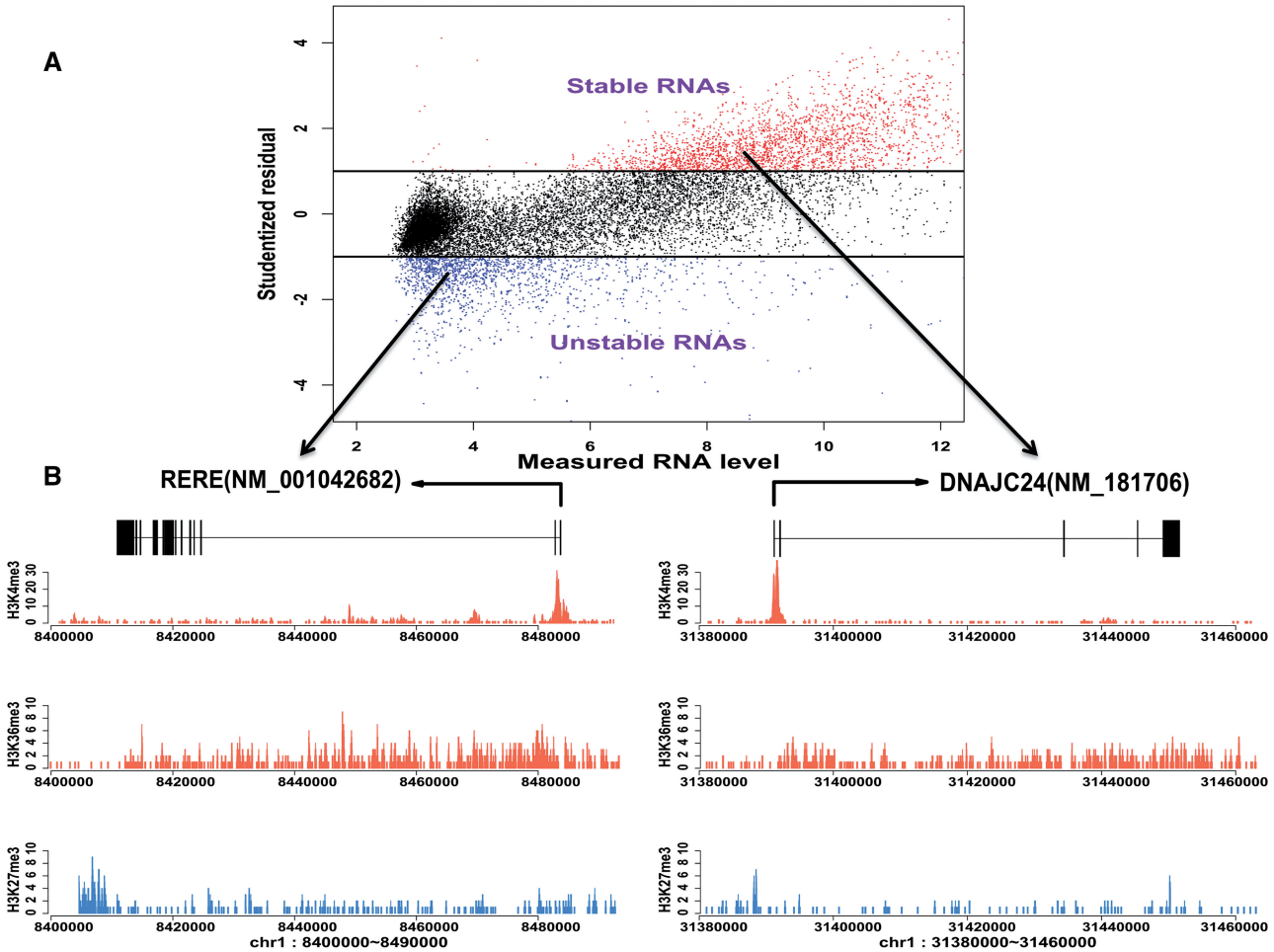


Figure 2. Definition of inferred stable RNAs and inferred unstable RNAs. (**A**) The mRNAs that are more highly expressed relative to the prediction values (studentized residual >1) are defined as stable mRNAs and shown as red dots. Conversely, the blue dots represent the mRNAs with studentized residual <−1, referred to as unstable. (**B**) Two randomly selected transcripts, NM_001042682 and NM_181706, separately from unstable RNAs and stable RNAs. The distributions of H3K36me3, H3K27me3 and H3K4me3 over these two genes are visualized below. The *y*-axes represent the read counts. Note that these two genes are quite similar in overall histone modification patterns and are predicted to have similar mRNA levels in the regression model (NM_001042682: 6.20; NM_181706: 5.95). Nevertheless, their real mRNA levels, as measured by microarray, are significantly different (NM_001042682: 3.58; NM_181706: 8.68).

magnitude (Figure 2B). Because our model predicts expression from histone marks, it resulted in a high positive residual for DNAJC24 and a high negative residual for RERE.

Previous studies have shown that the three histone modifications used in our model are tightly associated with the transcription cycle (9–14). Our computational model based on these histone marks is statistically correlated with mRNA levels. Therefore, we speculate that the model prediction somewhat represents a stable transcription level. In contrast, the mRNA level, as measured by microarrays or RNA-seq, reflects a balance between transcript production and degradation. Assuming that the degradation rate is independent of the transcription rate, we hypothesize that mRNAs with high negative and positive studentized residuals represent those that are degraded faster (defined as the unstable group, with residual <−1) and slower (defined as the stable group, with residual >1), respectively (Figure 2A).

Inferred unstable mRNAs harbor sequence signatures associated with more microRNA targeting

It is well known that microRNAs, a type of endogenous ~22-nt RNAs, play an important role in posttranscriptional gene regulation (42). These RNAs direct mRNA degradation, primarily by matching ~7 nt seeds to the 3′UTRs of target mRNAs (43). Although recent techniques, such as HITS-CLIP and PAR-CLIP, could accurately measure genome-wide microRNA target sites (44–48), such data are not available for the cell lines in our study (49). Because TargetScan is among the best and most widely used tools for the computational prediction of microRNA targets, we used its predictions to investigate the microRNA targeting of our inferred mRNAs (40).

Studies have shown that proliferating and cancer cells express mRNA isoforms with shorter 3′UTRs (50,51) and that mRNAs with longer 3′UTRs are more likely to be targeted by microRNAs (52). Indeed, we found that the inferred unstable mRNAs have significantly longer 3′UTRs

than the stable ones in all of the cell lines (Figure 3A–B and Supplementary Figure S3). In addition, the 3′UTRs of the unstable mRNAs harbor significantly larger numbers of all (conserved + non-conserved) microRNA-binding sites in multiple cell lines (Figure 3C–D and Supplementary Figure S4). Additionally, in the ENCODE cell lines, unstable mRNAs have larger numbers and a higher density of conserved microRNA binding sites within their 3′UTRs (density refers to the amount of binding sites divided by the 3′UTR length) than stable mRNAs (Supplementary Figures S5 and S6). These results suggest that the inferred unstable mRNAs are preferentially targeted by microRNAs.

Half-lives are highly correlated with residuals in an RNA-seq model but independent of those in a GRO-seq model

Recently, many studies (53–56) have used pulse labeling to examine the half-lives and degradation rates of mRNAs. In these studies, a pulse of radioactive nucleotides is used to label newly synthesized RNA. By measuring radioactive mRNA and total mRNA levels over a time course, researchers can accurately compute mRNA synthesis and degradation rates. Schwanhäusser B *et al.* (53) reported the half-life data for 5028 mRNAs in the mouse embryonic fibroblast cell line NIH3T3. Because ChIP-seq data for H3K4me3, H3K27me3 and H3K36me3 in NIH3T3 are not publicly available, we compared their results with our computationally predicted degradation rates from another mouse embryonic fibroblast cell line, MEFs.

In MEFs, the correlation between the mRNA levels measured by RNA-seq and our model predictions from the histone marks in MEFs is 0.7 for all of the transcripts. The reported half-lives (53) are significantly correlated with the residuals between the mRNA levels and the model predictions ($\rho = 0.322$, P-value $< 2.2e\text{-}16$, Figure 3E), supporting the rationale for our definitions of stable and unstable transcripts. When we incorporated the half-life information to re-derive a regression model for the 5028 mRNAs with half-life data, the model was able to better fit the mRNA levels ($R = 0.54$ for the model based only on histone modifications, $r = 0.62$ for the one based on both histone marks and half-lives).

To further validate our predictions of stable as compared with unstable transcripts, we examined the GRO-seq data from MEFs. GRO-seq is a global run-on experiment to measure nuclear nascent RNAs that are associated with transcriptionally engaged polymerases (57). We refitted the regression model using H3K4me3, H3K27me3 and H3K36me3 to the GRO-seq data in MEFs and examined the residual between GRO-seq and the model prediction. Because GRO-seq effectively

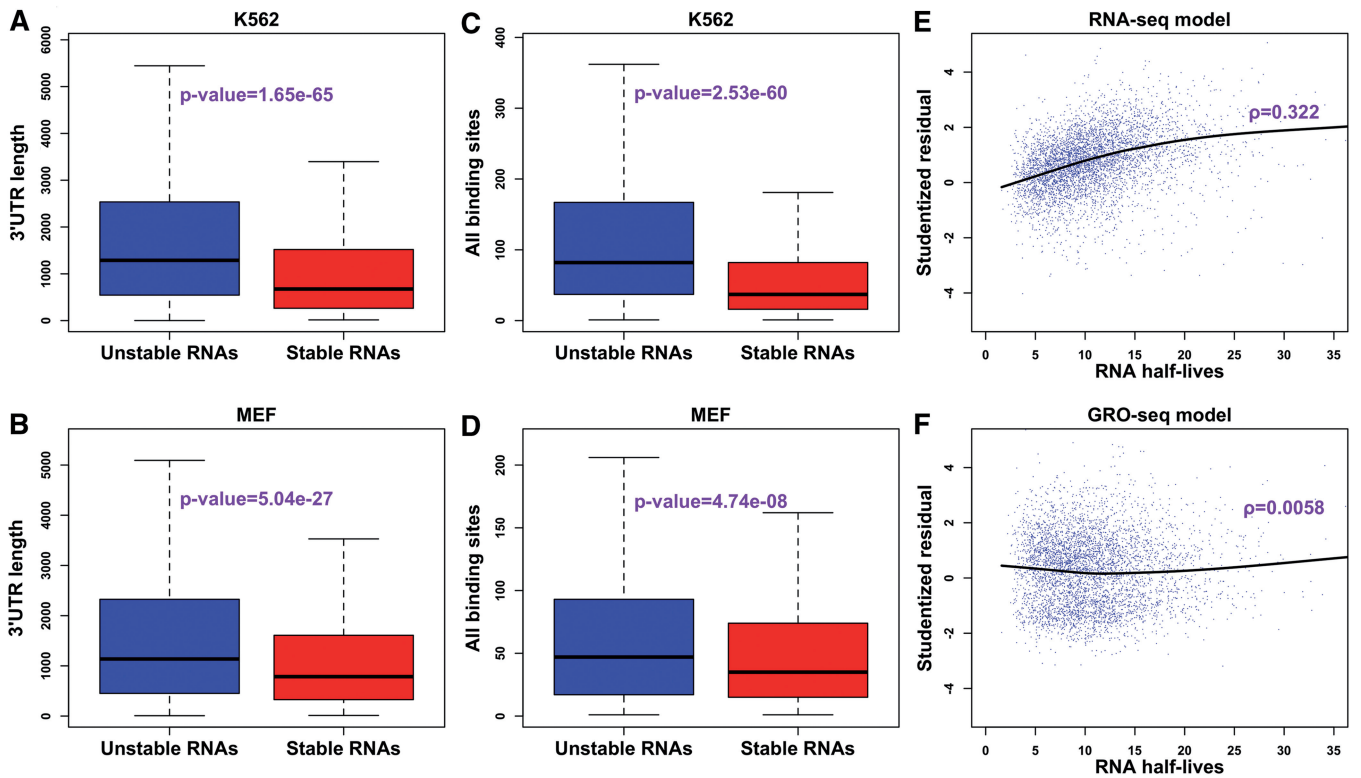


Figure 3. The sequence signatures and half-lives correlations of model inference. The boxplots in (**A–D**) represent the distribution of certain sequence signatures for specific groups (unstable or stable ones), and the P-values were calculated by the Wilcox-test. These boxplots indicate that the inferred unstable mRNAs have sequence signatures in favor of post-transcriptional degradation. (A–B) The comparison of 3′UTR length in K562 cells and in MEFs, separately. (C–D) The comparison of the number of all binding sites (conserved + non-conserved) within the 3′UTRs in K562 cells and in MEFs, separately. (**E**) The half-lives of mRNAs are positively correlated with studentized residuals in RNA-seq model, in which all of the transcripts are modeled on histone marks to fit the mRNA levels measured by RNA-seq. ρ refers to Pearson Correlation coefficients. (**F**) The half-lives are little correlated with the studentized residuals in a GRO-seq model, in which histone modifications are fitted to nascent mRNA levels.

measures nascent mRNA production, the residues derived from GRO-seq should be independent of the mRNA degradation rates and half-lives. Indeed, we found little correlation between the GRO-seq model residuals and the half-lives reported by pulse-labeling experiments ($\rho = 0.0058$, Figure 3F). This observation further validates the idea that the residuals can be used to infer transcript stability, which is independent of the transcription rate.

Inferred unstable and stable mRNAs are enriched for distinct biological processes

To investigate whether genes with varying mRNA degradation rates are enriched for specific biological processes or functions, we conducted a Gene Ontology (GO) analysis for the inferred stable and unstable mRNAs. We found that the inferred unstable mRNAs are consistently enriched in transcription, regulation of RNA metabolism and chromatin modification in all the cell lines examined (Figure 4A, the complete list is shown on Supplementary Table S1). This finding is in agreement with previous studies that identified many transcription factors (e.g. *Klf7, Dmtf1*), especially those targeted by microRNAs (e.g. *Foxo1, Hif1a, p53*), to have short mRNA half-lives (54).

In contrast, the inferred stable mRNAs are also consistently enriched in constitutive cellular processes in all of the cell lines. These processes include the generation of precursor metabolites and energy, translation, oxidation-reduction, oxidative phosphorylation, cellular respiration and the electron transport chain (Figure 4B, the complete list is shown in Supplementary Table S2). Translation and protein synthesis have been implicated to account for >90% of cellular energy consumption (53), and electron transport chains are also known to maximize electron flux and minimize energy expenditure (58). These findings suggest that the stable mRNAs are all enriched in constitutive cellular processes that are associated with energy. Notably, our GO analysis results for both stable and unstable mRNAs are very consistent with those from pulse-labeling experiments (53), which further validates our computational inference.

Inferred unstable mRNAs have higher cell-type specificity under functional constraint

To investigate the consistency of the stable and unstable mRNAs in different cell conditions, we compared the mRNAs with the inferred stability from different ENCODE cell lines that fall in the same GO terms and calculated their overlaps. We found that the unstable mRNAs under each enriched GO term significantly overlap among different cell lines (pairwise overlaps >50%, hypergeometric test, *P*-values <2.2e-16). In addition, the stable mRNAs also show significant pairwise overlap, and the overlap level is much higher than that in the unstable mRNAs (Figure 5). This difference might arise from the cell-type-specific expression of microRNAs and RNA-binding proteins involved in RNA degradation. These results suggest that, whereas the functional processes of stable and unstable mRNAs are consistent among different cell types, the unstable mRNAs have much higher cell-type specificity.

Histone modifications are informative for inferring mRNA stability

Because our model residuals have a positive correlation with the measured mRNA levels, as shown in Figure 2A, one might question of whether the inferred mRNA

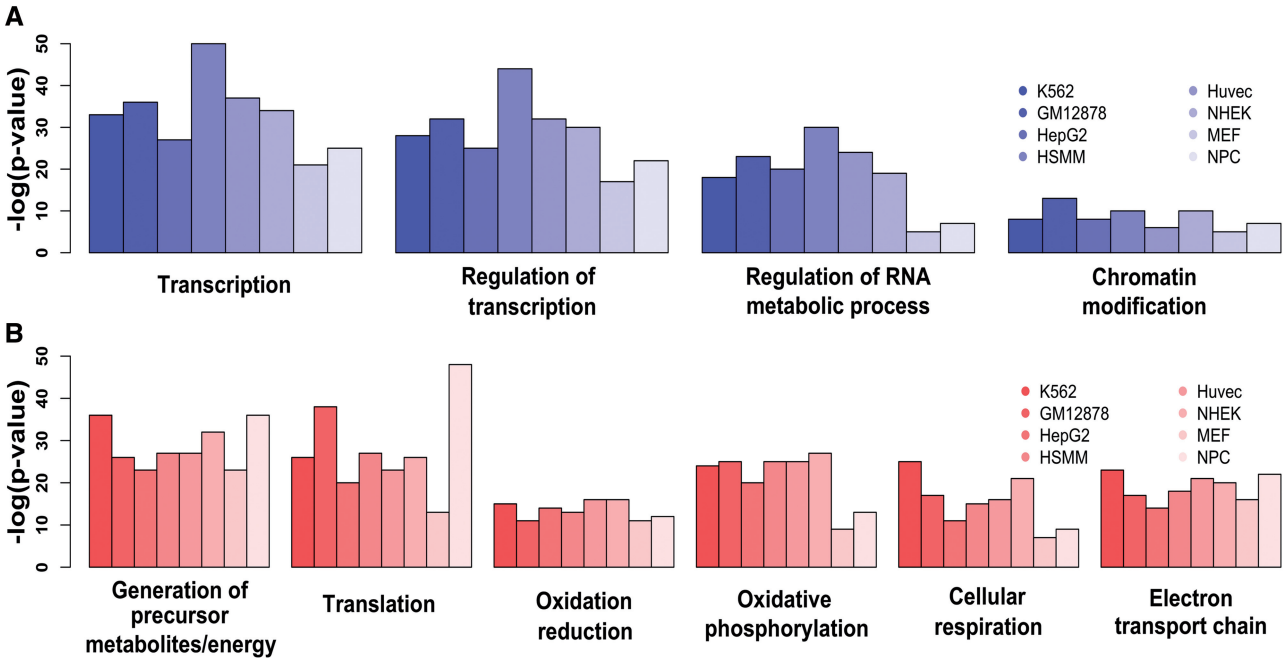


Figure 4. Differentially Enriched GO Terms for unstable mRNAs (**A**) and stable mRNAs (**B**) in multiple ENCODE and mouse cell lines. The *y*-axis represents a negative logarithmic scale of the *P*-values, so the higher they are, the more significantly the corresponding GO term is enriched.

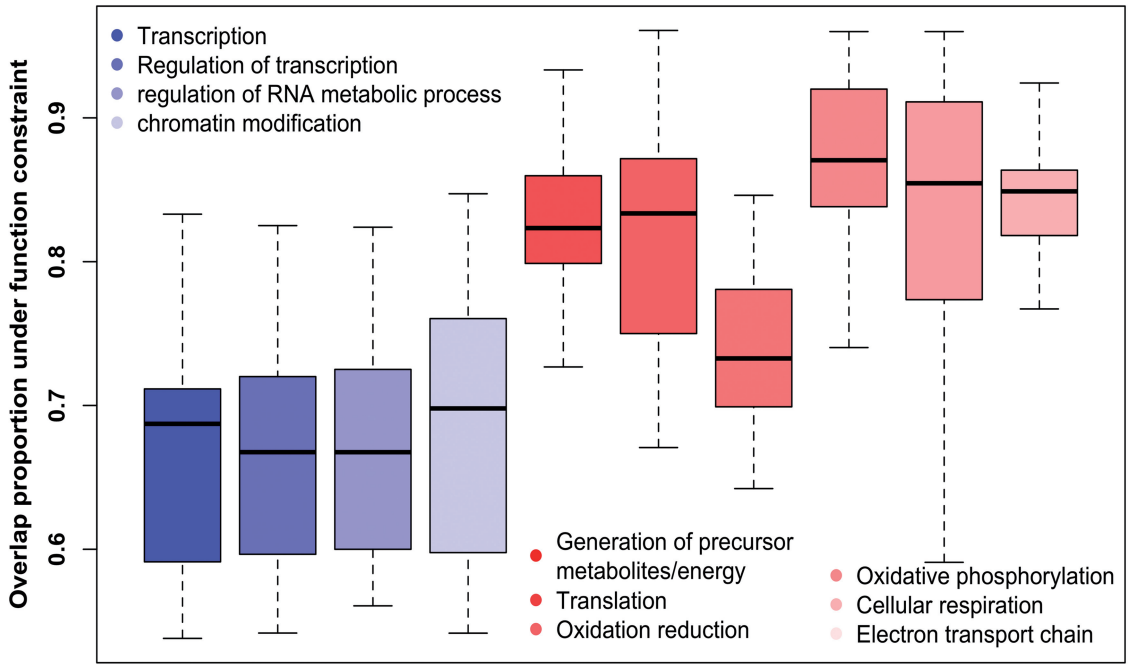


Figure 5. Unstable mRNAs have higher cell-type specificity than stable ones under functional constraint. The mRNAs with inferred stability from different ENCODE cells that fall onto the same GO terms are pairwise compared, and the overlapped proportions were calculated in each term separately (each overlap corresponds to two proportions based on a pair of compared sets). The unstable gene sets are shown in blue and the stable gene sets are shown in red. The boxplot represents the distribution of the pairwise overlap proportions under the GO term.

stability arises solely from gene expression levels. To assess this possibility, we defined four zones in the residual plot of MEFs (Figure 6A). Zone A accounts for 44% of the total 26 400 expressed transcripts in MEFs and is significantly enriched in functions with unclear half-lives (Figure 6A, top left). In fact, most of the mRNAs (10 859/11 572) in Zone A are not expressed at all (the GRO-seq FPKMs are equal to 0).

Compared with Zone A, the mRNAs in Zone B are expressed at similarly low levels. Nevertheless, significantly larger numbers of mRNAs in Zone B are actually expressed (the GRO-seq FPKMs of 53.81% of the transcripts in Zone B are not equal to 0 versus 6.16% for Zone A, two proportion z-test, *P*-value <2.2e-16). In addition, the functional enrichments from Zone B are consistent with those from the unstable transcripts obtained from pulse-labeling experiments (Figure 6A, bottom left). In fact, a large proportion of genes are expressed at very low levels in most of the cells, e.g. the RNA-seq levels of 55% of the mRNAs are less than one FPKM in MEFs. Our analysis suggests that our computational model from histone marks could differentiate two distinct classes of genes with low expression levels: the majority is transcriptionally silenced (Zone A), and the remainders are efficiently transcribed but rapidly degraded (Zone B).

Similarly, histone modification levels are informative to distinguish the more stable mRNAs from the highly expressed ones. The mRNAs in Zone C are enriched for constitutive gene functions, whereas the mRNA sets identified as unstable from pulse-labeling experiments are significantly enriched in Zone D (Figure 6A). Furthermore, the mRNAs in Zone C have significantly longer half-lives than those in Zone D (*P*-value = 1.6e-86, Figure 6B). In summary, the inclusion of histone modifications can differentiate between silenced mRNAs and unstable mRNAs and distinguish stable mRNAs from highly expressed ones.

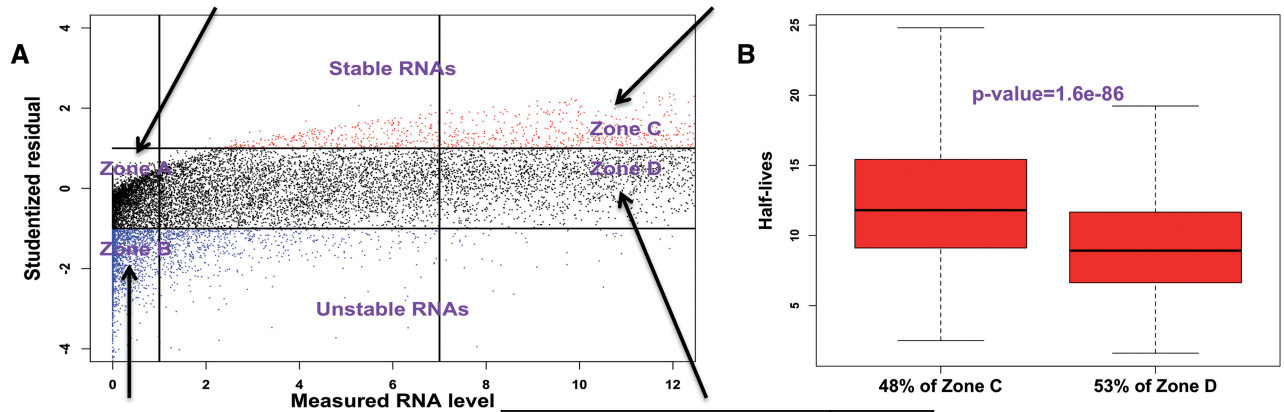
DISCUSSION

In this article, we report the first computational approach to systematically infer global mRNA stability on the basis of a comparison of mRNA levels and histone modification profiles. Three lines of evidence support our inference on mRNA stability. First, the inferred unstable mRNAs harbor sequence signatures associated with more microRNA targeting. Second, the half-lives are positively correlated with the residuals when our computational model is fitted to RNA-seq but independent of the residuals when fitted to GRO-seq. Third, functional annotations of the enriched gene sets produced consistent results with previous reports based on pulse-labeling experiments. Our analysis conducted on multiple human and a mouse cell lines suggests that unstable mRNAs have higher cell-type specificity than stable ones. Finally, we found histone modification levels can distinguish unstable mRNAs from silenced ones and differentiate stable mRNAs from highly expressed ones.

Histone modifications are implicated in transcriptional regulation in eukaryotic cells, and a number of reports have illustrated the predictive power of integrating multiple histone modifications on gene expression (25–27). Our study showed that carefully selected features combining only three histone modifications, H3K4me3, H3K27me3

G-protein coupled receptor protein signaling pathway	E-300
Sensory perception	E-300
Neurological system process	E-298
Cell surface receptor linked signal transduction	E-300
Ion transport	E-25
Cell cell signaling	E-10

Generation of precursor metabolites / energy	E-33
Protein localization	E-21
Electron transport chain	E-18
Translation	E-15
Oxidation reduction	E-12
Cellular respiration	E-8



Transcription	E-11
Regulation of transcription	E-8
DNA metabolic process	E-7
Chromatin organization	E-5

RNA processing	E-56
RNA splicing	E-38
Cell cycle	E-26
Translation	E-11

Red : stable
Blue: unstable
Green: neutral
Black : unidentified

Figure 6. Histone modification levels are informative for inferring mRNA stability. (**A**) Zones A–D are defined in the middle plot according to the measured RNA level (FPKM cutoff of the mRNAs with low expression is 1, whereas that of the highly expressed mRNAs is 7) and the studentized residual (cutoff is −1 and 1). Zone A refers to mRNAs with lower abundances and absolute value of studentized residuals <1. Those with lower abundances and residuals <−1 constitute Zone B. Zone C is defined as highly expressed RNAs with residuals >1, whereas Zone D is composed of those highly expressed mRNAs with absolute values of studentized residuals <1. The four tables contain GO significantly enriched for Zones A–D, respectively. The red color denotes that mRNAs involved in the GO term tend to have longer half-lives, whereas the blue color indicates shorter half-lives. Green denotes no tendency on half-lives, and black means that the information about half-lives is unclear. All the half-life data are from Schwanhäusser B *et al.* (**B**) Comparing mRNA half-lives between Zone C and Zone D. The half-life data are available for 48% of the mRNAs in Zone C and 53% in Zone D, separately.

and H3K36me3, could already explain 40–50% of the mRNA expression variance. We chose these three modifications because they are the most widely profiled and have been shown to be the most informative for gene expression prediction (26). As genome-wide histone modification profiles accumulate over time (59,60), we could improve our model by integrating more histone modifications. It is worth noting that in this study we evaluate significance mostly in the statistical sense, and biological significance needs to be confirmed by further experimental investigations.

Our computational model relies on two important assumptions, both of which depend on steady-state conditions. The first assumption is that histone modification is reflective of the steady-state transcription rate. In MCF7 cells, E2 treatment has significant transcriptional effect in only 10 min (61). Thus, transient transcriptional changes in response to outside stimuli may be faster than changes of the three histone marks we selected. The second assumption is that mRNA transcription rates and degradation rates are independent. This assumption is supported by the observation that the transcription rate measured by GRO-seq has little correlation (0.0058) with the half-lives measured by pulse-labeling experiments. Upon cell differentiation or environmental stimulation, the transcription rate could be coupled with the degradation rate (56,62–64). Further studies are needed to refine the computational models for better prediction of degradation during non-steady-state conditions.

Steady-state mRNA levels represent a balance between transcription and degradation. Although our analysis revealed that unstable mRNAs are significantly more likely to be targeted by microRNAs, incorporating microRNA binding sites or AU-rich elements within the 3′UTRs (65) only marginally improved the fitness of our model. For mRNAs with available half-life data, the most informative sequence feature (the number of all microRNA binding sites within the 3′UTR) merely explains 1% of the

variance of the residuals whereas half-lives could explain 13% of the variance. This finding suggests that other factors, such as RNA-binding proteins or the secondary structure of the 3′UTR, may also be involved in regulating mRNA stability and decay (65–67).

In summary, we propose the first computational method for inferring mRNA stability by comparing transcriptome and histone modification profiles. As histone mark ChIP-seq data continue to grow, our approach provides a cost-effective alternative to the direct measurement of RNA stability by pulse-labeling experiments (53–56) or transcriptional inhibition (62,68,69).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–2 and Supplementary Figures 1–6.

ACKNOWLEDGEMENTS

We thank Jianxing Feng, Ying Ge, Chao Chen, Qixuan Wang, Lin Wang, Yiqian Zhang and Wei Li for their helpful discussions about the project.

FUNDING

National Basic Research [973] Program of China No. [2010CB944904], National Natural Science Foundation of China NO. [31028011] and NIH grant [HG4069]. Funding for open access charge: National Basic Research [973] Program of China No. [2010CB944904], National Natural Science Foundation of China No. [31028011] and NIH grant [HG4069].

Conflict of interest statement. None declared.

REFERENCES

1. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
2. Li,B., Carey,M. and Workman,J.L. (2007) The role of chromatin during transcription. *Cell*, **128**, 707–719.
3. Heintzman,N.D., Stuart,R.K., Hon,G., Fu,Y., Ching,C.W., Hawkins,R.D., Barrera,L.O., Van Calcar,S., Qu,C., Ching,K.A. *et al*. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet*., **39**, 311–318.
4. Jenuwein,T. and Allis,C.D. (2001) Translating the histone code. *Science*, **293**, 1074–1080.
5. Berger,S.L. (2007) The complex language of chromatin regulation during transcription. *Nature*, **447**, 407–412.
6. Khan,A.U. and Krishnamurthy,S. (2005) Histone modifications as key regulators of transcription. *Front Biosci*, **10**, 866–872.
7. Egloff,S. and Murphy,S. (2008) Cracking the RNA polymerase II CTD code. *Trends Genet*., **24**, 280–288.
8. Svejstrup,J.Q. (2004) The RNA polymerase II transcription cycle: cycling through chromatin. *Biochim. Biophys. Acta*, **1677**, 64–73.
9. Guenther,M.G., Levine,S.S., Boyer,L.A., Jaenisch,R. and Young,R.A. (2007) A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, **130**, 77–88.
10. Kolasinska-Zwierz,P., Down,T., Latorre,I., Liu,T., Liu,X.S. and Ahringer,J. (2009) Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat. Genet*., **41**, 376–381.
11. Krogan,N.J., Kim,M., Tong,A., Golshani,A., Cagney,G., Canadien,V., Richards,D.P., Beattie,B.K., Emili,A., Boone,C. *et al*. (2003) Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II. *Mol. Cell. Biol*., **23**, 4207–4218.
12. Mikkelsen,T.S., Ku,M., Jaffe,D.B., Issac,B., Lieberman,E., Giannoukos,G., Alvarez,P., Brockman,W., Kim,T.K., Koche,R.P. *et al*. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.
13. Boyer,L.A., Plath,K., Zeitlinger,J., Brambrink,T., Medeiros,L.A., Lee,T.I., Levine,S.S., Wernig,M., Tajonar,A., Ray,M.K. *et al*. (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*, **441**, 349–353.
14. Zhou,W., Zhu,P., Wang,J., Pascual,G., Ohgi,K.A., Lozach,J., Glass,C.K. and Rosenfeld,M.G. (2008) Histone H2A monoubiquitination represses transcription by inhibiting RNA polymerase II transcriptional elongation. *Mol. Cell*, **29**, 69–80.
15. Ng,H.H., Robert,F., Young,R.A. and Struhl,K. (2003) Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Mol. Cell*, **11**, 709–719.
16. Schübeler,D., MacAlpine,D.M., Scalzo,D., Wirbelauer,C., Kooperberg,C., van Leeuwen,F., Gottschling,D.E., O'Neill,L.P., Turner,B.M., Delrow,J. *et al*. (2004) The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev*., **18**, 1263–1271.
17. Bernstein,B.E., Kamal,M., Lindblad-Toh,K., Bekiranov,S., Bailey,D.K., Huebert,D.J., McMahon,S., Karlsson,E.K., Kulbokas,E.J. 3rd, Gingeras,T.R. *et al*. (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, **120**, 169–181.
18. Kurdistani,S.K., Tavazoie,S. and Grunstein,M. (2004) Mapping global histone acetylation patterns to gene expression. *Cell*, **117**, 721–733.
19. Wang,Z., Zang,C., Rosenfeld,J.A., Schones,D.E., Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Peng,W., Zhang,M.Q. *et al*. (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet*., **40**, 897–903.
20. Xu,H. and Sung,W.K. (2012) Identifying differential histone modification sites from ChIP-seq data. *Methods Mol. Biol*., **802**, 293–303.
21. Tian,R., Feng,J., Cai,X. and Zhang,Y. (2012) Local chromatin dynamics of transcription factors imply cell-lineage specific functions during cellular differentiation. *Epigenetics*, **7**.
22. Mayr,C. and Bartel,D.P. (2009) Widespread shortening of 3′ UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, **138**, 673–684.
23. Zhang,Z. and Zhang,M.Q. (2011) Histone modification profiles are predictive for tissue/cell-type specific expression of both protein-coding and microRNA genes. *BMC Bioinformatics*, **12**, 155.
24. Yu,H., Zhu,S., Zhou,B., Xue,H. and Han,J.D.J. (2008) Inferring causal relationships among different histone modifications and gene expression. *Genome Res*., **18**, 1544.
25. Cheng,C., Yan,K.K., Yip,K.Y., Rozowsky,J., Alexander,R., Shou,C. and Gerstein,M. (2011) A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol*., **12**, R15.
26. Cheng,C. and Gerstein,M. (2012) Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res*., **40**, 553–568.
27. Karlić,R., Chung,H.R., Lasserre,J., Vlahovicek,K. and Vingron,M. (2010) Histone modification levels are predictive for gene expression. *Proc. Natl Acad. Sci. USA*, **107**, 2926–2931.
28. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al*. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
29. Lienert,F., Mohn,F., Tiwari,V.K., Baubec,T., Roloff,T.C., Gaidatzis,D., Stadler,M.B. and Schubeler,D. (2011) Genomic prevalence of heterochromatic H3K9me2 and transcription do not discriminate pluripotent from terminally differentiated cells. *PLOS Genet*., **7**, e1002090.

30. Min,I.M., Waterfall,J.J., Core,L.J., Munroe,R.J., Schimenti,J. and Lis,J.T. (2011) Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev.*, **25**, 742–754.
31. Guttman,M., Garber,M., Levin,J.Z., Donaghey,J., Robinson,J., Adiconis,X., Fan,L., Koziol,M.J., Gnirke,A., Nusbaum,C. *et al.* (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.*, **28**, 503–510.
32. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–U52.
33. Hawkins,R.D., Hon,G.C., Lee,L.K., Ngo,Q., Lister,R., Pelizzola,M., Edsall,L.E., Kuan,S., Luu,Y., Klugman,S. *et al.* (2010) Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell*, **6**, 479–491.
34. Schwanhäusser,B., Busse,D., Li,N., Dittmar,G., Schuchhardt,J., Wolf,J., Chen,W. and Selbach,M. (2011) Global quantification of mammalian gene expression control. *Nature*, **473**, 337–342.
35. Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article 3.
36. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
37. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
38. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
39. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
40. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
41. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, **4**, 44–57.
42. Bartel,D.P. (2004) MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
43. Inui,M., Martello,G. and Piccolo,S. (2010) MicroRNA control of signal transduction. *Nat. Rev. Mol. Cell. Biol.*, **11**, 252–263.
44. Licatalosi,D.D., Mele,A., Fak,J.J., Ule,J., Kayikci,M., Chi,S.W., Clark,T.A., Schweitzer,A.C., Blume,J.E., Wang,X. *et al.* (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, **456**, 464–469.
45. Chi,S.W., Zang,J.B., Mele,A. and Darnell,R.B. (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, **460**, 479–486.
46. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Ascano,M., Jungkamp,A.C., Munschauer,M. *et al.* (2010) PAR-CliP–a method to identify transcriptome-wide the binding sites of RNA binding proteins. *J Vis Exp.*, **41**, e2034.
47. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Ascano,M. Jr, Jungkamp,A.C., Munschauer,M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141.
48. Zisoulis,D.G., Lovci,M.T., Wilbert,M.L., Hutt,K.R., Liang,T.Y., Pasquinelli,A.E. and Yeo,G.W. (2010) Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans. *Nat. Struct. Mol. Biol.*, **17**, 173–179.
49. Yang,J.H., Li,J.H., Shao,P., Zhou,H., Chen,Y.Q. and Qu,L.H. (2011) starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. *Nucleic Acids Res.*, **39**, D202–D209.
50. Sandberg,R., Neilson,J.R., Sarma,A., Sharp,P.A. and Burge,C.B. (2008) Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science*, **320**, 1643–1647.
51. Mayr,C. and Bartel,D.P. (2009) Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, **138**, 673–684.
52. Cheng,C., Bhardwaj,N. and Gerstein,M. (2009) The relationship between the evolution of microRNA targets and the length of their UTRs. *BMC Genomics*, **10**, 431.
53. Schwanhäusser,B., Busse,D., Li,N., Dittmar,G., Schuchhardt,J., Wolf,J., Chen,W. and Selbach,M. (2011) Global quantification of mammalian gene expression control. *Nature*, **473**, 337–342.
54. Rabani,M., Levin,J.Z., Fan,L., Adiconis,X., Raychowdhury,R., Garber,M., Gnirke,A., Nusbaum,C., Hacohen,N., Friedman,N. *et al.* (2011) Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat. Biotechnol.*, **29**, 436–442.
55. Dölken,L., Ruzsics,Z., Rädle,B., Friedel,C.C., Zimmer,R., Mages,J., Hoffmann,R., Dickinson,P., Forster,T., Ghazal,P. *et al.* (2008) High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA*, **14**, 1959–1972.
56. Amorim,M.J., Cotobal,C., Duncan,C. and Mata,J. (2010) Global coordination of transcriptional control and mRNA decay during cellular differentiation. *Mol. Syst. Biol.*, **6**, 380.
57. Core,L.J., Waterfall,J.J. and Lis,J.T. (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, **322**, 1845–1848.
58. Kurakin,A. (2011) The self-organizing fractal theory as a universal discovery method: the phenomenon of life. *Theor. Biol. Med. Model.*, **8**, 4.
59. Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigo,R., Gingeras,T.R., Margulies,E.H., Weng,Z.P., Snyder,M., Dermitzakis,E.T., Stamatoyannopoulos,J.A. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
60. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F., Ren,B., Milosavljevic,A., Meissner,A., Kellis,M., Marra,M.A., Beaudet,A.L., Ecker,J.R. *et al.* (2010) The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.*, **28**, 1045–1048.
61. Hah,N., Danko,C.G., Core,L., Waterfall,J.J., Siepel,A., Lis,J.T. and Kraus,W.L. (2011) A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell*, **145**, 622–634.
62. Shalem,O., Dahan,O., Levo,M., Martinez,M.R., Furman,I., Segal,E. and Pilpel,Y. (2008) Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation. *Mol. Syst. Biol.*, **4**, 223.
63. Elkon,R., Zlotorynski,E., Zeller,K.I. and Agami,R. (2010) Major role for mRNA stability in shaping the kinetics of gene induction. *BMC Genomics*, **11**, 259.
64. Shalem,O., Groisman,B., Choder,M., Dahan,O. and Pilpel,Y. (2011) Transcriptome kinetics is governed by a genome-wide coupling of mRNA production and degradation: a role for RNA Pol II. *Plos Genet.*, **7**, e1002273.
65. Hollams,E.M., Giles,K.M., Thomson,A.M. and Leedman,P.J. (2002) mRNA stability and the control of gene expression: implications for human disease. *Neurochem. Res.*, **27**, 957–980.
66. Cheng,Z.F. and Deutscher,M.P. (2005) An important role for RNase R in mRNA decay. *Mol. Cell*, **17**, 313–318.
67. Deutscher,M.P. (2006) Degradation of RNA in bacteria: comparison of mRNA and stable RNA. *Nucleic Acids Res.*, **34**, 659–666.
68. Raghavan,A., Ogilvie,R.L., Reilly,C., Abelson,M.L., Raghavan,S., Vasdewani,J., Krathwohl,M. and Bohjanen,P.R. (2002) Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes. *Nucleic Acids Res.*, **30**, 5529–5538.
69. Barenco,M., Brewer,D., Papouli,E., Tomescu,D., Callard,R., Stark,J. and Hubank,M. (2009) Dissection of a complex transcriptional response using genome-wide transcriptional modelling. *Mol. Syst. Biol.*, **5**, 327.