pulseTD: RNA life cycle dynamics analysis based on pulse model of 4sU-seq time course sequencing data

Xin Wang[1,*], Siyu He[1,*], Jian Li[1], Jun Wang[1], Chengyi Wang[1], Mingwei Wang[1], Danni He[1], Xingfeng Lv[2], Qiuyan Zhong[3], Hongjiu Wang[4,5,6] and Zhenzhen Wang[4,5]

[1] College of Bioinformatics Science and Technology, Harbin Medical University, Heilongjiang, China
[2] College of Computer Science and Technology, Heilongjiang University, Harbin, China
[3] Dalian University of Technology, Dalian, China
[4] Key Laboratory of Tropical Translational Medicine of Ministry of Education, Hainan Medical University, Haikou, China
[5] School of Biomedical Information and Engineering, Hainan Medical University, Haikou, China
[6] College of Science, Heilongjiang University of Science and Technology, Harbin, China
* These authors contributed equally to this work.

ABSTRACT

The life cycle of intracellular RNA mainly involves transcriptional production, splicing maturation and degradation processes. Their dynamic changes are termed as RNA life cycle dynamics (RLCD). It is still challenging for the accurate and robust identification of RLCD under unknow the functional form of RLCD. By using the pulse model, we developed an R package named pulseTD to identify RLCD by integrating 4sU-seq and RNA-seq data, and it provides flexible functions to capture continuous changes in RCLD rates. More importantly, it also can predict the trend of RNA transcription and expression changes in future time points. The pulseTD shows better accuracy and robustness than some other methods, and it is available on the GitHub repository (https://github.com/bioWzz/pulseTD_0.2.0).

Subjects Bioinformatics, Computational Biology
Keywords 4sU-seq, RNA-seq, Pulse model, RNA life cycle dynamics

INTRODUCTION

The response of cell stimulation is mainly manifested in the transcription, processing and degradation levels of RNA (*Rabani et al., 2011*; *Thapar & Denmon, 2013*). The dynamic combination of these processes is known as RNA life cycle dynamics (RLCD), which controls the gene expression and steady state of the cells (*Zeisel et al., 2011*). The cells continuously produce new RNA (pre-RNA), which will be processed into mature RNA (mRNA), and the mRNA is continuously degraded. The balance among transcription, processing and degradation rates keep cells in a steady state. However, external environmental interference, the changes in cell signal transduction and the activity transcription factor may change the rates of transcription, processing and degradation, which destroy the balance of RCLD in some cells. After a period of adjustment, cell volume growth, protein activity changes, etc., the rates of transcription, processing and

Submitted 27 September 2019
Accepted 27 May 2020
Published 8 July 2020

Corresponding authors
Hongjiu Wang,
yourfriend9@sohu.com
Zhenzhen Wang,
wangzz@ems.hrbmu.edu.cn

Academic editor
Elena Papaleo

Additional Information and Declarations can be found on page 13

DOI **10.7717/peerj.9371**

degradation again reach a new equilibrium, which makes the RCLD steady in some cells again (*Lopez-Maury, Marguerat & Bahler, 2008*). Therefore, the identification of a continuous pattern of changes in the rates of RLCD is crucial for cell homeostasis analysis.

Currently, experimental techniques based on short pulse labeling, such as 4-thiouridine (4sU), provide the possibility to identify RLCD (*De Pretis et al., 2015*; *Garibaldi, Carranza & Hertel, 2017*; *Melvin et al., 1978*; *Rabani et al., 2014*; *Sabo et al., 2014*). Incorporation of thiol-containing nucleosides 4-thiouridine into nascent RNA of eukaryotic cells within a few minutes allows for non-destructive metabolic labeling of RNA. By sequencing the 4sU-labeled RNA (4sU-seq), it is possible to separate the newly generated RNA (labeled RNA) from the originally existing RNA (unlabeled pre-RNA). *Rabani et al. (2011)* proposed a dynamic model framework for calculating transcription and degradation, without providing software implementation. *Schwalb et al. (2012)* used a similar method to calculate the rates of transcription and degradation, but it lacks a calculation of the processing rates. *Rabani et al. (2014)* used the model of the splicing maturity process and analyzed it at the junction level. In 2015, based on the same model, a novel framework INSPEcT (*De Pretis et al., 2015*) was proposed, which can calculate the RLCD flexibly. In 2017, an R package of pulseR was developed based on a negative binomial distribution model (*Uvarovskii & Dieterich, 2017*). Although these methods identified transcriptional dynamic rates from different perspectives, they were limited to the experimental measurement time. If the measurement time is short or the number of time nodes is small, it is difficult to analyze the complete RNA life cycle. Therefore, there is an urgent need to develop a tool with the function of predicting RCLD trends, which is of great significance for RNA life cycle analysis. This function is convenient for researchers to understand the complete process of some RCLDs in cells under external stimulation, and even the complete process of cell changes which cannot be detected due to experimental limitations, which extends the current capabilities of RNA life cycle analysis tools.

Here, we developed an R package termed pulseTD that can serve as a powerful tool to identify RNA RLCD based on the pulse model. It can adequately capture the trend of RLCD, which is important to analyze the process of cells from homeostasis to new homeostasis in cell-stimulated responses. More importantly, pulseTD shows better performance in predicting RLCD and gene expression values than other methods.

MATERIALS AND METHODS

Description of the GEO dataset

The 4sU-seq experimental datasets were obtained from GEO database (https://www.ncbi.nlm.nih.gov/geo/). The first dataset was an RNA-seq dataset from mouse DC (GSE56977). RNA was sampled from mouse DC every 15 min for the first 3 h (total 13 samples) of their response to LPS, followed by a short (10 min) metabolic marker pulse(4sU) before the sampling time point (*Rabani et al., 2014*). The samples of GSE59717 dataset were the infection of primary human foreskin fibroblasts with a wild-type simplex virus strain 17 with a multiplicity of infection of 10 (*Rutkowski et al., 2015*).

The algorithm flow of pulseTD model

The rates of RLCD within 24 h of the pulse state were evaluated, which represented a complete RNA life cycle, and the mode of oscillation was discharged (*La Manno et al., 2018*). We assumed that the rates of transcription, processing and degradation conforms to the functional form of the pulse model in the pulse state. Expression value data of pre-mRNA and mRNA at any time point could be obtained by 4su-seq and RNA-seq. The parameters were estimated by using an optimization algorithm. Next, we introduced the principles of the model and the use of the software. After that we created simulation data and evaluate performance.

First, we used the R package (GenomicAlignments, GenomicFeatures, etc.) to analyze RNA-seq and 4sU-seq bam files, in order to quantify intronic (reads alignment to the intron region) and exonic (reads alignment to the exon region) in Reads Per Kilobase of exon model per Million mapped reads (RPKM), Transcripts Per Kilobase of exon model per Million mapped reads (TPM) or Fragments Per Kilobase of exon model per Million mapped fragments (FPKM) of each gene. Some methods (min–max normalization, log normalization) had been added to standardize expression profile data before evaluating RLCD. In the cell life cycle, transcription continuously generates new pre-mRNA expressed by $P(t)$. Processing converts newly generated pre-RNA into mature mRNA expressed by $M(t)$. Finally, mRNA is targeted for degradation (*Moore & Proudfoot, 2009*). The equation according with the process of RLCD as follows:

$$\begin{cases} \dot{P} = \alpha - \gamma \\ \dot{M} = \gamma - \beta \end{cases} \tag{1.1}$$

Among them, α, β, γ represent the expression level of synthesized, degraded and processed RNA per unit time, respectively. Since the labeling time (t_L) is very short during the 4sU experimental, we assumed that RNA was not degraded during experimental labeling time. The total labeled RNA expression level $T_L(t)$ can be expressed as:

$$\begin{cases} \dot{T}_L = \alpha \\ \beta = 0 \end{cases} \tag{1.2}$$

The transcription level of intracellular RNA was in a steady state without being stimulated by outside world. When some factors (transient pulse stimulation) disrupted the equilibrium, the rates of genes at different stages were changed to cushion the stimulus. After a period of time, most cells returned to a steady state due to some factors such as cell morphology and stress. As mentioned above, the change in gene expression values was mainly the result of the combination of transcription, processing and degradation. Therefore, external stimulus conditions, pulse stimuli directly affected RLCD. To this end, we hypothesized that the three processes of gene expression were transcription α, processing γ, degradation β and $\alpha = \beta$, $\gamma = \alpha$ in steady state, which were broken when the external stimulus pulse was stimulated. After a period of time, it reached steady state again, and $\alpha' = \beta'$, $\gamma' = \alpha'$. The functional form of α, β, γ needed to be determined.

In this case, the rates of transcription, processing and degradation changed from $\gamma \sim \alpha \sim \beta$ to $\gamma' \sim \alpha' \sim \beta'$, which could be considered as a process of RNA from a steady state to a new steady state under pulsed stimulation. So, we used the pulse model $f(x)$ (*Chechik & Koller, 2009*) to represent functional form of α, β, γ:

$$f_\theta(x) = \frac{1}{h_1}\left(h_0 + (h_1 - h_0)\frac{1}{1+e^{-\beta(x-t_1)}}\right)\left(h_2 + (h_1 - h_2)\frac{1}{1+e^{\beta(x-t_2)}}\right) \tag{1.3}$$

The θ is the parameter vector represented by $(h_0, h_1, h_2, t_1, t_2, \beta)$. h_0, h_1, h_2 represent the initial state rates value, the peak time rates value and the new steady state rates value. t_1, t_2 are the maximum time for the first and second rise or fall changes, and β is the slope of the both transitions.

Taking into account the effects of existing RNA, we first estimated the global normalization factors in the model. We assumed that real mRNA R levels were composed of the labeled mRNA S and pre-existing mRNA N:

$$R = S + N \tag{1.4}$$

Make the total mRNA observations proportional to S and N, and the scale factors were both$w$. At the same time, we assumed that the labeled mRNA observations included two parts, labeled S and unlabeled N, with scale factors of w_1 and w_2, respectively. Then:

$$\begin{cases} w(S+N) = O_T = T(t) \\ w_1 S + w_2 N = O_L = T_L(t) \end{cases} \tag{1.5}$$

Three scale factors are constant in one sample, and O_T, O_L represent observations for total RNA and labeled RNA, respectively. According to formula (1.5):

$$(w_1 - w_2)S + \frac{w_2}{w}O_T = O_L \tag{1.6}$$

Here $P(t)$, $T(t)$ and $T_L(t)$ are represented as a linear combination of pulse function integrals. According to formula (1.1~1.6):

$$\begin{cases} P(t) = P(0) + \int_0^t f_{\theta_\alpha}(t)dt - \int_0^t f_{\theta_\gamma}(t)dt \\ T(t) = T(0) + \int_0^t f_{\theta_\alpha}(t)dt - \int_0^t f_{\theta_\beta}(t)dt \\ T_L(t) = \frac{w_2}{w}T(t) + (w_1 - w_2)\int_{t-t_L}^t f_{\theta_\alpha}(t)dt \end{cases} \tag{1.7}$$

Where $P(t), T(t), T_L(t)$ are the observation data of the measurement time node, t_L is the labeled time, $\theta_x = \left(h_0^x, h_1^x, h_2^x, t_1^x, t_2^x, \beta^x\right)$, so the parameter vector $X = \left(\theta_\alpha, \theta_\gamma, \theta_\beta\right)$. The total objective function is

$$J(X) = \frac{1}{2}\sum_{j=0}^{t_{end}}\left(\hat{P}_j - NP_j\right)^2 + \left(\hat{T}_j - NT_j\right)^2 + \left(\hat{T}_{Lj} - NT_{Lj}\right)^2 \tag{1.8}$$

Among them $\hat{P}, \hat{T}, \hat{T}_L$ are the model prediction value and NP, NT, NT_L are the standardized observation data. Therefore, α, β, γ can be solved by the following constrained optimization problem:

$$X = \arg minJ(X)$$

$$\text{s.t. } X > 0,$$

$$h_0^{\alpha} \sim h_0^{\beta} \sim h_0^{\gamma},$$

$$h_2^{\alpha} \sim h_2^{\beta} \sim h_2^{\gamma},$$

First, we evaluated the global normalization factor using a gradient descent method with random initialization parameters. We minimized the objective function by using the nlminb method, which was available in the stats R packages. The best fit result was chosen from the results of 100 (default) random initial values. At the same time, the chi-square test of goodness of fit was used to verify the statistical validity of pulseTD.

The interpretability of the model output

Here, the rates of transcription, processing and degradation were defined as the expression level of RNA transcribed, processed or degraded per unit time (unit: RNA/min or RNA/hour). In order to increase the interpretability of pulseTD and make it easier to compare with other tools, a conversion method was applied to pulseTD output. Let $\gamma = \gamma^k P$ and $\beta = \beta^k (T - P)$ at any time t. We assumed that the processing rates were directly proportional to the pre-mRNA expression and the ratio was γ^k. Similarly, degradation rates were directly proportional to mature mRNA expression value, the ratio was β^k. According to formula (1.1):

$$\begin{cases} \dot{P} = \alpha - \gamma^k P \\ \dot{T} = \alpha - \beta^k (T - P) \end{cases} \tag{1.9}$$

This was similar to previous researches (*De Pretis et al., 2015*; *Rabani et al., 2011*, *2014*; *Sun et al., 2012*). α represents the transcription rates in units of mRNA/min. Where $\dot{P}$ and $\dot{T}$ represent the derivatives of the functions $P(t)$ and $T(t)$ at time t, respectively. And they are functions of time t. Here we used $\alpha, \gamma^k, \beta^k$ to compare with other methods, and it was easy to explain.

Generation of simulation data

To evaluate the effectiveness of the model, we generated simulation data for 1,000 genes by randomly drawing from a specific distribution. Transcription, processing, and degradation rates were first generated, as well as randomly generated scale factors. The simulation expression value was then created based on the rates using Runge–Kutta method.

First of all, based on previous researches, we knew that the half-life of RNA was considered to follow a normal distribution (*Friedel et al., 2009*). Next, we determined the

distribution of the transcription rates based on the mean (μ) and variance (σ) of the observations. And the transcription rates were randomly extracted t from the normal distribution $N(\mu, \sigma^2)$. We had expected to determine the degradation rates in the same way. However, the degradation rates and the transcription rate were dependent. To simulate this dependency, we evaluated the correlation between transcription and degradation rates at any time based on the pearson correlation coefficient, which was k_t. Therefore, the degradation rates approximately obeyed the $N\left(\mu/k_t, (\sigma/k_t)^2\right)$ distribution. Similarly, processing rates were randomly drawn in the same way. Finally, we also randomly generated global scale factors, which were used to simulate existing and newly generated RNA.

After determining all rates, we used the fourth-order Runge–Kutta method of the R package (deSolve) to evaluate the expression levels of pre-mRNA, mature mRNA and labeled RNA as a simulation data set. Among them, the initial value of the expression value was randomly selected from the distribution of the observed data.

In general, we determined the rate distribution based on the experimental data and generated simulated data parameters based on the mean and variance of the rate. Here, the time range of the experimental analysis was 0–180 min, sampling was performed every 15 min, and the 4sU marking was performed 10 min before sampling. Finally, we got a simulation data set of 1,000 genes.

RESULTS

Software framework and description

The pulseTD combines the pulse model to predict the steady state of the RLCD. We defined the rates of transcription processing and degradation as a pulse function which had 6 parameters, a total of 18 parameters. To standardize RNA expression levels, additional global scale factors needed to be evaluated. For the evaluation of each gene, 100 random initializations were required. We used a multi-threaded approach to reduce runtime. The software supports different ways to evaluate expression levels such as counts, RPKM, TPM, FPKM. The workflow of pulseTD software is as follows (Fig. 1): (i) The RNA-seq and 4sU-seq data are aligned to the reference genome, and the results are used as the input files of the software. Expression values of pre-RNA, mRNA and labeled-RNA can be calculated in the form of RPKM, corresponding to the R function named extimateExpression. (ii) Subsequently, the expression profile is used to optimize the parameters of the pulse model. This is an optimization problem with six parameters, which are determined by MSE (minimizing the mean square error). The extimateParams function can estimate the pulse parameters of the transcription, processing and degradation rates. (iii) The parameters can be re-estimated using the correctionParams function because of the influence of random initial values. (iv) Next, transcriptional dynamic rates are solved flexibly, including the rates of transcription, processing and degradation, or the steady state rates are predicted. Complete guide reference documentation.

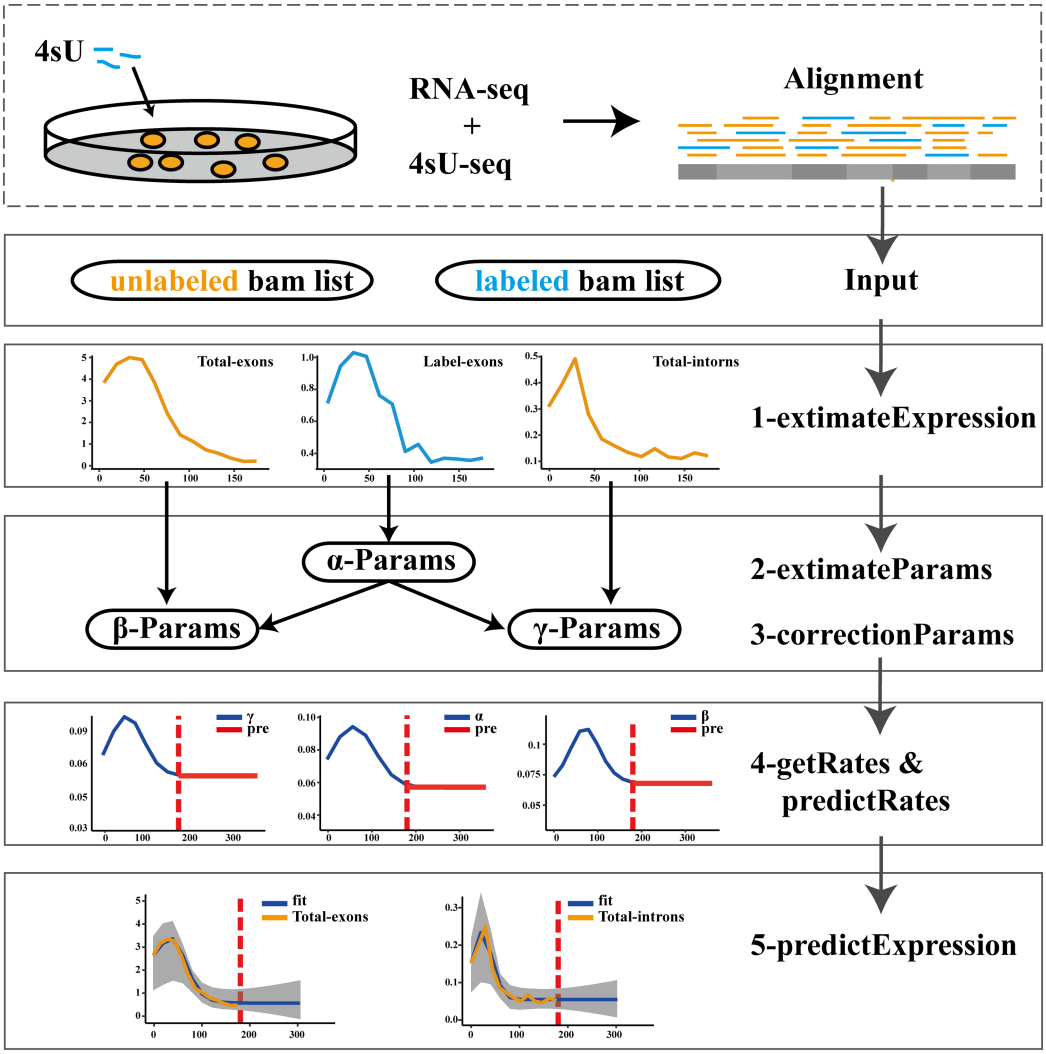

Figure 1 **Flowwork corresponding software function.** Full-size DOI: 10.7717/peerj.9371/fig-1

Compared with other methods

Several published studies had revealed RNA RLCD from different perspectives. We compared the following common software tools: INSPEcT (*De Pretis et al., 2015*), DRiLL (*Rabani et al., 2014*), DTA (*Schwalb et al., 2012*) and pulseR (*Uvarovskii & Dieterich, 2017*). pulseTD, INSPEcT and DRiLL provide evaluation functions for gene expression and processing rates, except for DTA and pulseR. The encoding of DRiLL is MATLAB, which requires a linux system and has a large operational limit. In terms of model, pulseTD directly applies the pulse model to different stages of RLCD, making the calculation of RLCD continuous. Although the pulse model was also used in INSPEcT and DRiLL, the purpose was to fit the expression pattern. Most importantly, pulseTD has the ability to predict RLCD (Table 1). Overall, pulseTD improves accuracy between biological repetitions and enhances the performance of evaluating low expression data. At the same time, it adds the ability to predict steady state. The assumption that the

Table 1 Detailed description of the software function.

	pulseTD	INSPEcT	DRiLL	DTA	pulseR
Code language	R	R	MATLAB	R	R
Gene expression	1	1	1	0	0
Processing rates	1	1	1	0	?
Prediction	1	0	0	0	0
Continuity	1	0	0	0	0

Note:
1, means available; 0, means unavailable; ?, representative unknown.

transcription rate is constant within the 4sU labeling time is abandoned in the solution process, and dynamic rates and gene expression values can be predicted at some future time points.

Performance analysis

The efficacy of pulseTD was evaluated on the simulation dataset. We calculated the Pearson correlation coefficient (PCC) for pulseTD between real and simulated transcriptional dynamic rates, and found that the PCC values for transcription, processing and degradation rates were 0.95, 0.95 and 0.77, respectively (Figs. 2A–2C). The PCCs for the transcription, processing, and degradation rates of INSPEcT were 0.94, 0.87 and 0.42, respectively. However, the rates of RCLD evaluated by INSPEcT were less than the true rates (Figs. 2D–2F). The MSE between real and simulated RCLD rates were calculated using pulseTD, INSPEcT and curve-fitting methods. The results showed that the MSE value of pulseTD was <0.1 (Fig. 3). These results suggested that pulseTD was accurate in assessing the RCLD rates.

To eliminate the bias of simulation data, simulation data produced by INSPEcT was used to evaluate the performance of pulseTD. We used the INSPEcT tool to generate two biological replications dataset, which contained 10 time nodes. The correlation of total RNA expression dataset was 0.98. Then, we used pulseTD to evaluate the RCLD of the biological replications, where the correlations of transcription, processing and degradation rates were 0.97, 0.90 and 0.92 (Figs. 4A–4C). As a comparison, we used INSPEcT to evaluate the RCLD of the simulation data. The correlations of transcription, processing, and degradation rates were 0.98, 0.86 and 0.84 (Figs. 4D–4F). In general, pulseTD had shown good performance in different simulation data.

The robustness of pulseTD was tested using two biologically repeated gene expression datasets from GEO (GSE59717). The PCC values of transcription, processing, and degradation rates between replications were 0.95, and the RCLD rates distribution also showed high consistency (Figs. 5A–5C). This reflected high robustness of pulseTD. For comparison, the PCC values of INSPEcT were calculated using the same data, they were 0.92, 0.93 and 0.84 (Figs. 5D–5F). The PCC for its degradation rates was low, which might be due to more outliers, during the evaluation process. And its density distribution produced a large deviation. pulseTD showed higher stability in biological

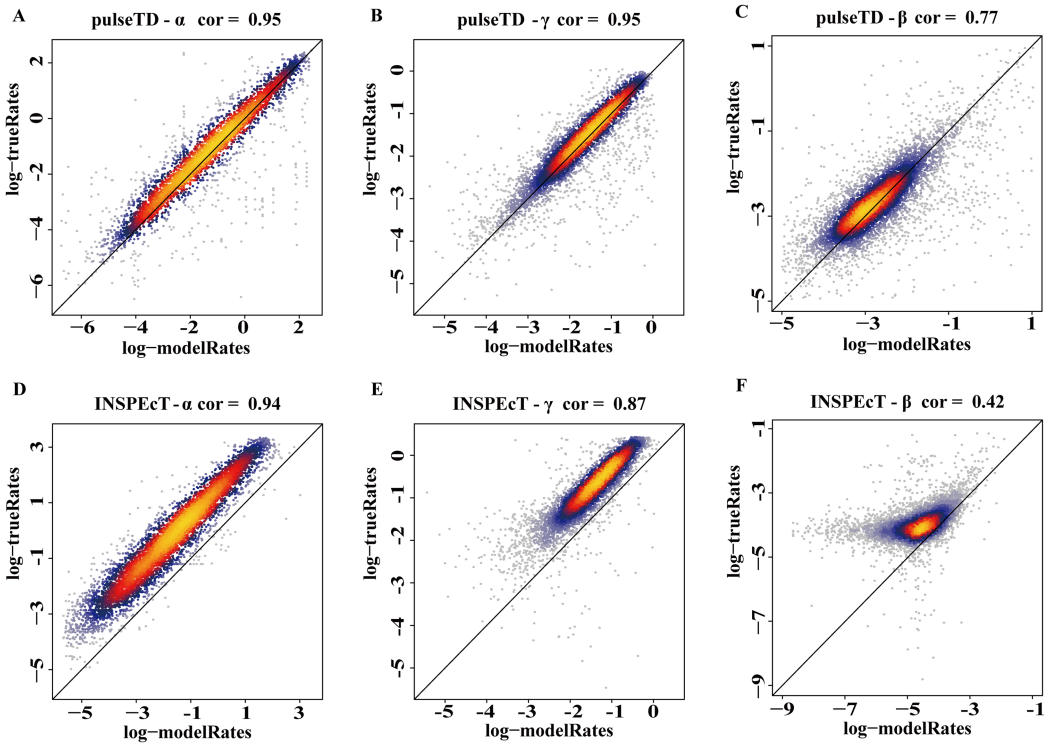

Figure 2 pulseTD and INSPEcT accuracy analysis and comparison. (A–C) Scatter plots of the correlation between the RLCD rates and the true rates evaluated by pulseTD, which represent the transcription, processing and degradation rates, respectively. The *x*-axis and *y*-axis are the logarithm of the simulated and real rates values. The color is closer to yellow, and the density of scatter is larger. (D–F) Scatter plots of the correlation between the RLCD rates and the true rates evaluated by INSPEcT, which represent the transcription, processing, and degradation rates, respectively. The *x*-axis and *y*-axis are the logarithm of the simulated and real rates values. The color is closer to yellow, and the density of scatter is larger. Full-size DOI: 10.7717/peerj.9371/fig-2

replicate data. At the same time, in order to judge the impact of expression levels of gene on software performance, we calculated the mean expression levels of the total exons, total introns and labeled exons of each gene during the detection time and ranked the mean. The first 1/3 genes were considered to be low expression, and the last 1/3 genes were considered to be high expression. Both of them were used to evaluate RCLD. The PCC value of degradation rates and processing rates obtained by INSPECT were 0.58 and 0.68, but the PCC value of pulseTD remained stable. This showed that pulseTD had a higher tolerance for poor quality data and the evaluation results were more stable. We also explored the effects of pre-RNA, total RNA and labeled RNA expression levels on model optimization. The correlation of the transcription, processing and degradation rates of each gene was calculated between the two biological replicate data sets. At the same time, we sorted them according to the mean expression levels of the total exons, total introns and labeled exons, respectively. We found that the number of genes with an average correlation greater than 0.5 was 96.21% and overall showed a high correlation (Figs. 6A–6C). This indicated the expression levels of total RNA, pre-RNA and labeled RNA had little effect on the optimization of the model.

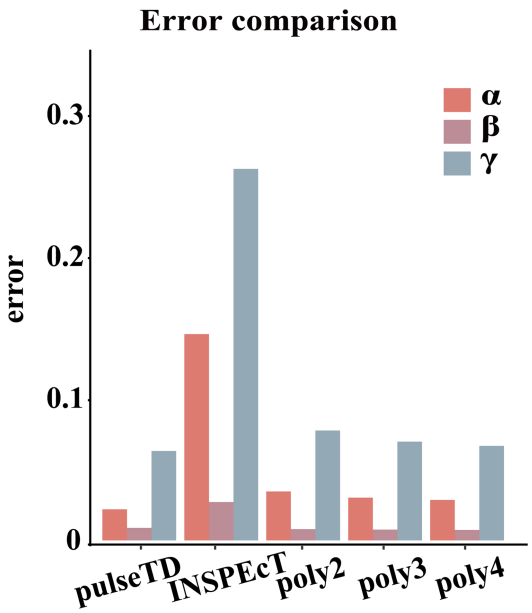

Figure 3 Compare RCLD errors between different software. The bar chart with error comparison of different methods, including pulseTD, INSPEcT, second-order (poly2), third-order (poly3) and fourth-order (poly4) polynomials. The *y*-axis is the error value of the RCLD, and α is the transcription rates, β is the degradation rates, and γ is the processing rates.

Full-size DOI: 10.7717/peerj.9371/fig-3

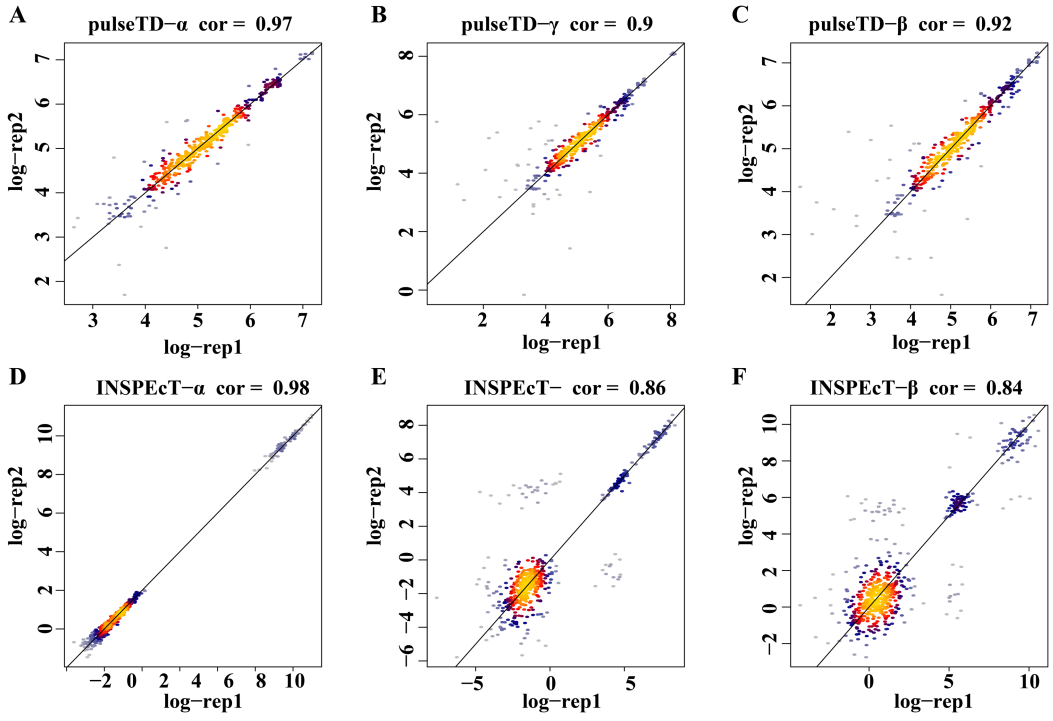

Figure 4 Accuracy analysis and comparison based on the simulation data generated by INSPEcT. (A–F) Scatter plots of the correlation. The *x*-axis is the logarithm of replication-1, the *y*-axis is the logarithm of replication-2. The color is closer to yellow, and the density of scatter is larger. (A–C) is using pulseTD software and (D–F) is using INSPEcT software.

Full-size DOI: 10.7717/peerj.9371/fig-4

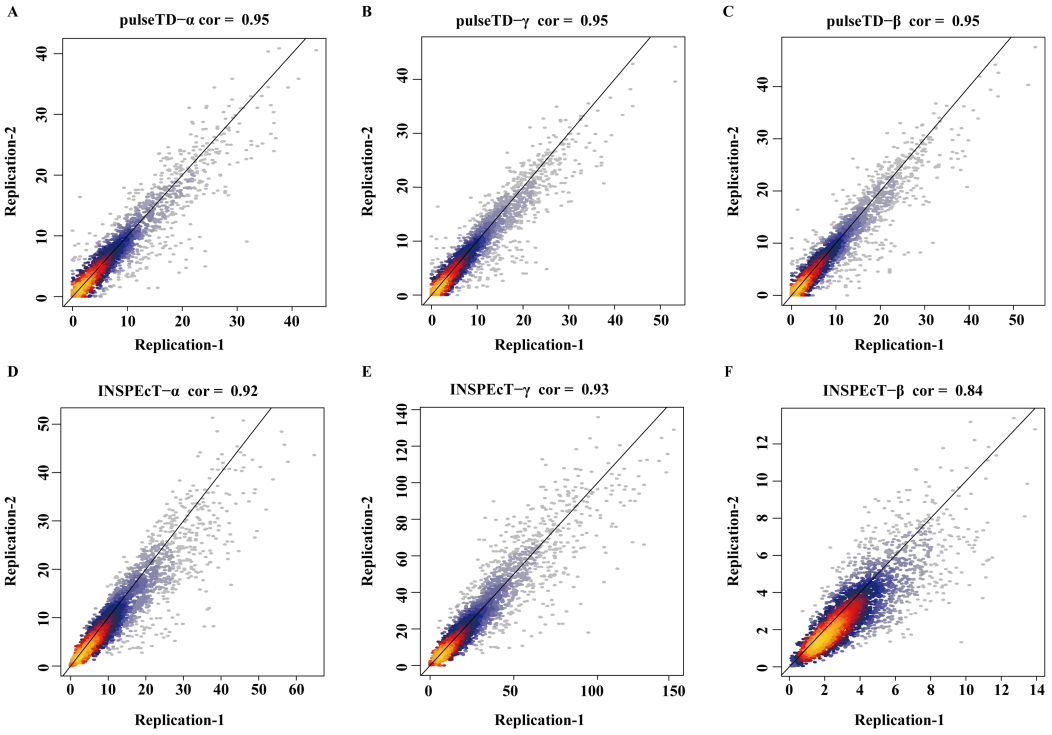

Figure 5 The correlation of biological duplicate data. (A–F) Scatter plots of the correlation. The *x*-axis is the replication 1, the *y*-axis is the replication 2. (A–C) is using pulseTD software and (D–F) is using INSPEcT software.
Full-size DOI: 10.7717/peerj.9371/fig-5

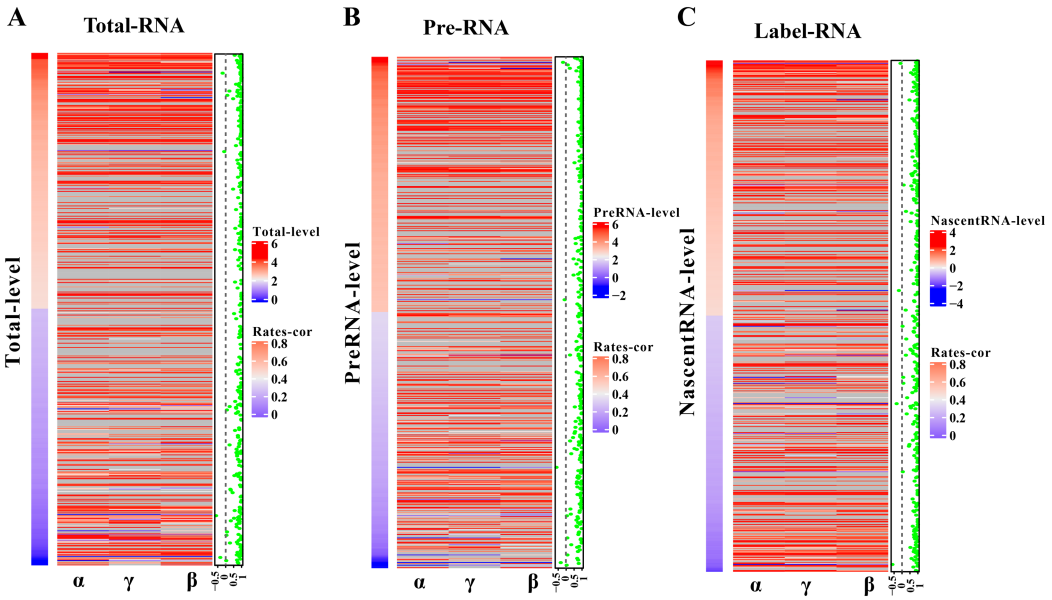

Figure 6 Influence of gene expression level on model optimization. (A–C) Heat maps of total RNA, pre-RNA and labeled RNA, respectively. (A) A heat map of expression levels. (B) The respective correlation of transcription, processing and degradation rates between two biological replicates. (C) A scatter plot of the mean correlation of transcription, processing, and degradation rates.
Full-size DOI: 10.7717/peerj.9371/fig-6

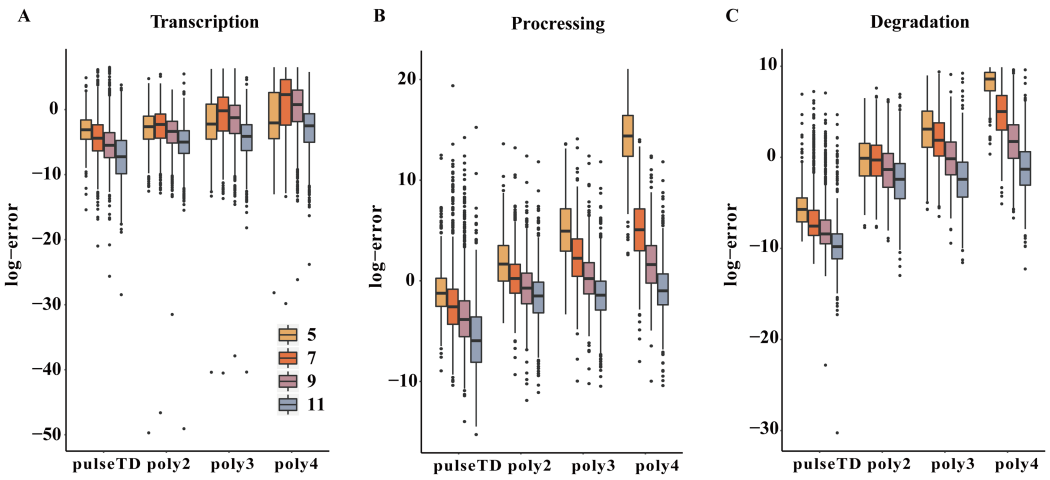

Figure 7 Comparison of the efficacy of different models for predicting the RCLD. (A–C) Box plots for the comparison of prediction errors of different methods, including pulseTD, INSPEcT, second-order (poly2), third-order (poly3) and fourth-order (poly4) polynomials. The smaller the error value, the higher the accuracy of the prediction. The *y*-axis is the negative logarithm of the mean square error between the predicted and true values, and 5, 7, 9 and 11 in the legend represent the number of experimental measurement time points, respectively. Full-size DOI: 10.7717/peerj.9371/fig-7

Predict RLCD at unknown time nodes

The pulseTD can predict dynamic transcription rates and gene expression because of its pulsed model characteristics. To evaluate prediction effectiveness, we divided the GEO (GSE56977) dataset into two parts. We selected the first 5, 7, 9 and 11 time points from 13 measurement time points as training samples (Assuming they are experimental measurement data) to estimate the rates of RCLD and used the remaining time points (test samples) for prediction. We estimated and predicted the rates of RCLD based on the training samples by using pulseTD and some other methods. Then, the MSE values between the predicted results and the test samples were calculated to evaluate the prediction performance. The results showed that pulseTD had the lower average MES values of RLCD rates than other methods (Figs. 7A–7C). These results showed that pulseTD had good prediction capabilities. We also found that as the training time points increase, the MSE values gradually decreased. When estimating RLCD, we recommend increasing the number of measurement time points in the experiment in order to have more accurate predictions.

CONCLUSIONS

In the field of bioinformatics, it is important to accurately identify the rates of RLCD and predict transcriptional stability. Few programs can identify the rates of RLCD, and none can provide predictions of the dynamic rates and steady state of RLCD. Here we use 4sU-seq and RNA-seq technology to analyze and predict RLCD. In summary, based on the pulse model, combined with the biological significance of RNA life cycle, we developed the R package named pulseTD. It only needs the alignment files of 4sU-seq and RNA-seq to calculate the expression value and the pulse model parameters in a simple manner. Here, we recommend using min–max normalization when comparing experiments with

different conditions to remove the dimension and logarithmic normalization when analyzing single experimental data to narrow the range of values. It can easily evaluate the RLCD at any time points. More importantly, it can predict the trend and the steady state of transcriptional dynamic rates. It has better accuracy and robustness than other methods. You can get source code on GitHub (https://github.com/bioWzz/pulseTD_0.2.0).

ACKNOWLEDGEMENTS

We would like to thank all members of the software engineering department and the teachers who provide guidance for the method.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the National Natural Science Foundation of China (No. 31701159), Harbin Science and Technology Bureau Project (No. 2017RAQXJ131), and Fundamental Research Service Fundamental Research Project of Heilongjiang Province (No. KJCX201815). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Natural Science Foundation of China: 31701159.
Harbin Science and Technology Bureau Project: 2017RAQXJ131.
Fundamental Research Service Fundamental Research Project of Heilongjiang Province: KJCX201815.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Xin Wang conceived and designed the experiments, performed the experiments, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Siyu He performed the experiments, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Jian Li performed the experiments, prepared figures and/or tables, and approved the final draft.
- Jun Wang performed the experiments, prepared figures and/or tables, and approved the final draft.
- Chengyi Wang analyzed the data, prepared figures and/or tables, and approved the final draft.
- Mingwei Wang analyzed the data, prepared figures and/or tables, and approved the final draft.
- Danni He analyzed the data, prepared figures and/or tables, and approved the final draft.

- Xingfeng Lv analyzed the data, prepared figures and/or tables, and approved the final draft.
- Qiuyan Zhong analyzed the data, prepared figures and/or tables, and approved the final draft.
- Hongjiu Wang conceived and designed the experiments, prepared figures and/or tables, and approved the final draft.
- Zhenzhen Wang conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The code is available on GitHub: https://github.com/bioWzz/pulseTD_0.2.0.

REFERENCES

Chechik G, Koller D. 2009. Timing of gene expression responses to environmental changes. *Journal of Computational Biology* **16(2)**:279–290 DOI 10.1089/cmb.2008.13TT.

De Pretis S, Kress T, Morelli MJ, Melloni GE, Riva L, Amati B, Pelizzola M. 2015. INSPEcT: a computational tool to infer mRNA synthesis, processing and degradation dynamics from RNA- and 4sU-seq time course experiments. *Bioinformatics* **31(17)**:2829–2835 DOI 10.1093/bioinformatics/btv288.

Friedel CC, Dolken L, Ruzsics Z, Koszinowski UH, Zimmer R. 2009. Conserved principles of mammalian transcriptional regulation revealed by RNA half-life. *Nucleic Acids Research* **37(17)**:e115 DOI 10.1093/nar/gkp542.

Garibaldi A, Carranza F, Hertel KJ. 2017. Isolation of newly transcribed RNA using the metabolic label 4-Thiouridine. *Methods in Molecular Biology* **1648**:169–176 DOI 10.1007/978-1-4939-7204-3_13.

La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, Lidschreiber K, Kastriti ME, Lonnerberg P, Furlan A, Fan J, Borm LE, Liu Z, Van Bruggen D, Guo J, He X, Barker R, Sundstrom E, Castelo-Branco G, Cramer P, Adameyko I, Linnarsson S, Kharchenko PV. 2018. RNA velocity of single cells. *Nature* **560(7719)**:494–498 DOI 10.1038/s41586-018-0414-6.

Lopez-Maury L, Marguerat S, Bahler J. 2008. Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nature Reviews Genetics* **9(8)**:583–593 DOI 10.1038/nrg2398.

Melvin WT, Milne HB, Slater AA, Allen HJ, Keir HM. 1978. Incorporation of 6-thioguanosine and 4-thiouridine into RNA. Application to isolation of newly synthesised RNA by affinity chromatography. *European Journal of Biochemistry* **92(2)**:373–379 DOI 10.1111/j.1432-1033.1978.tb12756.x.

Moore MJ, Proudfoot NJ. 2009. Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell* **136(4)**:688–700 DOI 10.1016/j.cell.2009.02.001.

Rabani M, Levin JZ, Fan L, Adiconis X, Raychowdhury R, Garber M, Gnirke A, Nusbaum C, Hacohen N, Friedman N, Amit I, Regev A. 2011. Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nature Biotechnology* **29(5)**:436–442 DOI 10.1038/nbt.1861.

Rabani M, Raychowdhury R, Jovanovic M, Rooney M, Stumpo DJ, Pauli A, Hacohen N, Schier AF, Blackshear PJ, Friedman N, Amit I, Regev A. 2014. High-resolution sequencing

and modeling identifies distinct dynamic RNA regulatory strategies. *Cell* **159(7)**:1698–1710 DOI 10.1016/j.cell.2014.11.015.

Rutkowski AJ, Erhard F, L'Hernault A, Bonfert T, Schilhabel M, Crump C, Rosenstiel P, Efstathiou S, Zimmer R, Friedel CC, Dolken L. 2015. Widespread disruption of host transcription termination in HSV-1 infection. *Nature Communications* **6(1)**:7126 DOI 10.1038/ncomms8126.

Sabo A, Kress TR, Pelizzola M, De Pretis S, Gorski MM, Tesi A, Morelli MJ, Bora P, Doni M, Verrecchia A, Tonelli C, Faga G, Bianchi V, Ronchi A, Low D, Muller H, Guccione E, Campaner S, Amati B. 2014. Selective transcriptional regulation by Myc in cellular growth control and lymphomagenesis. *Nature* **511(7510)**:488–492 DOI 10.1038/nature13537.

Schwalb B, Schulz D, Sun M, Zacher B, Dumcke S, Martin DE, Cramer P, Tresch A. 2012. Measurement of genome-wide RNA synthesis and decay rates with dynamic transcriptome analysis (DTA). *Bioinformatics* **28(6)**:884–885 DOI 10.1093/bioinformatics/bts052.

Sun M, Schwalb B, Schulz D, Pirkl N, Etzold S, Lariviere L, Maier KC, Seizl M, Tresch A, Cramer P. 2012. Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. *Genome Research* **22(7)**:1350–1359 DOI 10.1101/gr.130161.111.

Thapar R, Denmon AP. 2013. Signaling pathways that control mRNA turnover. *Cellular Signalling* **25(8)**:1699–1710 DOI 10.1016/j.cellsig.2013.03.026.

Uvarovskii A, Dieterich C. 2017. pulseR: Versatile computational analysis of RNA turnover from metabolic labeling experiments. *Bioinformatics* **33(20)**:3305–3307 DOI 10.1093/bioinformatics/btx368.

Zeisel A, Kostler WJ, Molotski N, Tsai JM, Krauthgamer R, Jacob-Hirsch J, Rechavi G, Soen Y, Jung S, Yarden Y, Domany E. 2011. Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. *Molecular Systems Biology* **7(1)**:529 DOI 10.1038/msb.2011.62.