Bioinformatics, 31(17), 2015, 2829–2835
doi: 10.1093/bioinformatics/btv288
Advance Access Publication Date: 7 May 2015
Original Paper

Gene expression

INSPEcT: a computational tool to infer mRNA synthesis, processing and degradation dynamics from RNA- and 4sU-seq time course experiments

Stefano de Pretis[1], Theresia Kress[1], Marco J. Morelli[1], Giorgio E. M. Melloni[1], Laura Riva[1], Bruno Amati[1,2] and Mattia Pelizzola[1,*]

[1]Center for Genomic Science of IIT@SEMM, Istituto Italiano di Tecnologia (IIT), 20139, Milano, Italy and [2]Department of Experimental Oncology, European Institute of Oncology (IEO), 20139, Milano, Italy

*To whom correspondence should be addressed.

Associate Editor: Ziv Bar-Joseph

Received on January 16, 2015; revised on April 13, 2015; accepted on May 3, 2015

Abstract

Motivation: Cellular mRNA levels originate from the combined action of multiple regulatory processes, which can be recapitulated by the rates of pre-mRNA synthesis, pre-mRNA processing and mRNA degradation. Recent experimental and computational advances set the basis to study these intertwined levels of regulation. Nevertheless, software for the comprehensive quantification of RNA dynamics is still lacking.

Results: INSPEcT is an R package for the integrative analysis of RNA- and 4sU-seq data to study the dynamics of transcriptional regulation. INSPEcT provides gene-level quantification of these rates, and a modeling framework to identify which of these regulatory processes are most likely to explain the observed mRNA and pre-mRNA concentrations. Software performance is tested on a synthetic dataset, instrumental to guide the choice of the modeling parameters and the experimental design.

Availability and implementation: INSPEcT is submitted to Bioconductor and is currently available as Supplementary Additional File S1.

Contact: mattia.pelizzola@iit.it

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The study of eukaryotic cellular transcriptional responses following external stimuli or during cell differentiation processes is typically based on profiling of mRNA abundances over time. This allows discriminating between early, intermediate and late responsive genes, but leaves undisclosed the dynamic transcriptional regulatory processes determining the resulting observed RNA level. mRNAs are synthesized within the nucleus thanks to a complex process that is controlled by numerous factors, including chromatin accessibility, transcription factors binding events and RNA Polymerase release (Orphanides and Reinberg, 2002). The joint action of these regulatory processes determines the efficiency of transcription, which can be recapitulated and measured as the RNA synthesis rate. Precursor mRNA molecules (pre-mRNAs), originated through the synthesis step, need to be processed in order to make a mature, functional mRNA molecule. Maturation involves the excision of the introns and the addition of a 5′-cap and a 3′ poly-adenine tail, modifications that co-occur at the transcription site and impact transcript stability.

Once mature mRNA is produced, it is translocated within the cytoplasm where it can be either translated or bound by RNA-binding proteins and targeted for degradation (Fu *et al.*, 2014; Houseley and Tollervey, 2009). Overall, synthesis, processing and degradation rates determine the levels of mRNA within the cell, and the combined modulation of these three elements determines changes in mRNA abundance over time (Braun and Young, 2014; Raghavan *et al.*, 2002; Shalem *et al.*, 2008).

Recently, an experimental technique based on a short pulse of a labeled nucleotide (4-thiouridine, 4sU) and consequent incorporation in the nascent RNA was developed, to measure the concentration of nascent mRNA and for the genome-wide inference of gene-level synthesis rates. During a short pulse (typically few minutes), cells medium is complemented with 4sU, a naturally occurring modified uridine that is incorporated within mRNA growing chains with minimal impact on cell viability (Melvin *et al.*, 1978). RNA chains having incorporated the uridine variant (newly synthesized) can be isolated from the total RNA population by biotinylation and purification with streptavidin-coated magnetic beads, followed by sequencing (4sU-seq). Consequently, various studies have elucidated the different roles that RNA synthesis, processing and degradation can have in response to various environmental conditions in yeasts and metazoans (Eser *et al.*, 2013; Miller *et al.*, 2011; Rabani *et al.*, 2011; Sun *et al.*, 2013; Zeisel *et al.*, 2011), illustrating how various and complex can be the transcriptional responses originated by the concerted action of these regulatory mechanisms.

Few analytical methods were proposed for the study of these regulatory mechanisms. The DTA Bioconductor package was used to determine synthesis and degradation rates in a number of works (Eser *et al.*, 2013; Miller *et al.*, 2011; Sun *et al.*, 2013). The main limitations of this tool are (i) the lack of inference of processing rates, (ii) the lack of computationally based normalization between 4sU and RNA-seq data, and more importantly (iii) the absence of a modeling framework. A similar method was described for the inference of synthesis and degradation rates, lacking a corresponding software implementation (Zeisel *et al.*, 2011). In a first study, Rabani *et al.* introduced a framework where rates are modeled to improve the confidence in the determination of rates absolute values. Additionally, this modeling approach was used to discriminate between constant and varying degradation processes (Rabani *et al.*, 2011). This method was extensively described in the publication but no software implementation was released. At the moment of this publication, the method described in Rabani *et al.* (2011) was extended to distinguish between constant and varying processing rates and a software implementation was released (DRiLL) (Rabani *et al.*, 2014). The software currently available has major limitations in terms of both documentation and implementation, and lacks functionalities to test its performance and provide guidance on the most suitable experimental design.

Given the limitations of these studies and their corresponding tools, we developed INSPEcT (INference of Synthesis, Processing and dEgradation rates in Time-course analysis), a computational tool for the joint analysis of 4sU- and RNA-seq data providing robust synthesis, processing and degradation rates over time, based on a system of differential equations describing mRNA production, maturation and degradation processes. INSPEcT can determine the probability of a given combination of rate(s) to regulate the gene expression during the time-course by modeling and comparing alternative scenarios of transcriptional regulation. Importantly, INSPEcT allows testing the performance of the tool given a dataset and predicting how many additional replicates and/or time points would be needed to reach the desired performance.

2 Software implementation and overview

INSPEcT is a computational tool for the analysis of RNA- and 4sU-seq time-course data, resulting in the inference of RNA synthesis, processing and degradation rates over time, and allowing to statistically assess their contribution in shaping the expression level of a gene. INSPEcT is based on estimation of total mRNA levels and pre-mRNA levels (from RNA-seq), synthesis rates and processing rates (from 4sU-seq), and degradation rates from the combined analysis of these two data types.

The INSPEcT R package is submitted to the Bioconductor project and was developed in compliance with the most common Bioconductor infrastructures. Specifically, classes inheriting from the ExpressionSet class are used to represent high-throughput gene expression data, and the TranscriptDb class (available for an extensive set of organisms) is adopted as a reference for gene models. Therefore, methods and functions available within this package can be easily integrated with other Bioconductor packages for up- or down-stream analysis steps. Parallel computation is used to minimize the computational time for most demanding tasks, as in the case of the *modelRates* INSPEcT method. As required in Bioconductor, each individual method and function is accompanied by specific documentation and working examples. Moreover, INSPEcT includes a vignette to interactively demonstrate the software key functionalities and a typical workflow (also reported as Supplementary Additional File S2). Methods, functions and classes available in the INSPEcT package matching the discussed functionalities are in italic throughout the text, and the main steps in the software workflow are outlined here:

- Exonic and intronic RPKMs (Reads Per Kilobase per Million mapped reads) for both RNA- and 4sU-seq datasets are determined for each gene (*makeRPKMs* function). Exonic and intronic RNA-seq RPKMs allow quantifying total mRNA and pre-mRNA, respectively.
- Normalized synthesis, processing and degradation rates are obtained integrating RNA- and 4sU-seq data (*newINSPEcT*).
- Total mRNA and pre-mRNA concentrations, and synthesis, processing and degradation rates are modeled for each gene to assess which of the rates (if any) is changing over time reconstructing observed abundances of mRNAs (*modelRates*).
- Simulated data that recapitulate rate distributions, their variation over time and their pair-wise correlations are created and used to evaluate the performance of the method (*makeSimModel*, *makeSimDataset*, *rocCurve*).

Figure 1 illustrates the overall INSPEcT design along with the main input and output.

3 Mathematical model

INSPEcT is based on a set of differential equations describing the process of production, maturation and degradation of pre-mRNA and mature mRNAs. The system of differential equations models the synthesis of new pre-mRNAs (P) by a process that occurs at rate $a(t)$ (synthesis). pre-mRNAs decay into mature mRNAs (M) exponentially following a rate $c(t)$ (processing). Mature mRNAs are exponentially degraded following rate $b(t)$ (degradation). Total mRNAs levels are defined as the sum of pre-mRNA and mature mRNA levels ($T = P + M$).

$$\emptyset \xrightarrow{a} P \xrightarrow{c} M \xrightarrow{b} \emptyset \quad (1)$$

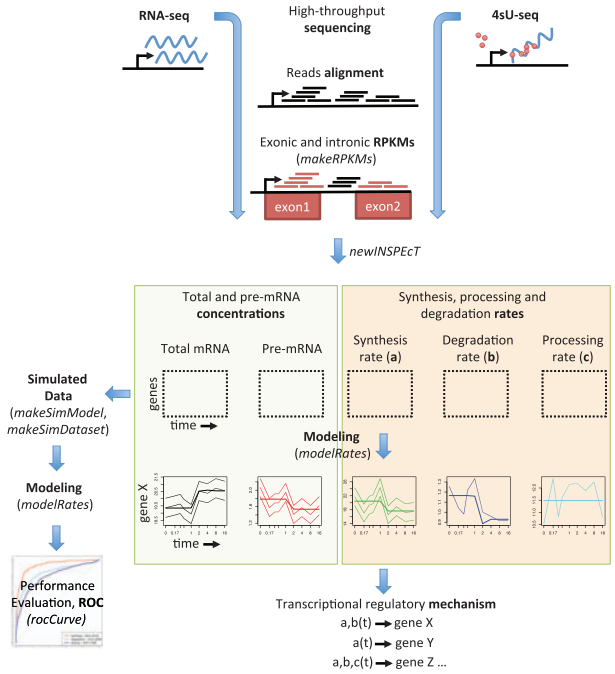

Fig. 1. Diagram illustrating the main steps of a typical INSPEcT workflow. INSPEcT functions and methods that can be used for specific steps are indicated in italic. Input RNA- and 4sU-seq sequencing data are both subjected to alignment and exonic and intronic RPKMs are computed. These data are passed to the *newINSPEcT* method that determines total and pre-mRNA concentrations and normalized rates. On one hand these data can be used for the creation of a simulated dataset; simulated RNA concentrations and rates are subjected to modeling, leading to evaluation of modeling performance through ROC analysis. On the other hand, real RNA concentrations and rates are subjected to modeling thus providing for each gene the most likely transcriptional regulatory mechanism. For example, as indicated in the figure the expression of gene X over time is mostly under control of varying synthesis (a) and degradation (b) rates

$$\begin{cases} \dot{P} = a(t) - c(t)P \\ \dot{T} = a(t) - b(t)(T - P) \end{cases} \quad (2)$$

This model makes the assumptions that pre-mRNAs are not degraded, and that translocation of mRNAs from nucleus (where they originate) to cytoplasm (where they can be degraded) occur immediately after maturation, or at a rate considerably faster than $b(t)$. These assumptions are typically considered acceptable and have been previously referred to (Rabani *et al.*, 2011). The model also imposes no spatial segregation of mRNA into cellular compartments. This phenomenon can have an impact on degradation of some mRNAs, but the experimental information is typically difficult to obtain (or not available).

3.1 Determination of normalized rates and concentrations

After having quantified data from RNA-seq (R) and 4sU-seq (labeled, L) libraries into intronic and exonic RPKMs (*makeRPKMs* function), the *newINSPEcT* method is used to estimate synthesis, processing and degradation rates by solving the system of differential equations (2) applied at every time point (t) to both the total and labeled fractions. When applied to the labeled RNA fraction, the system can be solved and integrated between $t - t_L$ and t, assuming that no labeled molecules existed before the labeling pulse (t_L). INSPEcT by default assumes that no degradation occurs during the short labeling time (typically 10 min). Consequently, we have four equations with three unknowns (a_t, b_t, and c_t):

$$\begin{cases} \dot{P}_{R_t} = a_t - c_t P_{R_t} \\ \dot{T}_{R_t} = a_t - b_t(T_{R_t} - P_{R_t}) \\ P_{L_t} = \frac{a_t}{c_t}(1 - e^{-c_t t_L}) \\ T_{L_t} = a_t t_L \end{cases} \quad (3)$$

For each gene, P_{R_t} is equal to its pre-mRNA level (intronic RNA-seq RPKM), T_{R_t} is equal to the total mRNA level (exonic RNA-seq RPKM), and P_{L_t} is equal to the pre-mRNA level as quantified in the labeled fraction (intronic 4sU-seq RPKM). Finally, T_{L_t} is equal to the total mRNA level as quantified in the labeled fraction (exonic 4sU-seq RPKM). $\dot{P}_{R_t}$, $\dot{T}_{R_t}$ are estimated from the interpolation of the $P_R(t)$ and $T_R(t)$ time-courses using cubic splines.

INSPEcT takes advantage of the over-determination of the system to identify for each time point a normalization factor n_t that is used to scale labeled data (P_{L_t},T_{L_t}). n_t is determined minimizing the difference between c_t as estimated from (i) equations (3.1) and (3.3), and (ii) equations (3.3) and (3.4) [see Supplementary Additional File S3, equations (s1)–(s2)].

Subsequently, while synthesis rates are calculated solving equation (3.4), processing and degradation rates are determined by *newINSPEcT* iteratively integrating equations (3.1) and (3.2), assuming linear behavior of synthesis rate, pre-mRNA and total mRNA between time-points.

$$\begin{cases} [P_R(t)e^{c_t \cdot t}]_{t_i}^{t_{i+1}} = \int_{t_i}^{t_{i+1}} a(t) \cdot e^{c_t \cdot t} dt \\ [T_R(t)e^{b_t \cdot t}]_{t_i}^{t_{i+1}} = \int_{t_i}^{t_{i+1}} a(t) \cdot e^{b_t \cdot t} dt + b_t \int_{t_i}^{t_{i+1}} P_R(t) \cdot e^{b_t \cdot t} dt \end{cases} \quad (4)$$

The solution of the first time point is obtained directly from the data assuming the steady state.

$$\begin{cases} c_0 = {}^{a_0}/_{P_{R_0}} \\ b_0 = {}^{a_0}/_{(T_{R_0} - P_{R_0})} \end{cases} \quad (5)$$

This procedure of estimation of degradation and processing rates is more robust compared with estimating derivatives of T and P directly from the data [for details see Supplementary Additional File S3, equations (s3)–(s7)].

In case of experiments based on longer 4sU pulses, the assumption that no degradation of newly synthesized transcript occurs does not hold anymore. To cope with this experimental design, the *newINSPEcT* method provides an alternative set of equations that take into account the degradation of transcripts during the pulse [see Supplementary Additional File S3, equation (s8)].

3.2 Modeling of rates and concentrations

Once a prior estimate is obtained for synthesis, processing and degradation rates over time for each gene, INSPEcT tests different models of transcriptional regulation to identify the most likely combination of rates explaining the observed changes in gene expression (*modelRates* method). To this purpose, a parametric function is fit on each rate over time, through minimization of residual sum of squares. For example in the case of RNA synthesis, after having selected function f for the fit of a synthesis rate for a given gene, a set of parameters p_a is chosen after n random initializations as the one that better explains the data [see Supplementary Additional File S3, equation (s9)].

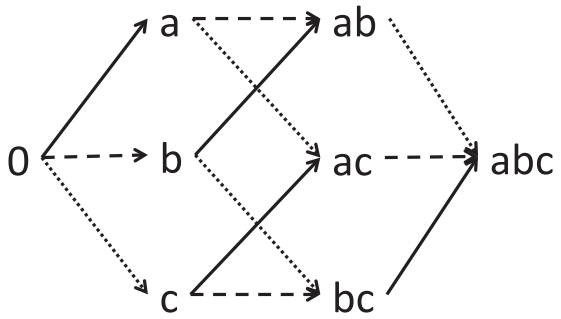

Fig. 2. Schema illustrating the models to be compared for determining if a transcriptional regulatory rate is variable or constant over time. mRNA synthesis (a), degradation (b) and processing rates (c) are individually tested for being variable or constant; models are named after rate(s) that are hypothesized to be variable. For each rate, multiple nested models can be compared, indicated by identical arrows. As an example, dashed arrows indicate the four models that could be tested to determine the likelihood that the degradation rate (b) is varying over time, given the provided data (see text for details)

Both f and n can be user-defined in INSPEcT. By default, rates are fit using both the sigmoid and the impulse model (Chechik and Koller, 2009) functions, and the one that better explains the data in terms of goodness of fit, measured by χ^2 test is chosen, thus penalizing the impulse model for the added complexity. For both functions, due to the reduced number of the time-points typically available in a biological time-course experiment, a strong priori to the random initialization of parameters is given based on the data. Once the parametric functionalization for synthesis, degradation and processing rates were obtained, it is possible to test how those parametric functions were able to recapitulate the experimental data they originated from after an additional minimization step [see Supplementary Additional File S3, equations (s10)–(s11)]. To identify the most likely mechanism of transcriptional regulation, INSPEcT tests the possibility that each rate is constant during the time course by building models that alternatively set as constant one, two or all the three rates. Regarding the degradation rate for example, four pairs of models (dashed arrows in Fig. 2) can be considered to test whether this rate is more likely to be considered variable or constant. All these pairs are nested models that can be evaluated via log likelihood ratio test against the null hypothesis that the degradation rate is constant. Only pairs of nested models where at least one model is successfully evaluated by χ^2 test are considered, and resulting *P*-values are combined using Brown's method (Brown, 1975) The same procedure is thus applied for the evaluation of synthesis rate (comparing nested models as indicated by the solid arrows in Fig. 2) and processing rate (dotted arrows). Alternatively, model selection through AIC (Akaike information criterion) is supported.

4 Evaluation of performance through simulated data

Simulated data were used to evaluate the performance of INSPEcT in classifying each rate as constant or variable, and to estimate the number of time points and replicates necessary to achieve a given performance. Simulated data are generated on the basis of real data (RNA concentrations and rates) and aim at: (i) reproducing the distributions of their absolute intensities, their variation over time, and the correlations between these features; and (ii) obtaining a dataset where the ground truth is known in terms of which genes are under the control of varying or constant rates.

4.1 Generation of the simulated data

The generation of the simulated data involves the following steps (see also Fig. 1 and Supplementary Additional File S3 for details):

i. Real rates quantification (*newINSPEcT* function): rates are quantified on a real time-course dataset (including N time points).
ii. Parametric functions of simulated rates (*makeSimModel* method): distributions of rates, their variation over time, and their pairwise correlations are evaluated. As a result, functions of the simulated rates and their parameters are returned for the required number of genes (G) (see an example for the synthesis rate in Supplementary Fig. S6 in Supplementary Additional File S3).
iii. Recapitulating the variance of real data (*makeSimModel* method): to sample the noise that will be added to the simulated data, a time-course of synthesis, degradation and processing rates is created for N time points, and pre-mRNA and total mRNA levels are coherently generated for G genes. Rates and concentrations are then used to derive simulated exonic and intronic signals of labeled and total RNA datasets. The resulting simulated datasets are intrinsically noise-free and indeed show a reduced variance compared with the real data. The missing variance is determined and returned for each gene G and dataset (see an example for the exonic signals from the labeled library in Supplementary Fig. S7 in Supplementary Additional File S3).
iv. Generation of simulated data (*makeSimDataset* method): given a required number of time-points and replicates, and the output of the *makeSimModel* method, for each gene G simulated rates and concentrations are calculated, white-noise is added (with the given noise variance), and exonic and intronic signals are coherently reconstructed and returned for both the total and the labeled dataset.

The obtained simulated dataset, subjected to the modeling by the *modelRates* method, is provided to the *rocCurve* method, which uses a ROC-curve analysis to measure classification performance in terms of sensitivity and specificity. False negatives (FN) represent cases where the rate is identified as constant while it was simulated as varying. False positives (FP) represent cases where INSPEcT identified a rate as varying while it was simulated as constant. On the contrary, true positives (TP) and negatives (TN) are cases of correct classification of varying and constant rates, respectively. Consequently, sensitivity and specificity are computed using increasing thresholds for the brown's *P*-values, and the ability of correctly classifying a rate is measured through the area under the curve (AUC) for each rate.

4.2 Evaluation of INSPEcT classification performance with different experimental designs

To evaluate INSPEcT performance, we generated a pilot experiment using mouse 3T9 cells transfected with the MycER construct so that the expression of the Myc transcription factor can be induced beyond endogenous levels. The transcriptional response was followed over time (9 time points) by RNA-seq and 4sU-seq sequencing (10 min 4sU pulse; the data are available within the INSPEcT package). Normalized rates and concentrations are determined applying the *newINSPEcT* method on the experimental data as illustrated earlier and were used to generate synthetic datasets with different numbers of replicates and time points that were subjected to ROC analysis. In all tested cases the performance achieved for the synthesis rate was higher in terms of AUC compared with degradation and processing rates. This is expected, since only the synthesis rate is determined directly from the labeled RNA fraction. When increasing the number of

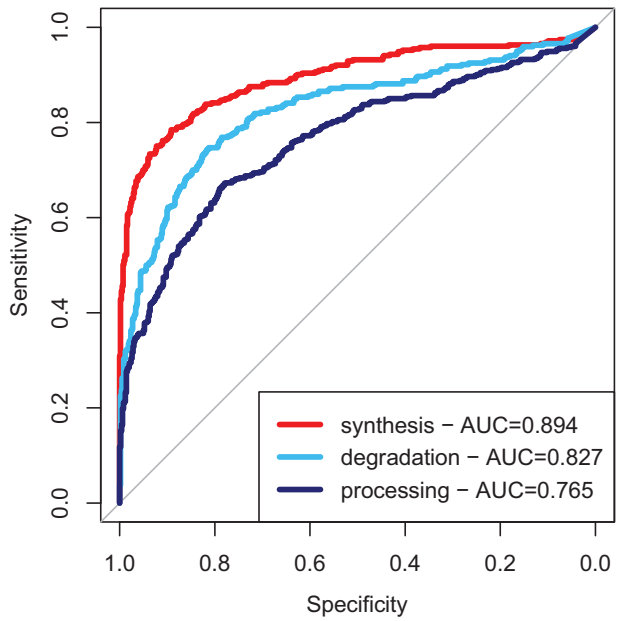

Fig. 3. The ability of identifying variable or constant rates in simulated data (twelve time-points, three replicated time-courses) is evaluated through ROC curves. Sensitivity and specificity are calculated for each of the synthesis, degradation and processing rates considering different p-values cutoffs. The area under the curve (AUC) is reported in the legend

time points from 9 to 12 and the number of replicates from 1 to 3, the performance of the classification significantly improved: AUC increased from 0.78 to 0.89 for the synthesis rate, from 0.65 to 0.83 for the degradation rate and from 0.61 to 0.77 for the processing rate (Fig. 3 and Supplementary Fig. S1 in Supplementary Additional File S3). As expected, using a simulated dataset including only two replicated time-courses provided intermediate performance, while increasing from three to four replicated time-courses did not significantly improve the results, thus not justifying the increase sequencing cost. We concluded that, given the simulated data created on the basis of the pilot experiment data, an experimental design comprising three replicated time-courses of 12 time points each would provide the best trade-off in terms of cost and performance.

5 Comparison with existing methods and tools

We compared INSPEcT theoretical and computational framework with available tools and data from the following studies:

i. (Rabani *et al.*, 2011); from here on data will be named Rabani2011; no software provided;
ii. (Miller *et al.*, 2011); data: Miller2011; software: DTA
iii. (Rabani *et al.*, 2014); data: Rabani2014; software: DRiLL

5.1 Rabani2011 and DRiLL

Compared with the method proposed in Rabani *et al.* (2011), INSPEcT implements and significantly extends the set of equations describing the dynamics of system, and the modeling and testing framework. Very recently, a revised version of the method originally proposed in Rabani *et al.* (2011) was published, extending the former approach to quantify pre-mRNAs and model processing rates (Rabani *et al.*, 2014). Both Rabani2011 and Rabani2014 are based on high-throughput measurements of the transcriptional response of mouse dendritic cells exposed to lipopolysaccharide, for both total and 4sU-labeled RNAs. Rabani2011 data were obtained through the Nanostring nCounter platform, thus providing digital measurements based on a limited number of probes for several hundreds genes. Thanks to INSPEcT flexibility, we could analyze these data despite the lack of probes covering introns, at the cost of an impaired quantification of pre-mRNA. Although we were not able to quantify pre-mRNA processing rates, we could recapitulate the published results in terms of degradation rates distribution and degradation dynamics of genes that were considered variable over time (Supplementary Figs S2 and S3 in Supplementary Additional File S3).

For a more quantitative comparison, we reanalyzed Rabani2014 data, for which rates resulting from DRiLL analysis are available. Data were provided in the form of pre-mRNA and mature mRNA expression intensities of the labeled and the total fraction. We derived intronic expression directly from their pre-mRNAs intensities and exonic expression as the sum of the pre-mRNAs and mature mRNAs levels. Intronic and exonic expressions of both labeled and total fractions were provided to the newINSPEcT method and rates were obtained and further subjected to modeling. According to Rabani *et al.* (2014) the correlation of DRiLL-derived rates with their previous study (Rabani *et al.*, 2011) was considered significant ($r = 0.39$ and 0.23 for degradation and processing rates, respectively). Similarly, rates estimated with INSPEcT on Rabani2014 modeled rates are correlated at 0.33 and 0.24 for degradation and processing, respectively. Importantly, after modeling these rates with INSPEcT, the correlations increased to 0.56 and 0.28, respectively. Synthesis rates are correlated at 0.65 and 0.69 pre- and post-modeling. We consider these correlations to be modest and we evaluated the possibility that this variation is due to the different normalization strategy adopted by INSPEcT and DRiLL. DRiLL introduced a normalization method that tries to estimate the contamination of total RNA in the labeled library. The authors estimated this contamination to be around 30% in Rabani2014 data, and they consequently adjusted the data for this factor. We adopted the same method they described to estimate this contamination factor on Rabani2014 and on the data in the current study. We could not confirm this level of contamination, which was estimated at 10% in their dataset and was absent in ours. When compensating for this factor, the correlation between INSPEcT and DRiLL only marginally increased to 0.75 and 0.36 for pre-model synthesis and processing rates, respectively. Eventually, we decided to refrain from implementing this correction, while ruling out that this might represent a reason for the observed difference in the estimated rates.

Similarly to DRiLL, INSPEcT provides a computational routine for the normalization of RNA- and 4sU-seq data, allowing an unbiased estimation of the scaling factors at each time point. This was previously achieved based on the yield of the 4sU-labeled mRNA recovered from the total RNA (Miller *et al.*, 2011; Rabani *et al.*, 2011). However, this procedure depends on the 4sU availability within the cell, which can be influenced by the rate of nucleotides transportation and metabolism. Considering the fact that genes involved in uridine transportation and metabolization are often regulated over time-resolved transcriptional responses (see Supplementary Figs S4 and S5 in Supplementary Additional File S3), the scaling factors computed using the amount of 4sU recovered and the one estimated computationally can be highly different and lead to considerably different global results. Therefore, we propose to rely on the computational scaling factor as a more robust alternative for the analysis of 4sU-seq data.

Importantly, Rabani2014 dataset lacks replicates, which is far from the optimal choice when using INSPEcT, especially for the

modeling framework. It is not clear to us how they managed the lack of replicated experiments. In fact, Supplementary Figure S1 in our manuscript displays the performance evaluation on the analysis of our own experimental data in single replicate with sequencing depth similar to the Rabani2014 dataset (86M versus 68M rRNA-depleted, unaligned reads in their dataset, compared to our own). We show there that the performance of the constant/varying classification is quite low with data of this depth and lack of replicates (Supplementary Fig. S1). Our ROC analysis suggests that their dataset did not have enough replicates. For this reason, we decided to avoid to compare the classification between the two methods.

Eventually, the ROC analysis based on the simulated data generated by INSPEcT, could have been conclusive on the relative performance of the two tools. Unfortunately, we were unable to use DRiLL neither on their nor on our own data, having this software major limitations in terms of both documentation and implementation.

5.2 DTA

DTA is an alternative method available for the joint analysis of RNA- and 4sU-seq data (Miller *et al.*, 2011). Compared with INSPEcT, DTA is limited to the inference of RNA synthesis and degradation rates in terms of variation between two conditions. Moreover, it does not offer the possibility of determining pre-mRNA levels and computing pre-mRNA processing rates. Importantly, DTA lacks any modeling, which is central in INSPEcT, and consequently is not able to infer which layer(s) of transcriptional regulation (synthesis, degradation or processing in the case of INSPEcT) are most likely to be responsible for the final mRNA level. Regarding the normalization of RNA- and 4sU-seq data, an expanded version of DTA was released (Eser *et al.*, 2013; Sun *et al.*, 2012), proposing an experimental procedure to scale the data from the two sequencing libraries, which is based on the adoption of internal standards quantified along with the sample, thus allowing the estimation of absolute rate intensities. However, this procedure does not eliminate the bias due to 4sU-metabolism discussed earlier.

In order to directly compare INSPEcT to DTA, we used our tool to reanalyze the Miller2011 dataset. As previously discussed for Rabani2011, the Miller2011 dataset does not include expression intensities for introns and only synthesis and degradation rates can be determined. Leveraging on INSPEcT flexibility, we were able to reanalyze these data, and we obtained a correlation of 0.91 and 0.70 for pre-model synthesis and degradation rates, respectively. Importantly, Miller2011 degradation rates are validated comparing with experiments performed in response to drugs blocking transcription, thus indirectly validating the rates obtained with INSPEcT.

5.3 Advantages of INSPEcT compared with available methods (DRiLL and DTA)

Recapitulating, the INSPEcT advantages compared with DTA are:

- the ability of dealing with the intronic signal to determine pre-mRNA levels and quantify mRNA processing rates
- the presence of a comprehensive modeling framework, based on ODE modeling of RNA dynamics
- a computational normalization method to scale labeled and total mRNA intensities that compensates for 4sU-metabolism bias

INSPEcT advantages compared with DTA and DRiLL are:

- the INSPEcT ability to generate simulated data and perform a ROC analysis to evaluate the performance of the classification and guide on the experimental design
- the flexibility of working with both exonic only and exonic plus intronic signals
- the ability of dealing with long 4sU pulses, taking into account degradation during the pulse (which is in our experience a relatively common experimental choice among our collaborators)
- the possibility of fitting either impulse or sigmoid functions on the data to reduce overfitting
- the possibility of using Akaike information criterion (AIC) for model selection
- the availability as open-source R package in the Bioconductor project and of the extended documentation for both individual methods and the overall workflow (see Supplementary Additional File S4). This only applies with respect to DRiLL.

6 Conclusions

In conclusion, INSPEcT provides an R/Bioconductor compliant solution for the study of dynamic transcriptional regulatory processes. Based on RNA- and 4sU-seq time-course datasets, which can be jointly analyzed thanks to a computational normalization routine, INSPEcT determines mRNA synthesis, degradation and pre-mRNA processing rates over time for each gene, genome-wide. The INSPEcT modeling framework allows the identification of gene-level transcriptional regulatory mechanisms, determining which combination of constant or variable synthesis, degradation and processing rates is most likely to be responsible for the observed mRNA level over time. Importantly, given a dataset, INSPEcT allows testing its performance in classifying rates as constant or varying, and in predicting how many additional replicates and/or time points would be needed to reach the desired performance.

Acknowledgements

The authors would like to thank Magnus Rattray, Valerio Bianchi, and Anna Russo for critical feedback and discussions, and all R/Bioconductor developers.

Funding

This work was supported by the European Community's Seventh Framework (FP7/2007-2013) project RADIANT (grant number 305626) to M.P., and a grant from the Italian Association for Cancer Research (AIRC) to B.A.

Conflict of Interest: none declared.

References

Braun,K.A. and Young,E.T. (2014) Coupling mRNA synthesis and decay. *Mol. Cell. Biol.*, **34**, 4078–4087.

Brown,M. (1975) A method for combining non-independent, one-sided tests of significance. *Biometrics*, **31**, 987–992.

Chechik,G. and Koller,D. (2009) Timing of gene expression responses to environmental changes. *J. Comput. Biol.*, **16**, 279–290.

Eser,P. *et al.* (2013) Periodic mRNA synthesis and degradation co-operate during cell cycle gene expression. *Mol. Syst. Biol.*, **10**, 717.

Fu,Y. *et al.* (2014) Gene expression regulation mediated through reversible m6A RNA methylation. *Nat. Rev. Genet.*, **15**, 293–306.

Houseley,J. and Tollervey,D. (2009) The many pathways of RNA degradation. *Cell*, **136**, 763–776.

Melvin,W.T. *et al.* (1978) Incorporation of 6-thioguanosine and 4-thiouridine into RNA. Application to isolation of newly synthesised RNA by affinity chromatography. *Eur. J. Biochem.*, **92**, 373–379.

Miller,C. *et al.* (2011) Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Mol. Syst. Biol.*, **7**, 1–13.

Orphanides,G. and Reinberg,D. (2002) A unified theory of gene expression. *Cell*, **108**, 439–451.

Rabani,M. *et al*. (2011) Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat. Biotechnol.*, **29**, 436–442.

Rabani,M. *et al*. (2014) High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. *Cell*, **159**, 1698–1710.

Raghavan,A. *et al*. (2002) Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes. *Nucleic Acids Res.*, **30**, 5529–5538.

Shalem,O. *et al*. (2008) Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation. *Mol. Syst. Biol.*, **4**, 223.

Sun,M. *et al*. (2012) Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. *Genome Res.*, **22**, 1350–1359.

Sun,M. *et al*. (2013) Global analysis of EukaryoticmRNA degradation reveals Xrn1-dependent buffering of transcript levels. *Mol. Cell*, **52**, 52–62.

Zeisel,A. *et al*. (2011) Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. *Mol. Syst. Biol.*, **7**, 529.