Genome-wide dynamics of RNA synthesis, processing and degradation without RNA metabolic labeling

Supplemental Methods

M. Furlan, E. Galeota, N. Del Gaudio,
E. Dassi, M. Caselle, S. de Pretis, M. Pelizzola

Contents

1 Quantification of premature and mature RNA expression levels from RNA-seq data

The framework of INSPEcT analysis is based on the joint study of premature (P) and mature (M) RNA expression levels. Premature RNA can be defined as the ensemble of transcripts that requires additional structural modifications, i.e. splicing. Alternatively, it can be defined as the transcripts located in the nuclear compartment at the time of sequencing. INSPEcT exploits the structural information and quantifies P as the (length and library size normalized) read counts that overlap gene introns. Additionally, INSPEcT quantifies total RNA (T = M + P) as the (length and library size normalized) read counts that overlap gene exons. As a consequence of this, mature RNA descends from the difference between T and P.

The quantification procedure of INSPEcT is implemented by the embedded functions "quantifyExpressionsFromBAMs" (for BAM input files) or "quantifyExpressionsFromBWs" (for BigWig). The count of intronic and exonic reads is inherently an ambiguous task, mainly because of two issues: (i) the presence of multiple isoforms for the same gene, and (ii) the possible overlap between the annotation of different genes. To cope with the first issue, INSPEcT collapse the exons of transcripts belonging to the same gene and internally defines introns as the gaps between adjacent collapsed exons. Regarding the second issue, INSPEcT do not assign reads overlapping to multiple exonic or intronic features. Conversely, reads overlapping to both an intronic and an exonic feature are assigned to the exonic feature.
The following genomic annotations, in the form of R/Bioconductor Annotation packages were used throughout the text to retrieve exonic genomic coordinates in the analysis of time-course data:

- TxDb.Athaliana.BioMart.plantsmart28 (Arabidopsis Thaliana)
- TxDb.Hsapiens.UCSC.hg19.knownGene (Homo Sapiens)
- TxDb.Mmusculus.UCSC.mm9.knownGene (Mus Musculus)

For the steady-state analysis, we used the recount2 package function "recount_exons", which is based on the GRCh38 annotation.

1.1 Validation using fractionated RNA

We took advantage of fractionated RNA-seq data to validate how we determine premature and mature RNA expression based on intronic and exonic

signals of total RNA-seq data, by comparing them to nuclear and cytoplasmic RNA. To carry out these analyses, we re-analyzed a previously published dataset composed of nuclear, cytoplasmic and total RNA-seq data from a human fetal frontal cortex tissue sample [1]. We downloaded the BAM files released by the study (ArrayExpress database, accession E-MTAB-1898) and we quantified the gene counts for each RNA-seq library according to the standard procedure implemented in the INSPEcT package. Then, we checked the fraction of intronic reads in the three libraries (Supplemental Fig. 15A). As expected, a relevant fraction of whole-cell total RNA reads were intronic (43%). This percentage was even higher for nuclear RNA (58%), while it dropped to 14% in the cytoplasmic library.
We performed a ranked correlation study on more than 9000 genes, comparing the abundance of nuclear (N) versus premature (P) RNA, and of cytoplasmic (C) versus mature (M) RNA, where P and M were quantified through intronic and exonic counts in the total RNA-seq data, respectively. Spearman's correlation coefficients were $r_{P,N} = 0.75$, and $r_{M,C} = 0.88$ (Supplemental Fig. 15B). We repeated this analysis using partial correlations, to account for the information shared between nuclear and cytoplasmic RNA ($r_{N,C} = 0.86$). The resulting correlations were only partially reduced ($r_{P,N|C} = 0.42$, and $r_{M,C|N} = 0.74$), and were still highly significant ($P - values < 1e - 16$). Moreover, they were much higher that the correlations obtained comparing P versus C, and M versus N ($r_{P,C|N} = 0.13$, and $r_{M,N|C} = -0.17$). To conclude, this analysis shows that premature and mature RNA expression levels, estimated with whole-cell total RNA-seq intronic and exonic reads, are in good agreement with nuclear and cytoplasmic RNA expression levels respectively.

2 Mathematical modelling of RNA life-cycle (Main Figure 1)

We model the dynamics of premature (P) and mature (M) RNA according to the following well established schema

$$\emptyset \xrightarrow{k_1} P \xrightarrow{k_2} M \xrightarrow{k_3} \emptyset,$$

where k_1 and k_2 are the rates of synthesis and processing of the premature RNA, respectively, while k_3 is the rate of degradation of the mature RNA.
Using mass action kinetics, the above system translates in a system of two ordinary differential Equations:

$$\begin{cases} \dot{P}(t) = k_1(t) - k_2(t) \cdot P(t), & \text{(1a)} \\ \dot{M}(t) = k_2(t) \cdot P(t) - k_3(t) \cdot M(t). & \text{(1b)} \end{cases}$$

The system can also be trivially rewritten in terms of total RNA, by exploiting the linearity of the derivatives:

$$\begin{cases} \dot{P}(t) = k_1(t) - k_2(t) \cdot P(t), & \text{(2a)} \\ \dot{T}(t) = k_1(t) - k_3(t) \cdot (T(t) - P(t)). & \text{(2b)} \end{cases}$$

The model presented in Equation 1 can be also simplified to describe the steady-state of our process. In this condition, by definition, the time derivatives of premature and mature RNA are equal to zero ($\dot{P} = 0, \dot{M} = 0$), and all the kinetic rates are not changing over time:

$$\begin{cases} P = \dfrac{k_1}{k_2}, & \text{(3a)} \\ M = \dfrac{k_1}{k_3}. & \text{(3b)} \end{cases}$$

3 Time-course experimental design (Main Figure 3)

A general solution for the ODEs systems (Equations 1 and 2) is not possible without making assumptions on the functional forms of expression data and/or kinetic rates. Naturally, countless parameterizations could be used, and INSPEcT is restricted on few of them. The procedure of INSPEcT is divided into two steps:

1. The first step models k_1 as a piecewise linear, and k_2, k_3 as piecewise constant, all defined between consecutive experimental time points. This procedure, described in Section 3.1.2, exploits all the degrees of freedom of the time-course, fits perfectly to the experimental data, with the consequence of fitting also the experimental noise. Nonetheless, this procedure provides a fast solution to check for the quality of the input data. Moreover, the estimated rates are used to initialize the parameters of the second modeling step.

2. The second step can be achieved by means of three different implementations: (i) non-functional approach, described in Section 3.2.1, (ii) integrative functional approach, and (iii) derivative functional approach.

The aim of this step of modelling is to control the noise associated to the experimental data and to statistically assess which are the rates that are shaping the variation, if any, in premature and mature RNA. Each of these methods has a peculiar ability in describing the experimental data. Summarizing, the non-functional approach is able to detect the gene responses of any shape, as it maintains the piecewise parametrization, but is more affected by the noise than the other two. The integrative functional approach restricts the shape of k_1, k_2, and k_3 to either sigmoid or impulse functions. This approach requires to solve the ODE system by integrating it and is the most expensive in terms of computational time. The derivative functional approach restricts the shape of M, k_2, and k_3 to either sigmoid or impulse functions. Due to the fact that k_1 is not constrained to any a-priori functional shape, this approach could originate over complicated transcriptional responses. Despite this, the analysis based on simulated datasets did not show a marked reduction in the performance compared to the integrative functional approach. Conversely, this method is computationally less expensive and performs more than one order of magnitude faster (Supplemental Fig. 1). For these reasons, INSPEcT runs the derivative functional approach by default, and unless differently specified in the text it has been used to produce all time-course analysis of the paper.

In addition to what previously described, when INSPEcT runs without nascent RNA (INSPEcT-), there is an additional step prior to all others. This step, which is described in Section 3.1.1, models constant k_2 and k_3 and it is necessary to supply the lack of the nascent RNA information.

General notes The $\hat{X}$ symbol will identify the experimental data of the variable X, the σ_X the associated standard deviations, the i index a specific time point between 1 and n. In the following sections, we will refer to many numerical optimizations of different cost functions. We always perform them exploiting the *R* functions *optimize* and *optim* which are implemented in the built-in *stat* package and perform univariate and multivariate optimization, respectively.

3.1 First step of modelling

3.1.1 Constant post-transcriptional rates

As anticipated in the introduction of Section 3, the method described here is applied only in the absence of nascent RNA (INSPEcT-). In this case,

the routine provides an initial guess of $k_1(t)$, assuming that k_2 and k_3 are constant throughout the entire time-course. Despite these assumptions are probably too strict for a considerable fraction of genes, they allow to solve the Equation 1b and express M in function of the time t, and two unknowns k_2 and k_3.

$$
\begin{aligned}
&\dot{M}(t) = k_2 \cdot P(t) - k_3 \cdot M(t), \\
&\dot{M}(t) + k_3 \cdot M(t) = k_2 \cdot P(t), \\
&e^{k_3 \cdot t} \cdot \left(\dot{M}(t) + k_3 \cdot M(t)\right) = e^{k_3 \cdot t} \cdot k_2 \cdot P(t), \\
&d_t\left(M(t) \cdot e^{k_3 \cdot t}\right) = k_2 \cdot P(t) \cdot e^{k_3 \cdot t}, \\
&\int_{t_i}^{t_{i+1}} d_t\left(M(t) \cdot e^{k_3 \cdot t}\right) dt = \int_{t_i}^{t_{i+1}} k_2 \cdot P(t) \cdot e^{k_3 \cdot t} dt, \\
&M(t) \cdot e^{k_3 \cdot t}\Big|_{t_i}^{t_{i+1}} = k_2 \cdot \int_{t_i}^{t_{i+1}} P(t) \cdot e^{k_3 \cdot t} dt.
\end{aligned}
$$

The solution of Equation 1b requires the integration of $P(t)$ along the time series $\{t_1, \ldots, t_n\}$, meaning that $P(t)$ should be expressed in functional form. We chose to model $P(t)$ as a piecewise linear function, defined at each pair of consecutive experimental observations:

$$
\begin{cases}
P(t) = a_i + b_i \cdot t, & \text{(4a)} \\
a_i = \widehat{P}(t_i) \cdot b_i \cdot t_i, \qquad \forall \;\; t \in [t_i, t_{i+1}] \wedge i \in \{1, \ldots, n-1\} & \text{(4b)} \\
b_i = \dfrac{\widehat{P}(t_i) - \widehat{P}(t_{i+1})}{t_i - t_{i+1}}, & \text{(4c)}
\end{cases}
$$

allowing to complete the solution of Equation 1b for a generic interval defined by t_i and t_{i+1}:

$$
\begin{aligned}
M(t) \cdot e^{k_3 \cdot t}\Big|_{t_i}^{t_{i+1}} &= k_2 \cdot \int_{t_i}^{t_{i+1}} (a_i + b_i \cdot t) \cdot e^{k_3 \cdot t} dt, \\
M(t) \cdot e^{k_3 \cdot t}\Big|_{t_i}^{t_{i+1}} &= \frac{k_2 \cdot a_i}{k_3} \cdot e^{k_3 \cdot t}\Big|_{t_i}^{t_{i+1}} + k_2 \cdot b_i \cdot \int_{t_i}^{t_{i+1}} t \cdot e^{k_3 \cdot t} dt, \\
M(t) \cdot e^{k_3 \cdot t}\Big|_{t_i}^{t_{i+1}} &= \frac{k_2 \cdot a_i}{k_3} \cdot e^{k_3 \cdot t}\Big|_{t_i}^{t_{i+1}} + k_2 \cdot b_i \cdot \left(t \cdot \frac{e^{k_3 \cdot t}}{k_3}\Big|_{t_i}^{t_{i+1}} - \frac{e^{k_3 \cdot t}}{k_3^2}\Big|_{t_i}^{t_{i+1}} \right),
\end{aligned}
$$

which yields to

$$M(t_{i+1}) = M(t_i) \cdot e^{-k_3 \cdot (t_{i+1} - t_i)} + \frac{k_2 \cdot a_i}{k_3} \cdot \left(1 - e^{-k_3 \cdot (t_{i+1} - t_i)}\right) + \frac{k_2 \cdot b_i}{k_3} \cdot \left[\left(t_{i+1} - t_i \cdot e^{-k_3 \cdot (t_{i+1} - t_i)}\right) - \frac{1}{k_3} \cdot \left(1 - e^{-k_3 \cdot (t_{i+1} - t_i)}\right)\right]. \quad (5)$$

The dependence of $M\left(t_{i+1}\right)$ from $M\left(t_i\right)$ is solved recursively, starting from the experimental value $\widehat{M}\left(t_0\right)$. Moreover, applying the post-transcriptional ratio formula (Equation 42) to the steady-state observations of P and M at time $t = t_0$:

$$\frac{P(t_0)}{M(t_0)} = \frac{k_3}{k_2}, \quad (6)$$

it is possible to express M in Equation 5 solely in terms of the time t and a single unknown (either k_2 or k_3). To facilitate the inference of the model parameters, we perform two distinct optimizations of the mature RNA cost function (standard χ^2): (i) the first one is unidimensional and regards k_3 only, while k_2 is expressed by means of Equation 6, (ii) the second one is performed on the bi-dimensional k_2, k_3 space, using the values estimated in the previous step as a starting point for the minimization of the χ^2.

After the estimation of constant k_2 and k_3, we compute for each gene $k_1(t)$ exploiting the optimized k_2 and Equation 1a:

$$k_1\left(t\right) = \dot{P}\left(t\right) + k_2 \cdot P\left(t\right), \quad (7)$$

where the function $\dot{P}(t)$ is approximated by fitting cubic splines to the experimental data.

3.1.2 Piecewise constant post-transcriptional rates

The aim of the procedure described in this section is to estimate $k_2(t)$ and $k_3(t)$ as piecewise constant functions, with intervals defined at the boundaries of the experimental time points:

$$\begin{aligned} k_2(t) &= k_{2_i} \quad \forall \;\; t \in (t_i, t_{i+1}] \wedge i \in \{1, \ldots, n-1\}, \\ k_3(t) &= k_{3_i} \quad \forall \;\; t \in (t_i, t_{i+1}] \wedge i \in \{1, \ldots, n-1\}. \end{aligned} \quad (8)$$

$\widehat{M}(t_i)$, $\widehat{P}(t_i)$, and $\widehat{k_1}(t_i)$ are required ($i \in \{1, \ldots, n\}$). In case of INSPEcT+, the quantification of $\widehat{k_1}(t_i)$ is available through of the measure of nascent

RNA, as will be described in Equations 43 and 44. In case of INSPEcT-, $\widehat{k_1}(t_i)$ is obtained through the procedure described in Section 3.1.1.

In order to estimate $k_2(t)$, we define $k_1(t)$ as piecewise linear:

$$\begin{cases} k_1(t) = c_i + d_i \cdot t, & \text{(9a)} \\ c_i = \widehat{k_1}(t_i) \cdot d_i \cdot t_i, \qquad \forall\ t \in [t_i, t_{i+1}] \wedge i \in \{1, \ldots, n-1\} & \text{(9b)} \\ d_i = \dfrac{\widehat{k_1}(t_i) - \widehat{k_1}(t_{i+1})}{t_i - t_{i+1}}. & \text{(9c)} \end{cases}$$

Following this, it is possible to solve the Equation 1a as follows:

$$\begin{aligned}
&\dot{P}(t) = k_1(t) - k_{2_i} \cdot P(t), \\
&\dot{P}(t) + k_{2_i} \cdot P(t) = k_1(t), \\
&e^{k_{2_i} \cdot t} \cdot \left(\dot{P}(t) + k_{2_i} \cdot P(t)\right) = e^{k_{2_i} \cdot t} \cdot k_1(t), \\
&d_t\left(P(t) \cdot e^{k_{2_i} \cdot t}\right) = k_1(t) \cdot e^{k_{2_i} \cdot t}, \\
&\int_{t_i}^{t_{i+1}} d_t\left(P(t) \cdot e^{k_{2_i} \cdot t}\right) dt = \int_{t_i}^{t_{i+1}} k_1(t) \cdot e^{k_{2_i} \cdot t} dt, \\
&P(t) \cdot e^{k_{2_i} \cdot t}\Big|_{t_i}^{t_{i+1}} = \int_{t_i}^{t_{i+1}} (c_i + d_i \cdot t) \cdot e^{k_{2_1} \cdot t} dt, \\
&P(t) \cdot e^{k_{2_i} \cdot t}\Big|_{t_i}^{t_{i+1}} = \frac{c_i}{k_{2_i}} \cdot e^{k_{2_i} \cdot t}\Big|_{t_i}^{t_{i+1}} + d_i \cdot \int_{t_i}^{t_{i+1}} t \cdot e^{k_{2_i} \cdot t} dt, \\
&P(t) \cdot e^{k_{2_i} \cdot t}\Big|_{t_i}^{t_{i+1}} = \frac{c_i}{k_{2_i}} \cdot e^{k_{2_i} \cdot t}\Big|_{t_i}^{t_{i+1}} + d_i \cdot \left(t \cdot \frac{e^{k_{2_i} \cdot t}}{k_{2_i}}\Big|_{t_i}^{t_{i+1}} - \frac{e^{k_{2_i} \cdot t}}{k_{2_i}^2}\Big|_{t_i}^{t_{i+1}} \right),
\end{aligned}$$

which yields to

$$\begin{aligned}
P(t_{i+1}) = &P(t_i) \cdot e^{-k_{2_i} \cdot (t_{i+1} - t_i)} + \frac{c_i}{k_{2_i}} \cdot \left(1 - e^{-k_{2_i} \cdot (t_{i+1} - t_i)}\right) + \\
&\frac{d_i}{k_{2_i}} \cdot \left[\left(t_{i+1} - t_i \cdot e^{-k_{2_i} \cdot (t_{i+1} - t_i)}\right) - \left(\frac{1 - e^{-k_{2_i} \cdot (t_{i+1} - t_i)}}{k_{2_i}} \right) \right]. \qquad \text{(10)}
\end{aligned}$$

The dependence of $P(t_{i+1})$ from $P(t_i)$ is solved recursively, starting from the experimental value $\widehat{P}(t_0)$. At each time interval, k_{2_i} is found as the one that minimizes of the χ^2 error with $\widehat{P}(t_{i+1})$.

In order to estimate $k_3(t)$, we define both $k_1(t)$ and $P(t)$ as piecewise linear (Equations 4 and 9). Following this, we solve Equation 2b for a generic

interval defined by t_i and t_{i+1}, assuming piecewise constant k_2 and k_3 (Equation 8):

$$
\begin{aligned}
&\dot{T}(t) = k_1(t) - k_{3_i} \cdot (T(t) - P(t)),\\
&\dot{T}(t) + k_{3_i} \cdot T(t) = k_1(t) + k_{3_i} \cdot P(t),\\
&e^{k_{3_i} \cdot t} \cdot \left(\dot{T}(t) + k_{3_i} \cdot T(t)\right) = e^{k_{3_i} \cdot t} \cdot (k_1(t) + k_{3_i} \cdot P(t)),\\
&d_t\left(T(t) \cdot e^{k_{3_i} \cdot t}\right) = (k_1(t) + k_{3_i} \cdot P(t)) \cdot e^{k_{3_i} \cdot t},\\
&\int_{t_i}^{t_{i+1}} d_t\left(T(t) \cdot e^{k_{3_i} \cdot t}\right) dt = \int_{t_i}^{t_{i+1}} (k_1(t) + k_{3_i} \cdot P(t)) \cdot e^{k_{3_i} \cdot t} dt,\\
&T(t) \cdot e^{k_{3_i} \cdot t}\Big|_{t_i}^{t_{i+1}} = \int_{t_i}^{t_{i+1}} (c_i + d_i \cdot t) \cdot e^{k_{3_1} \cdot t} dt + k_{3_i} \cdot \int_{t_i}^{t_{i+1}} (a_i + b_i \cdot t) \cdot e^{k_{3_1} \cdot t} dt,\\
&T(t) \cdot e^{k_{3_i} \cdot t}\Big|_{t_i}^{t_{i+1}} = \frac{(c_i + k_{3_i} \cdot a_i)}{k_{3_i}} \cdot e^{k_{3_i} \cdot t}\Big|_{t_i}^{t_{i+1}} + (d_i + k_{3_i} \cdot b_i) \cdot \int_{t_i}^{t_{i+1}} t \cdot e^{k_{3_1} \cdot t} dt,\\
&T(t) \cdot e^{k_{3_i} \cdot t}\Big|_{t_i}^{t_{i+1}} = \frac{(c_i + k_{3_i} \cdot a_i)}{k_{3_i}} \cdot e^{k_{3_i} \cdot t}\Big|_{t_i}^{t_{i+1}} + (d_i + k_{3_i} \cdot b_i) \cdot \left(t \cdot \frac{e^{k_{3_i} \cdot t}}{k_{3_i}}\Big|_{t_i}^{t_{i+1}} - \frac{e^{k_{3_i} \cdot t}}{k_{3_i}^2}\Big|_{t_i}^{t_{i+1}}\right),
\end{aligned}
$$

which yields to

$$
\begin{aligned}
T(t_{i+1}) =& T(t_i) \cdot e^{-k_{3_i} \cdot (t_{i+1} - t_i)} + \left(\frac{c_i}{k_{3_i}} + a_i\right) \cdot \left(1 - e^{-k_{3_i} \cdot (t_{i+1} - t_i)}\right) +\\
& \left(\frac{d_i}{k_{3_i}} + b_i\right) \cdot \left[\left(t_{i+1} - t_i \cdot e^{-k_{3_i} \cdot (t_{i+1} - t_i)}\right) - \left(\frac{1 - e^{-k_{3_i} \cdot (t_{i+1} - t_i)}}{k_{3_i}}\right)\right].
\end{aligned} \tag{11}
$$

The dependence of $T(t_{i+1})$ from $T(t_i)$ is solved recursively, starting from the experimental value $\widehat{T}(t_0)$. At each time interval, k_{3_i} is found as the one that minimizes of the χ^2 error with $\widehat{T}(t_{i+1})$.

Linear approximation of premature RNA In the solution of Equation 11, we imposed that $k_1(t)$ and $P(t)$ are both linear functions defined independently for each time window. These assumptions are compatible with our model as we can easily prove substituting the two lines in equation 1a. However, nothing guarantees that a linear solution of equation 1a provides the best fit of the experimental data. Indeed, this particular solution induces some constrains on the parameters of $P(t)$ and disregards the exponential terms when Equation 10 is plugged into Equation 11. We tested the goodness of our particular solution (linear) comparing it against the most general

one (exponential) on a set of simulated data. Specifically, we inferred the best k_3 for both the exponential and linear frameworks minimizing the error of the model defined as,

$$\left| 1 - \frac{T(t_{i+1})}{\widehat{T}(t_{i+1})} \right|. \tag{12}$$

The distributions of this quantity in the two modeling frameworks resulted very similar and much lower than the same error estimated at one standard deviation (the medians were 4 orders of magnitude far). We concluded that the linear solution provides a good fit of the data very similar to the one obtained including the exponential term. Then, we compared the optimum degradation rates for the linear solution against the counterpart estimated taking into account the exponential term and we found very high concordance (Spearman's correlation coefficient = 1) except for few conditions characterized by very slow processing rates.
We concluded that the differences between linear and exponential solutions are negligible from the practical point of view.

3.2 Second step of modelling

As anticipated in the introduction of Section 3, the aim of the second step of modelling is to control the noise associated to the experimental data and to statistically assess which are the rates that are shaping the variation, if any, in premature and mature RNA.

3.2.1 Non-functional approach

In this approach, we extend the rates estimated in the first step by calculating their confidence intervals and we identify, for each gene, the most probable regulatory scenario. To do that, we developed a profile likelihood based technique.

Maximum Likelihood Estimation of model parameters Kinetic rates estimated as described in Section 3.1.2 are further optimized in order to maximize the Log-Likelihood of the model.
In case if INSPEcT-, we compute this quantity by comparing premature and total RNA estimated from Equations 10 and 11, expressed as a function of

the parameters (θ), and their experimental counterparts $\widehat{P}$ and $\widehat{T}$.

$$
\begin{aligned}
&f(\theta) \in \{P_1, \ldots, P_n, T_1, \ldots, T_n\},\\
&\widehat{x} \in \{\widehat{P}_1, \ldots, \widehat{P}_n, \widehat{T}_1, \ldots, \widehat{T}_n\},\\
&\sigma_x \in \{\sigma_{P_1}, \ldots, \sigma_{P_n}, \sigma_{T_1}, \ldots, \sigma_{T_n}\},
\end{aligned}
$$

In case of INSPEcT+, we add to the estimation of likelihood also the comparison between k_1 (estimated from Equations 43 and 44) and $\widehat{k_1}$:

$$
\begin{aligned}
&f(\theta) \in \{P_1, \ldots, P_n, T_1, \ldots, T_n, k_{1_1}, \ldots, k_{1_n}\},\\
&\widehat{x} \in \{\widehat{P}_1, \ldots, \widehat{P}_n, \widehat{T}_1, \ldots, \widehat{T}_n, \widehat{k}_{1_1}, \ldots, \widehat{k}_{1_n}\},\\
&\sigma_x \in \{\sigma_{P_1}, \ldots, \sigma_{P_n}, \sigma_{T_1}, \ldots, \sigma_{T_n}, \sigma_{k_{1_1}}, \ldots, \sigma_{k_{1_n}}\},
\end{aligned}
$$

Following this, we define the Likelihood function:

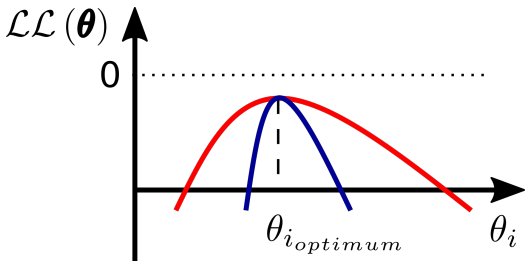
$$
\mathcal{L}\left(\theta_j|\widehat{x}_j, \sigma_{x_j}\right) = 2 \cdot \int_{|f(\theta)-\widehat{x}_j|}^{\infty} \mathcal{N}\left(h|\mu = 0, \sigma = \sigma_{x_j}\right) dh \quad \forall j \in \{1, \ldots, k\}, \tag{13}
$$

and its cumulative logarithmic counterpart:

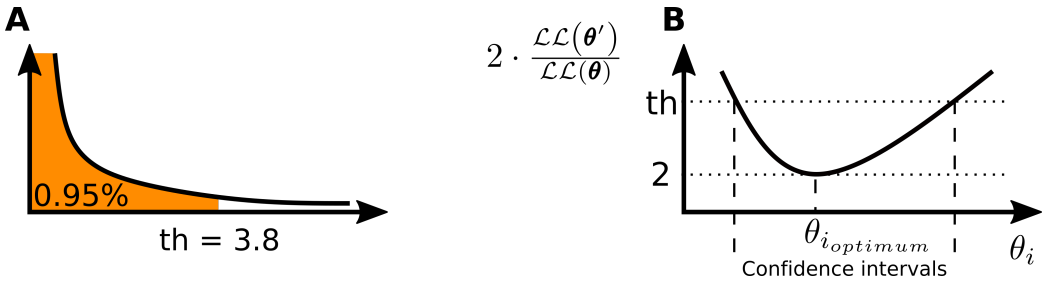
$$
\mathcal{LL}\left(\theta|\widehat{x}, \sigma_x\right) = \sum_{j=1}^{k} ln\left(\mathcal{L}\left(\theta_j|\widehat{x}_j, \sigma_{x_j}\right)\right). \tag{14}
$$

where k is the number of experimental observations of the model, i.e. $2 \cdot n$ in case of INSPEcT- and $3 \cdot n$ in case of INSPEcT+.

Log-Likelihood based confidence intervals The Log-Likelihood function (Equation 14) is defined in the space of the model parameters. In the proximity of a maximum is a concave function, meaning that any perturbation of the parameters results in a decrease of the Log-Likelihood. Parameters of the model whose perturbation cause a small reduction of the Log-Likelihood can be considered uncertain. Conversely, parameters whose perturbations cause a sharp reduction of the Log-Likelihood are considered well characterized (Method Fig. 1). In order to estimate rigorous confidence intervals, we exploited the Log-Likelihood ratio test [2]. This method compares two nested models, which differ in the number of parameters, assessing whether the more complex model explains the data better than the simpler one. In our set-up, we consider the model corresponding to the maximum



Method Figure 1: **Different shapes of the Log-Likelihood function according to the precision of the perturbed parameter θ_i.** Cartoon showing the Log-Likelihood curves for a parameter well (blue) and approximately (red) characterized.



Method Figure 2: **Confidence intervals definition.** Cartoon showing the method we adopt to define the confidence intervals. Box A shows the identification of the significance threshold while box B shows its application.

likelihood as the more complex, and the perturbed version obtained changing the value of a single parameter the simpler ($\theta_i \rightarrow \theta'_i$), with one parameter less. From the theory of Log-Likelihood ratio test we know that

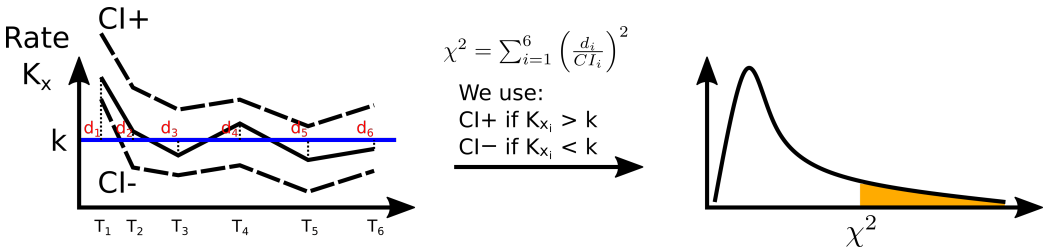
$$2 \cdot \frac{\mathcal{LL}\left(\boldsymbol{\theta'_i}\right)}{\mathcal{LL}\left(\boldsymbol{\theta_i}\right)} \sim \chi^2\left(1\right). \tag{15}$$

Therefore we seek for θ'_i that comprehend the 95% of the $\chi^2(1)$ distribution ($th \approx 3.8$) in both directions (left C.I. $\theta'_i < \theta_i$, right C.I. $\theta'_i > \theta_i$, Method Fig. 2).

Selection of the regulatory scenario The confidence intervals are exploited to identify the most probable regulatory scenario, by testing independently the variability of k_1, k_2, and k_3. The parameters of our model correspond to the kinetic rate values estimated at each experimental time interval:

$$\theta = \{k_{1_1}, \ldots, k_{1_n}, k_{2_1}, \ldots, k_{2_n}, k_{3_1}, \ldots, k_{3_n}\}). \tag{16}$$

For each kinetic rate, we optimize a constant model minimizing its squared distance from each data point normalized over the associated confidence interval and we interpret the cost function as a $\chi^2(n-1)$. Eventually, we use the expected distribution to compute a p-value for the model, which is corrected for multiple testing (R stat p.adjust function, BH method) and used to decide if the null hypothesis should be rejected (P-value < 0.05, variable rate) or not.
In case of absence of nascent RNA (INSPEcT-), the variability of the rates estimated in Section 3.1.2 is highly dependent upon the order through which they are estimated. For instance, the $k_1(t)$ is the first rate whose variability is estimated and explains the largest part of the variance of $P(t)$. The variance of $M(t)$ left unexplained by $k_1(t)$ can be absorbed only by $k_3(t)$, therefore $k_2(t)$ is usually largely constant. For this reason, we devised a second set of piecewise kinetic rates, whose order of computation and variability are different. For instance, k_2 is the first rate whose variability is estimated and explains the largest part of the variance of $P(t)$. Following this, $k_1(t)$ and $k_2(t)$ are estimated. The model with the lowest cumulative Log-fold changes of the kinetic rates calculated over their value at time zero, i.e. the simpler one, is selected and confidence intervals and model selection are performed.



Method Figure 3: **Identification of the most probable transcriptional and post-transcriptional regulatory scenario.** Cartoon showing the confidence intervals fit to test the variability of a given rate. We start from the definition of the confidence intervals with parameters perturbation (left) and we use this information to fit the profile with a constant minimizing a χ^2 function which is then used to estimate e p-value to accept or reject the constant null hypothesis (right).

3.2.2 Functional approaches

In addition to the non-functional approach (described in Section 3.2.1), INSPEcT provides an alternative modelling framework based on smooth functional forms. As for the non-functional approach, the aim of functional ap-

proaches is to reduce the impact of noise associated to the experimental data and identify a regulatory scenario.

INSPEcT provides two ways of solving the ODE system exploiting smooth functional forms, named integrative and derivative functional approaches. Each of them applies with different variants when applied to a dataset with or without nascent RNA, i.e. INSPEcT+ or INSPEcT-, and will be described in a dedicated paragraph. As introduced in Section 3, the integrative functional approach restricts the shape of k_1, k_2, and k_3 to one of the adopted smooth functions, while the derivative functional approach restricts the shape of M, k_2, and k_3. As a consequence of this, the integrative approach requires to solve the ODE system by integrating it and is the most expensive in terms of computational time, while the derivative functional approach has an analytical solution and can be solved faster. The main difference between the application of these approaches to INSPEcT+ and INSPEcT- resides in the choice of the regulatory scenario. While INSPEcT+ exploits the confidence intervals of model parameters, INSPEcT- exploits the fit of models with different degrees of freedom and performs model selection.

In all cases, the adopted functional forms are constants (Equation 17), sigmoids (Equation 18 and Method Fig. 4) or impulses functions (Equation 19 and Method Fig. 4).

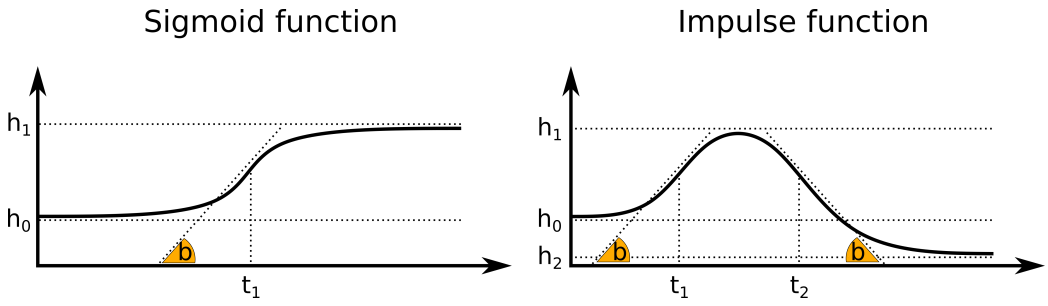
$$constant\,(t, k) = k, \tag{17}$$

$$sigmoid\,(t, h_0, h_1, t_1, b) = h_0 + \frac{h_1 - h_0}{1 + e^{-b\cdot(t-t_1)}}, \tag{18}$$

$$impulse\,(t, h_0, h_1, h_2, t_1, t_2, b) = \frac{1}{h_1} \cdot \left(h_0 + \frac{h_1 - h_0}{1 + e^{-b\cdot(t-t_1)}}\right) \cdot \left(h_2 + \frac{h_1 - h_2}{1 + e^{b\cdot(t-t_2)}}\right). \tag{19}$$

These functions are expected to recapitulate biological responses and are widely used in the field of expression dynamics [3] [4] [5] [6].

Integrative modelling with nascent RNA (INSPEcT+) For each gene in the dataset, we estimate the mean of k_{1_i}, k_{2_i} and k_{3_i} calculated in the first step of modelling (Section 3.1) and we optimize them in order to fit the steady-state solution presented in Equations 3a, 3b. Additionally, the experimental value of $\widehat{k_1}(t)$ is calculated as will be described in Equation 45.



Method Figure 4: **Sigmoid and impulse functions.** Cartoon showing an example of sigmoid (left) and impulse (right), in this picture, we highlighted the role of each parameter in the definition of the shape of the two curves.

We minimize the cumulative χ^2 function:

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(P(t=t_i) - \widehat{P_i}\right)^2}{\sigma_{P_i}^2} + \frac{\left(M(t=t_i) - \widehat{M_i}\right)^2}{\sigma_{M_i}^2} + \frac{\left(k_1(t=t_i) - \widehat{k_{1i}}\right)^2}{\sigma_{k_{1_i}}^2}. \tag{20}$$

The optimized steady-state rates $k_{1_{SS}}$, $k_{2_{SS}}$ and $k_{3_{SS}}$ are then used as the initial condition for two new optimizations that involve non steady-state rates, i.e. $k_1(t, \theta_{k_1})$, $k_2(t, \theta_{k_2})$ and $k_3(t, \theta_{k_3})$. In one case the rates are fitted by sigmoids functions:

$$k_1(t, \theta_{k_1}) \sim k_2(t, \theta_{k_2}) \sim k_3(t, \theta_{k_3}) \sim sigmoid(t, h_0, h_1, t_1, b), \tag{21}$$

in the other by impulse functions:

$$k_1(t, \theta_{k_1}) \sim k_2(t, \theta_{k_2}) \sim k_3(t, \theta_{k_3}) \sim impulse(t, h_0, h_1, h_2, t_1, t_2, b) \tag{22}$$

In particular, we parametrize each rate with a sigmoid or impulse function initially defined as flat curve equal to the corresponding steady-state rate. For sigmoids h_0 and h_1 are initially set equal to the steady-state rates value ($k_{x_{SS}}$), t_1 to $(t_n - t_0)/2$ and b to 1. For impulses, h_0, h_1 and h_2 are initially set equal to the steady-state rates value ($k_{x_{SS}}$), t_1 to $(t_n - t_0) \cdot 1/3$, t_2 to $(t_n - t_0) \cdot 2/3$, and b to 1. Then, we optimize all the parameters of the sigmoid or impulse to minimize the χ^2 introduced in Equation 20. Within the integrative functional framework an analytical solution of the ODEs system presented in Equations 1a and 1b is not possible, therefore $P(t)$ and $M(t)$ are estimated by numerical

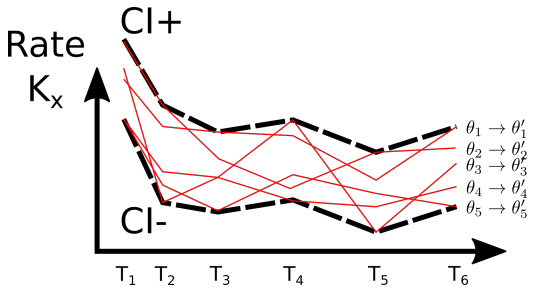
integration:

$$\begin{cases} P(t) = \dfrac{k_1(0,\theta_1)}{k_2(0,\theta_2)} + \displaystyle\int_0^t k_1(x,\theta_1)\cdot dx - \int_0^t k_2(x,\theta_2)\cdot P(x)\cdot dx, \\ M(t) = \dfrac{k_1(0,\theta_1)}{k_3(0,\theta_3)} + \displaystyle\int_0^t k_2(x,\theta_2)\cdot P(x)\cdot dx - \int_0^t k_3(x,\theta_3)\cdot M(x)\cdot dx \end{cases} \tag{23}$$

using the *ode* function of the *deSolve* R-package. Eventually, the p-value of the sigmoidal and impulsive models are calculated assuming $2 \cdot n - 12$ and $2 \cdot n - 18$ degrees of freedom, respectively. The model with the lower associated p-value is chosen.
In order assess which rates are shaping the changes in premature and mature RNA, we evaluate the confidence intervals of the model parameters, as described in Section 3.2.1. In this case, the model parameters do not correspond directly to kinetic rates, but are the parameters of sigmoid or impulse function. To propagate the uncertainties from the model parameters to the corresponding kinetic rate, we identify the portion of space, per each kinetic rate, containing all the trajectories computed perturbing the model parameters within their 95% confidence interval (Method Fig. 5). Once obtained the confidence intervals of each kinetic rate, we fit a linear function, as described in Method Fig. 3, to estimate a p-value that we use, after correction for multiple testing, to decide if the constant rate hypothesis should be rejected (adjusted p-value< 0.05) or not.

Derivative modelling with nascent RNA (INSPEcT+) For each gene in the dataset, we fit mature RNA expression level both with a sigmoid and an impulse model. We optimize each fit by minimizing a standard χ^2 cost function and select the model with the minimum associated p-value. The same functional form selected for $M(t)$ is also used to define the dynamics of $k_2(t)$ and $k_3(t)$, which are initially defined as flat curves equal to the mean rate estimated in the first step of modelling (see previous paragraph for more specifications). Following this, we optimize all the parameters of the sigmoid or impulse to minimize the χ^2 introduced in Equation 20. In order to define the χ^2 function, $P(t)$ is obtained through Equation 1b:

$$P\left(t\right) = \frac{\dot{M}\left(t\right) + k_3\left(t\right)\cdot M\left(t\right)}{k_2\left(t\right)}. \tag{24}$$



Method Figure 5: **Uncertainties propagation.** Cartoon showing the idea behind the propagation of the uncertainty from the parameters we use to describe a kinetic rate to its numerical values along the time-course. Each solid line represents a trajectory obtained perturbing one parameter of the function used to describe the generic rate k_x ($\theta_i \rightarrow \theta_i'$). We define the rate's confidence intervals as the smallest portion of space which includes all the trajectories (black dashed lines).

Analogously, $k_1(t)$ is obtained through Equation 1a:

$$k_1(t) = d_t\left(\frac{\dot{M}(t) + k_3(t)\cdot M(t)}{k_2(t)}\right) + \dot{M}(t) + k_3(t)\cdot M(t),$$

$$k_1(t) = \frac{\ddot{M}(t)}{k_2(t)} + \dot{M}(t)\cdot\left(1 + \frac{k_3(t)}{k_2(t)} - \frac{\dot{k_2}(t)}{k_2^2(t)}\right) + M(t)\cdot\left(\frac{\dot{k_3}(t)}{k_2(t)} - \frac{\dot{k_2}(t)\cdot k_3(t)}{k_2^2(t) + k_3(t)}\right). \quad (25)$$

The selection of the regulatory scenario follows the same rationale presented for the integrative approach.
As previously mentioned, the major improvement of the derivative approach on the integrative one is the reduction of the computational cost of models optimization due to the elimination of any numerical solution of the ODEs system. However, a drawback of such a speed-up is the introduction of first and second orders time derivatives in the χ^2 computation, which results in an increase of its complexity. For instance, we report the first and second derivatives of the sigmoid and impulse functions (Equations 26, 27, 28 and 29). During the optimization via the Nelder-Mead procedure, configurations of parameters that do not returns finite values for $M(t)$, $P(t)$ and $k_1(t)$ are

discarded and the optimization proceed with a new set of parameters.

$$\frac{\mathrm{d}\ \mathrm{sigmoid}\,(t,h_0,h_1,t_1,b)}{\mathrm{dt}} = \frac{h_1-h_0}{h_1}\cdot\frac{b}{\frac{1}{e^{-b\cdot(t-t_1)}}+2+e^{-b\cdot(t-t_1)}}, \tag{26}$$

$$\frac{\mathrm{d}^2\mathrm{sigmoid}\,(t,h_0,h_1,t_1,b)}{\mathrm{dt}^2} = \frac{2\cdot b^2\cdot(h_1-h_0)\cdot e^{-2\cdot b\cdot(t-t_1)}}{(e^{-b\cdot(t-t_1)}+1)^3} - \frac{b^2\cdot(h_1-h_0)\cdot e^{-b\cdot(t-t_1)}}{(e^{-b\cdot(t-t_1)}+1)^2}, \tag{27}$$

$$\begin{aligned}\frac{\mathrm{d}\ \mathrm{impulse}\,(t,h_0,h_1,h_2,t_1,t_2,b)}{\mathrm{dt}} = \frac{1}{h_1}\cdot\Bigg[&\left((h_1-h_0)\cdot\frac{b}{\frac{1}{e^{-b\cdot(t-t_1)}}+2+e^{-b\cdot(t-t_1)}}\right)\cdot\\ &\cdot\left(h_2+(h1-h_2)\cdot\frac{1}{1+e^{b\cdot(t-t_2)}}\right)+\\ &+\left(h_0+(h1-h_0)\cdot\frac{1}{1+e^{-b\cdot(t-t_1)}}\right)\cdot\\ &\cdot\left((h_1-h_2)\cdot\frac{-b}{\frac{1}{e^{b\cdot(t-t_2)}}+2+e^{b\cdot(t-t_2)}}\right)\Bigg],\end{aligned} \tag{28}$$

$$\begin{aligned}\frac{\mathrm{d}^2\mathrm{impulse}\,(t,h_0,h_1,h_2,t_1,t_2,b)}{\mathrm{dt}^2} = &-\frac{2\cdot b^2\cdot(h1-h0)\cdot(h1-h2)\cdot e^{b\cdot(t-t2)-b\cdot(t-t1)}}{h1\cdot\left(e^{-b\cdot(t-t1)}+1\right)^2\cdot\left(e^{b\cdot(t-t2)}+1\right)^2}+\\ &+\frac{1}{h_1}\cdot\left(\frac{2\cdot b^2\cdot e^{2\cdot b\cdot(t-t2)}}{\left(e^{b\cdot(t-t2)}+1\right)^3}-\frac{b^2\cdot e^{b\cdot(t-t2)}}{\left(e^{b\cdot(t-t2)}+1\right)^2}\right)\cdot\\ &\cdot\left(\frac{(h_1-h_2)\cdot(h1-h0)}{e^{-b\cdot(t-t1)}+1}+h0\right)+\\ &+\frac{1}{h_1}\cdot\left(\frac{2\cdot b^2\cdot e^{-2\cdot b\cdot(t-t1)}}{\left(e^{-b\cdot(t-t1)}+1\right)^3}-\frac{b^2\cdot e^{-b\cdot(t-t1)}}{\left(e^{-b\cdot(t-t1)}+1\right)^2}\right)\cdot\\ &\cdot\left(\frac{(h1-h0)\cdot(h1-h2)}{e^{b\cdot(t-t2)}+1}+h2\right)\end{aligned} \tag{29}$$

Integrative modelling without nascent RNA (INSPEcT-) Without an experimental measure of nascent RNA, models where all the kinetic rates are shaped as sigmoid or impulse functions, as commonly done in INSPEcT+, are overcomplicated and lead in several cases to data overfitting. In fact, the χ^2 cost function of INSPEcT- is simpler than the one of INSPEcT+

(Equation 20), as it does not include the fit of the synthesis rate profile:

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(P(t=t_i) - \widehat{P}_i\right)^2}{\sigma^2_{P_i}} + \frac{\left(M(t=t_i) - \widehat{M}_i\right)^2}{\sigma^2_{M_i}}. \tag{30}$$

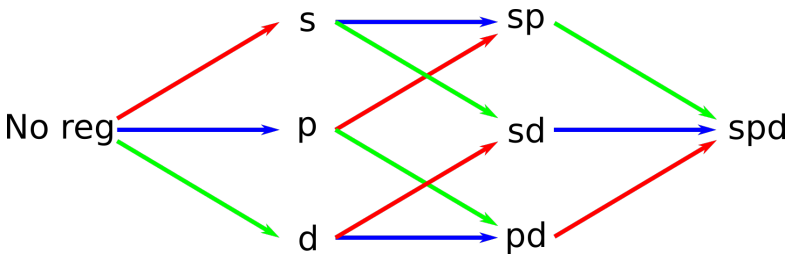
To solve this problem, INSPEcT- fits the $\widehat{P}$ and $\widehat{M}$ experimental profiles of each gene with eight models of different complexity, which represent all the possible combination of each kinetic rate being constant or variable (Method Fig. 6). At the end of the procedure, the model with the best trade-off between fit quality and simplicity is chosen to describe the kinetics of the gene.

For each gene in the dataset, we estimate the mean of k_{1_i}, k_{2_i} and k_{3_i} calculated in the first step of modelling (Section 3.1) and we optimize them in order to fit the steady-state solution presented in Equations 3a, 3b, through the minimization of the cumulative χ^2 function described above. This represent the simplest model (No-reg in Method Fig. 6) and its optimized rates ($k_{x_{SS}}$) are used to initialize the parameters of the following models. Remaining models are fitted to $\widehat{P}$ and $\widehat{M}$ profiles, through the minimization of the cumulative χ^2 function. As for INSPEcT+, we parametrize non-constant rates with both a sigmoid and an impulse functions and we select the best model according to the minimum χ^2 p-value. For sigmoids h_0 and h_1 are initially set equal to the steady-state rates value ($k_{x_{SS}}$), t_1 to $(t_n - t_0)/2$ and b to 1. For impulses, h_0, h_1 and h_2 are initially set equal to the steady-state rates value ($k_{x_{SS}}$), t_1 to $(t_n - t_0) \cdot 1/3$, t_2 to $(t_n - t_0) \cdot 2/3$, and b to 1. Following the optimization, we estimate per each model the Log-Likelihood (Equation 14) and the Akaike Information Criterion (AIC):

$$AIC = 2 \cdot k - 2 \cdot \mathcal{LL}\left(\theta | \widehat{x}, \sigma_x\right). \tag{31}$$

where θ is the vector of model parameters and k the complexity of the model, i.e. $dim(\theta)$. The model with the lowest AIC is compared with its nested neighbours (Method Fig 6) by means of the LLR test (Equation 15) to estimate a p-value for the variability of each kinetic rate.

Derivative modelling without nascent RNA (INSPEcT-) Similarly to the integrative approach of INSPEcT-, without the information of nascent RNA it is necessary to fit different models and perform model selection to discriminate between constant and non-constant rates and avoid overfitting. Differently from the integrative approach, within the derivative framework



Method Figure 6: **Diagram showing the eight models considered in our pipeline and their nested relationships.** Models are named after the non-constant rate(s) of the RNA life-cycle. *s* stands for synthesis (k_1), *p* for processing (k_2), and *d* for degradation (k_3). Each arrow links two nested models. The direction points to the more complex, while the color identifies the kinetic rate which is differentially regulated in the two scenarios: synthesis (red), processing (blue) and degradation (green).

the form of $k_1(t)$ is not restricted to any of the functional forms but is obtained assigning a functional form to $M(t)$, $k_2(t)$ and $k_3(t)$, as described in Equation 25. As a consequence of this, it is not possible to model a constant k_1 as explicitly required for models *p*, *d* and *pd* (Method Fig. 6). For this reason, we took advantage of the ODEs system expressed in terms of P and T (Equations 2a and 2b) to model an explicitly constant k_1.
In particular, in the model *p* (variable $k_2(t)$ and constant k_1 and k_3), $P(t)$ and $k_2(t)$ are expressed in function of $T(t)$, k_1 and k_3:

$$\begin{aligned}\dot{T}(t) =& k_1 - k_3 \cdot (T(t) - P(t)),\\ P(t) =& T(t) + \frac{\dot{T}(t) - k_1}{k_3},\end{aligned}$$

$$\begin{aligned}k_2(t) =& \frac{k_1 - \dot{P}(t)}{P(t)},\\ k_2(t) =& -\frac{\ddot{T}(t) - k_3 \cdot \dot{T}(t) - k_1 \cdot k_3}{\dot{T}(t) + k_3 \cdot T(t) - k_1}.\end{aligned} \tag{32}$$

Similarly, in the the model d (variable $k_3(t)$ and constant k_1 and k_2), $P(t)$ and $k_2(t)$ are expressed in function of $T(t)$, k_1 and k_2:

$$\begin{aligned}\dot{P}\left(t\right) =& k_1 - k_2 \cdot P\left(t\right) = 0,\\ P =& \frac{k_1}{k_2},\end{aligned}$$

$$\begin{aligned}\dot{T}\left(t\right) =& k_1 - k_3\left(t\right) \cdot \left(T\left(t\right) - P\right),\\ k_3(t) =& \frac{k_1 - \dot{T}\left(t\right)}{T\left(t\right) - P.}\end{aligned} \tag{33}$$

In the model pd (variable $k_2(t)$ and $k_3(t)$, and constant k_1), $P(t)$ and $k_2(t)$ are expressed in function of $T(t)$, k_1 and $k_3(t)$:

$$\begin{aligned}\dot{T}\left(t\right) =& k_1 - k_3\left(t\right) \cdot \left(T\left(t\right) - P\left(t\right)\right),\\ P\left(t\right) =& T\left(t\right) + \frac{\dot{T}\left(t\right) - k_1}{k_3\left(t\right)},\end{aligned}$$

$$\begin{aligned}k_2\left(t\right) =& \frac{k_1 - \dot{P}\left(t\right)}{P\left(t\right)},\\ k_2\left(t\right) =& -\frac{\ddot{T}\left(t\right) + \dot{T}\left(\frac{k_3^2(t) - \dot{k}_3(t)}{k_3(t)}\right) + \frac{k_1 \cdot \dot{k}_3(t)}{k_3(t)} - k_1 \cdot k_3\left(t\right)}{\dot{T}\left(t\right) + k_3\left(t\right) \cdot T\left(t\right) - k_1}.\end{aligned} \tag{34}$$

These three models are optimized by fitting a sigmoid and an impulse function on the profile of $\widehat{T}$. After choosing the best functional form in terms of the χ^2 p-value, the parameters of $T(t, \theta_T)$, are optimized together with the parameters associated to the kinetic rates by minimizing the χ^2 cost function defined over $\widehat{P}$ and $\widehat{T}$:

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(P(t = t_i) - \widehat{P}_i\right)^2}{\sigma_{P_i}^2} + \frac{\left(T(t = t_i) - \widehat{T}_i\right)^2}{\sigma_{T_i}^2}. \tag{35}$$

All other models (s, sp, sd, spd) are optimized by fitting a sigmoid and an impulse function on the profile of $\widehat{M}$. After choosing the best functional form in terms of the χ^2 p-value, the parameters of $M(t, \theta_T)$, are optimized together with the parameters associated to the kinetic rates by minimizing the χ^2 cost function defined in Equation 30 and expressing $P(t)$ and $k_1(t)$ as defined in the Equations 24 and 25.

In all cases, parameters associated to the kinetic rates are initialized as the population median of the rates calculated in first step of the modeling (i.e. median(k_x), see Section 3.1), optimized in order to fit the steady-state solution presented in Equations 3a, 3b, through the minimization of the cumulative χ^2 function described in Equation 30, i.e. $k_{x_{SS}}$. For sigmoids h_0 and h_1 are initially set equal to $k_{x_{SS}}$, t_1 to $(t_n - t_0)/2$ and b to 1. For impulses, h_0, h_1 and h_2 are initially set equal to $k_{x_{SS}}$, t_1 to $(t_n - t_0) \cdot 1/3$, t_2 to $(t_n - t_0) \cdot 2/3$, and b to 1. The choice of using population medians instead of gene specific parameters at the beginning of the optimization procedure is motivated by the fact that the derivative approach in the absence of nascent RNA may result in undefined models for specific set of parameters. By empirical observation, the optimized population medians $k_{x_{SS}}$ returned a lower fraction of undefined models compared to optimized gene medians. In any case, if the cost function is not defined in the initial condition of the optimization we search for the minimum increase of $k_{x_{SS}}$ which guarantees the correct initialization. We account for this correction and we added a term in the cost function which penalizes the distance of $k_1(0)$, $k_2(0)$ and $k_3(0)$ from their steady state values (i.e. "no-reg" model).
After the optimization of the eight models, we perform the model selection following the same procedure presented for the integrative approach of INSPEcT-.

4 Validation of INSPEcT- time-course framework (Main Figures 2, 4, 5, 6)

4.1 Contamination between unlabelled and labelled RNA

In order to estimate the amount of unlabelled RNA in the labelled fraction, we quantified the amount labelled RNA extracted at increasing labelling times, namely 0, 10, 20, 30, 60, 120 min. We reasoned that in the absence of contamination from the unlabelled fraction, labelled RNA (T_L) should be absent at labelling time equal to 0 and increase following the kinetics of 4sU incorporation (k) and preexisting RNA degradation (j):

$$\begin{cases} \dot{T_L} = k - j \cdot T_L, \\ T_L(0) = 0. \end{cases} \tag{36a}$$

The model was not able to recapitulate the experimental data (red line in Main Fig. 2A, Likelihood lower than $1e-80$), mainly due to a lower amount labelled RNA recovered in the early time points compared to modeled. We

reasoned that 4sU incorporation could be limited at initial time-points by the intracellular 4sU availability. For this reason, we modeled 4sU incorporation with an exponential increase with parameter b:

$$\begin{cases} \dot{T_L} = k \cdot \left(1 - e^{-b \cdot t}\right) - j \cdot T_L, \\ T_L(0) = 0. \end{cases} \tag{37a}$$

This model (green line in Main Fig. 2A) had a Likelihood of about $1e-20$. To establish if the increase in performance was sufficient to justify the new degree of freedom, we performed a Log-Likelihood ratio test. The test returned a very significant p-value ($2.0e-27$), therefore we concluded that the assumption of an exponential 4sU incorporation rate was reasonable. Nonetheless, this model is not able to recapitulate the amount of labeled RNA at early time pints. To test if a constant contamination term (C) could recapitulate our experimental observations during the whole time frame of our analysis, we devised the following model:

$$\begin{cases} \dot{T_L} = k \cdot \left(1 - e^{-b \cdot t}\right) - j \cdot T_L, \\ T_L(0) = C. \end{cases} \tag{38a}$$

This model fitted the early data point as well as later time points (blue line in Main Fig. 2A), which reflected in an increase of the Likelihood. Moreover, this model performed significantly better than the previous one, as attested by the Log-Likelihood ratio test ($3.1e-7$). Finally, we tried to model a variable contamination coefficient linearly increasing over time:

$$\begin{cases} \dot{T_L} = k \cdot \left(1 - e^{-b \cdot t}\right) - j \cdot T_L + a, \\ T_L(0) = C. \end{cases} \tag{39a}$$

In this case, the additional complexity did not resulted in a significantly better fit (Log-Likelihood ratio test p-value = 1).
Concluding, based on the modeling of labeled RNA extracted at different labelling times, we estimated a 30% of unlabeled RNA contamination at 10 minutes of 4sU pulse.

4sU labeling, extraction and quantification protocol. Detection of nascent RNA by metabolic labeling using 4-thiouridine (4sU, Sigma T4509) has been performed as previously described (Sabo' et al 2014). Briefly, cells were labeled with 300 μM 4sU for 10, 20, 30, 60, 120 min respectively. RNA was extracted using the Qiagen miRNeasy kit according to the manufacturer's instructions. 50 μg of total RNA were subject to biotinylation reaction: RNA was diluted in 100 μl of RNase-free water, 100 μl of biotinylation

buffer (2.5 x stock: 25 mM Tris pH 7.4, 2.5 mM EDTA) and 50 μl of EZ-link biotin-HPDP (1 mg ml-1 in DMF; PierceThermo Scientific 21341) were then added to diluted RNA and incubated for 2 h at room temperature (RT). RNA was precipitated and unbound biotin-HPDP was removed by chloroformisoamylalcohol (24:1) and purified using MaXtract high density tubes (Qiagen). Resulted biotinylated RNA was purified using Dynabeads MyOne Streptavidin T1 (Invitrogen). 50 μl of beads were washed 2 times in washing buffer A (100 mM NaOH, 50 mM NaCl), once in washing buffer B (100 mM NaCl) and resuspended in 100 μl of buffer C (2 M NaCl, 10 mM Tris pH 7.5, 1 mM EDTA, 0.1 Tween-20). RNA was added in an equal volume and rotated at RT for 20 min. Next, beads were washed 3 times with washing buffer D (1 M NaCl, 5 mM Tris pH 7.5, 0.5 mM EDTA, 0.05% Tween-20). RNA was eluted from the beads in 100 μl of 10 mM EDTA in 95% formamide (65 $^{\circ}C$, 10 min). After that, RNA was extracted with the RNeasy MinElute Spin columns (Qiagen) according to the manufacturer's instruction and eluted in 15 μl of RNase-free water. RNA quality was assessed using the Agilent 2100 Bioanalyzer (Agilent Technologies). 4sU-labeled purified RNA was quantified by the Qubit®2.0 Fluorometer according to manufacturer's instruction.

4.2 Simulation of RNA life-cycle data

In this section, we present our routine for simulated data genesis (graphically represented in Supplemental Fig. 2). The possibility to model the contamination of nascent RNA with unlabelled RNA is the major improvement of data simulation routine, compared to the version provided in the first release of the package [7].

Sampling of RNA life-cycle kinetic rates and fold-changes. Gene expression levels and RNA life-cycle kinetic rates (estimated in the first step of modelling, described in Section 3.1) are retrieved from an object of class INSPEcT. From the empirical distribution $P(k_1)$, one value is extracted through random sampling (k_1^g). Then, we subdivide $P(k_1)$ in quantiles to select the genes with a similar k_1. From those genes three empirical distributions are generated: $P(k_2|k_1^g)$, $P(k_3|k_1^g)$ and $P(k_{1_{FC}}|k_1^g)$. We iterate the procedure to sample the maximum Log_2 fold change distributions for the rates of processing and degradation conditioned to the corresponding rate and $k_{1_{FC}}^g$. At the end of this process, we obtain, for each artificial gene, a set of six numerical values that provide its RNA life-cycle dynamics. The conditional sampling guarantees to preserve the correlations which are known to

exist between rates absolute levels and Log_2 fold changes (see the paragraph *Comparison between simulated and real data* in this section).

Selection of RNA life-cycle functional forms. We proceed defining the dynamic response of each gene in the simulated dataset. For any rate, we select one among three possible functional forms (constant, sigmoid, impulse), according to a specific probability provided by the user (by default 0.5, 0.3, 0.2), and we shape it using the parameters previously defined. In this manner, we set the order of magnitude of the rate and, for sigmoids and impulses, also the maximum fold change; the response times are randomly selected according to the time window of the INSPEcT object used as reference. After the kinetic rates parametrization, we can solve numerically the ODEs system (Equations 1 and 2) to estimate an expression profile for nascent, premature and total RNA.

Sampling of gene specific contamination. We simulate a gene specific contamination of the nascent portion of the transcriptome from the pre-existing counterpart, modelled as a linear combination of the two fractions, the latter one scaled by a numerical corruption coefficient $b_g \sim N(\mu_{b_g}, \sigma_{b_g})$. This coefficient serves as a fraction of contamination, therefore values lower than 0, or larger than 1 are set equal to the extremes of the interval. The parameters μ_{b_g} and σ_{b_g} can be set by the user. By default, μ_{b_g} is set to reproduce the 30% average contamination estimated experimentally in Section 4.1, while σ_{b_g} is set to reproduce a correlation score of 0.7 [8] between the estimated k_3 and the true ones, as often reported for k_3 calculated by independent methods.

Sampling of variability between replicates. We exploit the Plgem Bioconductor package [9] to estimate a power law relation between each experimental expression data (premature RNA, total RNA and nascent RNA) and the associated variances. These functions allow to associate a standard deviation to each simulated expression level. In this manner, we pass from a set of numerical values, the simulated expression levels, to a group of Gaussian distributions. Finally, we sample them to simulate a real sequencing experiment. The number of samplings is defined by the user according to the number of biological replicates to simulate. All these values are then used to evaluate the mean simulated expression data profiles and the associated standard deviations.

Comparison between simulated and real data. We modelled a simulated dataset of 1000 genes in 11 time points and 3 replicates from a small set of genes released with the INSPEcT package and we compared the expression values and the rates first guess against the counterparts obtained from the analysis of the experimental data (Supplemental Fig. 3). Specifically, we compared the Log_2 distributions of total RNA, premature RNA, synthesis, processing and degradation rates first guess estimated at the steady-state. For the same quantities, we also compared the distributions of mean Log_2 fold change against the 0 minutes condition. Finally, we confronted the mean-variance relations of the experimental data. For each couple of box plots we used a non-parametric Two-Sided Wilcoxon test and we always got either not or barely significant p-values which attested very small differences between the two distributions. This is evident for the steady-state absolute values while the mean Log_2 fold changes distributions appear to be more different. This higher variability could be explained by different proportions of constant and variable rates in the two conditions. In the simulated dataset, each rate has the same probability to be constant or variable while the real genes we used to generate the data are mainly regulated in synthesis.

4.3 Measure of the classification performance

Simulated data created as described in the previous section are used as a benchmark to measure the ability of INSPEcT in discriminating between constant versus variable rates of the RNA life-cycle throughout a time course analysis. In particular, we measured the ability of INSPEcT in the identification of the variability of individual rates (i.e. k_1, k_2 or k_3) for all the genes of the simulated dataset by comparing INSPEcT results with the true class that generated the simulated profile.

ROC curves analysis Considering that we analyze the classification performance individually for each rate, we used the Receiver Operating Characteristic (ROC) curves, which are commonly used for binary classification. In our case, they were created by measuring, at each p-value threshold resulting from the model selection of INSPEcT, the True Positive Rate (TPR, sensitivity) and the False Positive Rate (FPR, 1 - specificity) defined as follow:

$$\mathrm{TPR} = \frac{\text{true positive}}{\text{condition positive}} = \frac{\mathrm{TP}}{\mathrm{P}},$$
$$\mathrm{TNR} = \frac{\text{true negative}}{\text{condition negative}} = \frac{\mathrm{TN}}{\mathrm{N}}.$$

The TPR and FPR calculated for each kinetic rate defined the ROC curves plotted in Main Figure 6C. The integral of a ROC curve is called Area Under the Curve (AUC), and is an indicator of the performance. AUC ranges from 0 (completely wrong classifier) to 1 (perfect classifier), and 0.5 is the score expected from a random classification. In this research article, we often reported the AUC score without showing the original ROC curve for graphical reasons (Main Fig. 2E, Supplemental Fig. 10, Supplemental Fig. 7 - 9).

Measure of cross-classification performance The performance analysis reported in Supplemental Fig. 10 is atypical because we classified the variability of one rate (i.e. k_1) according to the p-values estimated by INSPEcT relative to the variability of the other two (k_2 and k_3). In this situation, our expectation was to observe AUCs close to 0.5 as the variability of each rate was selected independently to the others.

Measure of the performance at fixed threshold The ROC analysis provides an overview of the performance of a binary classifier. However, it lacks measuring the classification performance at a given threshold. In fact, INSPEcT returns a classification based on the p-value of each rate. Specifically, if the (adjusted) p-value is lower than a certain threshold the rate is considered variable, otherwise it is constant. For this reason, INSPEcT classification performance was measured at the default threshold of adjusted p-value ≤ 0.05. The classification was measured by means of both sensitivity and specificity (Supplemental Fig. 6), or by means of a single indicator, i.e. the F_1 score (Main Fig. 4E), which is their harmonic mean:

$$F_1 = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FN} + \mathrm{FP}}.$$

4.4 Impact of time series design on classification performance

In order to elucidate the relevant features of a time series design that are able to impact INSPEcT classification, we created and analyzed with INSPEcT-simulated datasets with specific features.

Optimal design for the identification of sigmoidal k_1 responses We created a dataset of 100 simulated genes, including 50 constant genes, and 50 genes transcriptionally regulated through sigmoidal modulations covering a time lapse of 16 hours (sigmoid dataset). We sampled the sigmoid dataset with several time series of three points spanning different "time windows"

between 0 and 16 hours (Supplemental Fig. 7A). We modeled each gene with INSPEcT- and estimated the classification performance of the synthesis rate through ROC analysis (AUC score - red line in Supplemental Fig. 7B). INSPEcT- classification performance increased along with the portion of the time-course covered by the three time-points "time window". In particular, the AUC is linearly related to the fraction of regulated genes whose k_1 half response time (τ_{k_1}) is included in the "time window" (Supplemental Fig. 7C, D). The half response time of mature RNA (τ_M) can be used as a proxy of τ_{k_1}, which is not directly available to the user. In particular, τ_M is always greater or equal than τ_{k_1}, following the relation:

$$\tau_M \approx \tau_{k_1} + \frac{Log(2)}{k_2} + \frac{Log(2)}{k_3} \tag{40}$$

and can be used to design the experiment (Supplemental Fig. 7E).

Optimal design for the identification of k_1 impulsive responses We created a dataset of 100 simulated genes, including 50 constant genes, and 50 genes transcriptionally regulated through impulsive modulations covering a time lapse of 16 hours (impulse dataset). Again, we sampled the dataset with several time series of three points spanning different "time windows" between 0 and 16 hours (Supplemental Fig. 7A). Differently from the sigmoid dataset, the classification of the impulse dataset was not affected by the temporal positioning of the three time points (AUC score - green line in Supplemental Fig. 7B). Instead, we reasoned that this dataset might benefit from an increasing number of time points. We considered between 3 and 15 time points linearly distributed in the 0-16 hours time lapse, i.e. having decreasing "time steps" between time points (Supplemental Fig. 8A). Indeed, when we modeled these data with INSPEcT- we found that the AUCs of the synthesis rates increased along with the number of time points (Supplemental Fig. 8B). Impulse functions are characterized by a double response. We estimated for each variable gene in the dataset a "response interval" (δ_{k_1}) defined as the difference between the second and the first half response times (exemplified in Supplemental Fig. 8E). Then, we computed for each series the fraction of variable genes with a "response interval" larger than the "time step" (exemplified in Supplemental Fig. 8A). We found a clear relation between this quantity and the associated AUC relative to the synthesis rate (Supplemental Fig. 8C, D). Unfortunately, the "response interval" of M (δ_M), which is easier to directly measure, is always larger or equal then δ_{k_1} and, for this reason, the prediction based on this datum is an upper limit of the real classification power. However, it is possible to approximate δ_{k_1} from

δ_M, k_2 and k_3 by exploiting the empirical formula:

$$\delta_M \approx \delta_{k_1} + \frac{1}{2} \cdot \left(\frac{Log(2)}{k_2} + \frac{Log(2)}{k_3} \right) \tag{41}$$

This relation is not exact and its error is inversely proportional to the rates of processing and degradation. However, it could be a reasonable starting point for the careful design of an RNA profiling time-course experiment (Supplemental Fig. 8E). Following this, we reasoned that the time points overlapping the k_1 response should be the most important. To verify this hypothesis, we retrieved the first half response times of the synthesis rate, which showed that the largest part of the response of the simulated dataset resides between 0 and 4 hours (histogram in Supplemental Fig. 8E). The ROC analysis showed in the right side of Supplemental Fig. 8E indicates that the AUC of the 15 points time series described above (green line, AUC=0.90) is only marginally reduced when an halved number of time points covers the 0-4h time span (red line, AUC=0.86). Rather, a similar drop in the number of time points affects the AUC performance when these cover the less informative portion of the time-course (gold line, AUC=0.66). An reduction in the number of time points with homogeneous distribution could be a trade-off between costs and performance (blue line, AUC=0.74). Finally, an experimental design with logarithmic distributed time points could be an even better cost-saving strategy (black line, AUC=0.82).

Optimal design for mixed responses of the kinetic rates So far, we worked in an ideal situation where only transcriptional regulation occurs. To validate our findings in the most general case we generated two sets of 100 simulated genes, each independently regulated in synthesis, processing and degradation with the same probability (Probability of regulation = 0.5). The first dataset was composed of sigmoidal modulations while the other one was composed of impulsive regulations only. We sampled these two systems according to the time series presented in Supplemental Fig. 8A, we modeled each dataset with INSPEcT-, and we estimated the corresponding AUCs (Supplemental Fig. 9). As expected, the classification performance of INSPEcT- on the sigmoid dataset was independent on the number of time points (Pearson correlation coefficient of 0.08, P=0.61), while the classification of all kinetic rates benefits from an increasing number of time points in the impulse dataset (Pearson correlation coefficient of 0.67, $P < 1e-5$).

Summary of the key results of these analyses

- INSPEcT– classification performance depends on the composition of the simulated dataset in terms of sigmoidal and impulsive responses;
- To adequately cover simple sigmoidal responses, it is sufficient to build a time series with a limited number of time points which cover a large portion of the regulation. The latter can be estimated by the half-response time of mature RNA;
- To adequately cover complex impulsive responses, it is necessary to design a sufficiently dense time series. The time step between time points directly relates to the classification performance, especially for genes characterized by high processing and degradation rates;
- Not all the data points of a time series are equally relevant. The more the time points cover the portion of the time course where most of the modulations occur, the higher is their added value.

4.5 Comparison with independent quantification of k_1 and k_3

To further validate INSPEcT- models, we compared the synthesis and degradation rates estimated from the MYC activation dataset (3T9 mouse fibroblasts) with independent quantifications.

Regarding the synthesis rate, we decided to compare the values returned by the modelling procedure against nascent RNA expression levels (RPKM) estimated merging intronic and exonic reads. Specifically, we performed ranked correlation analyses on: the untreated condition (Supplemental Fig. 5A), following 4 hours of MYC activation (Supplemental Fig. 5B), and on the Log_2 fold changes between those conditions (Supplemental Fig. 5C). The comparisons were done on a set of 6'446 genes identified as transcriptionally regulated, the resulting Spearman's correlations were 0.88, 0.90 and 0.87 respectively.

To benchmark INSPEcT- degradation rates, we focused on the TimeLapse-seq method [10] that provides, as a Supplemental of the original publication (Supplemental Table 2), a set of RNA half-lives ($Log(2) \cdot k_3^{-1}$) estimated from the analysis of Mouse Embryonic Fibroblasts (MEF) cells; 2'992 genes in two replicates. This cell type is biologically close to the 3T9 murine fibroblasts we used in the MYC activation dataset, although not identical. To quantify this similarity, we downloaded two samples of wild type MEF cells total RNA-seq (SRR125393 and SRR125394) and we quantified the counts for 24'528 genes aligning the reads on the mm10 reference genome quantifying the exonic reads

through INSPEcT. We checked the Spearman's correlation between replicates (0.98) and we estimated gene's mean values. Then, we did the same for the counts estimated by INSPEcT on the untreated 3T9 samples (21'677 genes - replicates Spearman's correlations around 0.99). Finally, we performed a ranked correlation analysis between the mean expressions of MEF and 3T9 cells on 21'324 common genes finding a value of 0.89. To perform the comparison between degradation rates, we converted the original MGI symbols provided in the TimeLapse-seq dataset to Entrez IDs exploiting the biomaRt R package (dataset GRCm38.p6); we managed to fine a unique correspondence for 2'869 genes (2'737 part of the INSPEcT- dataset). Finally, we found a Spearman's correlation of 0.5 (p-value $< 1e-173$) - Supplemental Fig. 5D) between TimeLapse and INSPEcT- degradation rates. Interestingly, it was higher than the counterpart estimated confronting TimeLapse-seq rates against the INSPEcT+ models (0.46 - p-value $< 1e{-}148$) - Supplemental Fig. 5E). We would like to stress a key point for the interpretation of these results. While MEF cells are similar to 3T9 cells, they are not exactly the same cell type. Indeed, the genes considered for the comparison of degradation rates had a Spearman's correlation of 0.67 in their expression (Supplemental Fig. 5F).

5 Steady-state experimental design (Main Figure 7)

5.1 Without nascent RNA data (INSPEcT-)

5.1.1 General framework

The knowledge of the steady-state values of P and M is not sufficient to estimate the k_1, k_2 and k_3 that generated them. In fact, the system of equations 3 is composed of two Equations and three unknowns, and can be solved by infinite combinations of rates. INSPEcT- exploits P and M expression levels to retrieve the maximum information regarding the RNA life-cycle, calculating the ratio between processing and degradation rates from steady-state observations (post-transcriptional ratio, or PT-ratio, Equation):

$$\frac{P}{M} = \frac{k_3}{k_2}. \tag{42}$$

The variation of the PT-ratio of a gene among conditions is sufficient, although not necessary, to identify post-transcriptional regulation. In fact, while a significant variation of the PT-ratio might only be explained by the

variation of either k_2 or k_3, concordant and equal variations of k_2 and k_3 balance each other leaving the PT-ratio unchanged. Instead, variations in the k_1 do not impact the PT-ratio, as they modulate both P and M to the same extent.

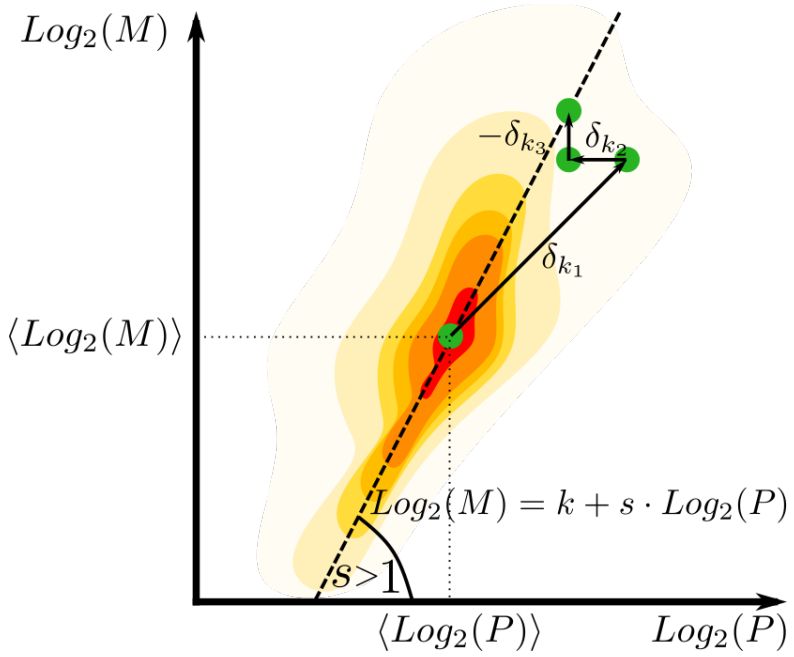
INSPEcT- goes beyond the pure identification of post-transcriptional regulations. Several works in the literature showed that the rates of the RNA metabolism are coupled [11]. As a consequence of this, it is expected that a certain regulation in k_2 and k_3 might follow the regulation of transcription. In order to discriminate between post-transcriptional regulations that are the consequence of rate coupling and other kind of post-transcriptional regulations (henceforth atypical), INSPEcT- estimates a linear relation in the $Log_2(P) - Log_2(M)$ space, i.e. a power law, among the conditions medians of P and M (Main Fig. 7C). The slope (s) indicates whether the PT-ratio increases ($s > 1$), decreases ($s < 1$) or is invariant ($s = 1$) at increasing expression levels. While a slope = 1 could be explained by an absence of a general coupling between k_1 and the post-transcriptional rates, in other cases it is reasonable to suppose that the trend is measuring the coupling between the rates of the RNA metabolism. In fact, in the P-M graph (Method Fig. 7):

- a modulation in k_1 results in an equal displacement of P and M,
- a modulation in k_2 results in an opposite displacement in P,
- a modulation in k_3 results in an opposite displacement in M,

Obviously, it is not possible to infer the exact relation between k_1, k_2 and k_3. For instance, the simplest explanation for PT-ratio increasing with the expression regime is that increases of k_1 are generally followed by an increase in k_2 or a decrease in k_3 (or a combination of the two), but multiple and more complex scenarios can fit the same observation. For these reasons, INSPEcT- does not guess which rate is modulated following a specific perturbation of P and M, but just identifies atypical regulations.

5.1.2 Inference of the P-M trend

The linear relation in the $Log_2(P)$-$Log_2(M)$ space is inferred based on the the median P and M observations across all the conditions available, individually for each gene. INSPEcT- defines a set of lines with different slopes (any integer between -89 and 90 degrees) interpolating the median point of the P and M distributions. For each line, INSPEcT- counts the number of median P and M observations included in a portion of the $Log_2(P)$-$Log_2(M)$ space

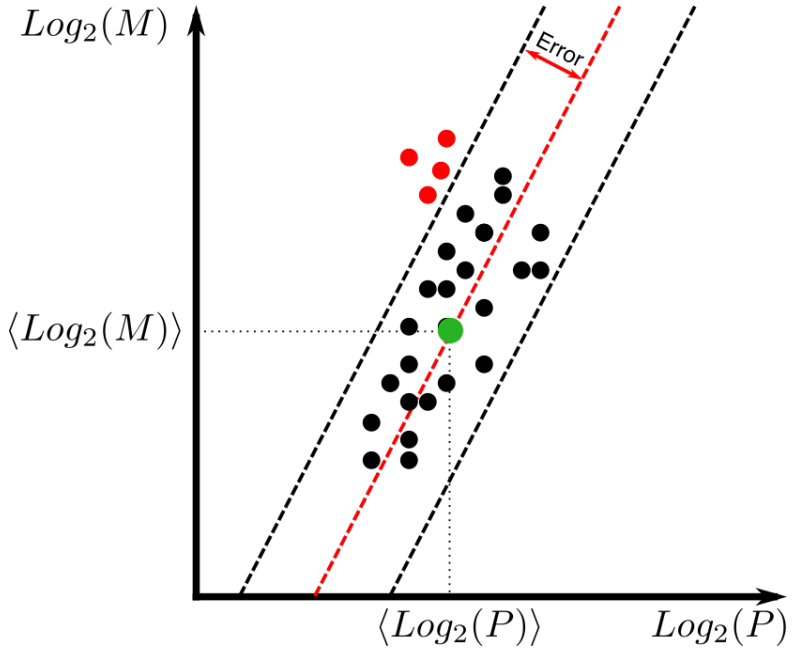


Method Figure 7: **Effects of transcriptional and post-transcriptional regulations on the steady-state premature and mature RNA expression levels.** Starting from the circle located in the median point of the cloud, we show with black arrows the effects of a modulation on the rate of synthesis δ_{k_1}, processing δ_{k_2} and degradation $-\delta_{k_3}$. The black dashed line represents the linear trend of the cloud, which represent P and M condition medians distribution.

defined by the two lines parallel to the one previously defined and distant $\pm Error$ from it (user defined parameter, 2 by default). Finally, the model which maximises the number of observations contained, i.e. that explains the higher number of data points, is chosen (Main Fig. 7C).
The P-M trend can be affected by different aspects, like technical issues, cell type or the gene class analyzed. For this reason, INSPEcT- calculates the P-M trend in a dataset specific manner.

5.1.3 Identification of atypical regulations

After the estimation of the typical regulatory strategy, the same linear model is applied to describe the $Log_2(P)$-$Log_2(M)$ of individual genes. The general P-M trend is translated to interpolate the P and M medians of the gene. All the conditions that are not compatible with the expected behaviour (points which are outside the portion of plane defined by the parallel lines defined by the $\pm Error$) are identified as atypically regulated (Method Figure 8).



Method Figure 8: **Identification of atypical regulations at the single gene level.** We spot conditions where a gene is atypically post-transcriptionally regulated (red dots) exploiting the model fitted on a population of genes (dashed lines) forced to interpolate the gene median point of the gene (green dot).

5.2 With nascent RNA data (INSPEcT+)

5.2.1 General framework

The post-transcriptional ratio is an aggregated quantity which does not provide information on the value of each kinetic rate. To uniquely solve the system introduced in Equation 3 and determine the complete set of kinetic rates another datum is required, for example the amount of nascent RNA. The metabolism of this portion of the transcriptome (T_L) can be described with the same ODEs systems introduced in Equations 1 and 2. For a labelling time (τ) sufficiently brief, the level of T_L closely reflects the action of RNA synthesis because the contribute of degradation can be disregarded. In this approximation, and assuming the steady-state condition for the rate of synthesis, Equation 2b can be rewritten as

$$\dot{T}_L(t) = k_1, \tag{43}$$

which is readily integrable in a time window of length τ, assuming the absence of endogenous 4sU ($T_L(0) = 0$), as

$$T_L = k_1 \cdot \tau. \tag{44}$$

As previously mentioned, the esteem of k_1 from nascent RNA allows to quantify processing an degradation rates from pre-existing RNA according to Equation 3:

$$k_1 = \frac{T_L}{\tau}, \tag{45}$$

$$k_2 = \frac{T_L}{\tau \cdot P}, \tag{46}$$

$$k_3 = \frac{T_L}{\tau \cdot M}. \tag{47}$$

The estimation of k_2 and k_3 requires the direct comparison of quantities derived from different sequencing libraries, i.e. T_L from the nascent RNA fraction, and P, M from the total RNA fraction. The library size normalization assumes that the libraries under comparison are derived from a similar amount of cellular RNA, which is not the case when analysing different fractions of the same RNA. For this reason, INSPEcT implements an internal normalization procedure between nascent and total RNA libraries based on the modeling of RNA metabolism, described in the following Section (5.2.2).

5.2.2 Scaling between total and nascent libraries

The goal of the normalization procedure implemented within INSPEcT is the estimation of an individual scaling factor that normalizes each nascent RNA

library to its total RNA library of reference. In order to do this, INSPEcT takes advantage of the ODEs system introduced in Equation 2. In particular, by solving the Equation 1a between time 0 and τ assuming the absence of endogenous nascent RNA, it is possible to model the amount of premature RNA in the nascent RNA fraction as a function of the sole unknown k_2, τ and T_L:

$$P_L = \frac{T_L}{\tau \cdot k_2} \cdot \left(1 - e^{-k_2 \cdot \tau}\right). \tag{48}$$

The rate k_2 can be estimated by minimizing the error between the modeled P_L and the measured one.

$$k_2 = \underset{k_2}{\operatorname{argmin}} \left[\widehat{P_L} - \frac{\widehat{T_L}}{\tau \cdot k_2} \cdot \left(1 - e^{-k_2 \cdot \tau}\right)\right]^2. \tag{49}$$

Additionally, the rate k_2 can be obtained also by Equation 46. INSPEcT exploits this redundancy to estimate a normalization factor α that can be used to scale the library of nascent RNA:

$$\alpha = \underset{\alpha}{\operatorname{argmin}} \left(k_2 - \alpha \cdot \frac{\widehat{T_L}}{\tau \cdot \widehat{P}}\right)^2, \tag{50}$$

where k_2 is calculated as described in Equation 49. An individual α can be calculated for each gene, nonetheless, the aim of INSPEcT is to estimate a single value that will be used to scale all the observations from the nascent RNA library. For this reason, INSPEcT uses the α that minimizes the squared median of the residuals of Equation 50 calculated for all genes.

6 INSPEcT- analysis of a large dataset of publicly available RNA-seq samples (Main Figure 7)

6.1 Description of the RNA-seq dataset

Overall, with this procedure we obtained premature and mature expression quantifications (RPKMs) for 35'125 ensemble genes in 669 conditions, that were classified in 41 cell lines and 29 diseases.

SRAdb query to retrieve RNA-seq samples We selected 3'856 human RNA-seq samples, non poly(A)-selected, using SRAdb (version 1.40.0), by querying for the following not case sensitive values in the corresponding fields:

- library_strategy = 'RNA-Seq',
- taxon_id = 9606,
- library_construction_protocol, design_description and sample_attribute containing at least one among the following values: "ribozero", "ribo0", "ribominus" and "ribo-"; and not containing: "cytoplasm", "nascent rna", "poly-a", "polya", "mrna", "ffpe" (i.e. paraffin conserved samples).

Samples annotation via *Onassis* package We used the R/Bioconductor package *Onassis* (version 3.8), which leverages natural language processing and biological ontologies, to associate relevant annotation terms to the selected samples based on their description in the SRA database. Onassis annotated 3'724 experiments using either "disease" [12] or "cell line" [13] ontologies, after removing some uninformative terms ("cell line" ontology: 'cell', 'tissue', 'homo sapiens', 'molecule', 'female organism', 'male organism', 'protein', 'cell line cell', 'chromatin', 'signaling', 'cultured cell', 'multicellular organism', 'compound organ', 'organ', 'nucleus', 'primary cultured cell', 'diploid', 'Bos taurus', 'process', 'chromatin', 'protein', 'size', 'ribosome' , 'organ part', 'time', 'body proper', 'multicellular organism'), and assigning the term "healthy" to all samples containing the following strings within the fields "sample_attribute" or "experiment_attribute": 'healthy', 'disease: none', 'disease: normal', 'disease: presumed normal', 'disease: no ad present', 'disease: no ad evident', 'disease state: normal', 'tissuetype: normal', 'no ad present', 'disease: healthy', 'disease: normal', 'disease: presumed normal', 'disease: none', 'disease: null', 'disease: na', 'disease status: normal', 'tumor: none'.

Quantification of exonic and intronic features via *recount* package Among the annotated samples, 1'140 experiments from 103 projects were found to be part of the recount2 database, which comprehends SRA samples uploaded before February 3, 2016. We selected 1'004 experiments from 100 projects annotated in SRAdb with at least 7.5 milion aligned reads. Finally, we successfully downloaded from recount2 (version 1.4.5) the expression data for 669 experiments from 75 projects. Exonic and intronic quantifications were obtained using the normalized coverage computed by

the R/Bioconductor package *recount*, using the function "coverage_matrix". For experiments with multiple runs associated, the mean intronic and exonic quantification were obtained.

6.2 Gene class specific P-M trend

We subdivided the 35'125 genes quantified from recount2 in three classes according to their "gene_type" tag in Gencode annotation (version 25), obtaining 18'729 protein coding genes (corresponding to the term "protein_coding"), 3'945 pseudogenes (corresponding to the term "pseudogene") and 12'451 non-coding (corresponding to all the other terms). The P-M trend was calculated independently for each gene class, as described in Section 5.1.2 and used as a prototypical combined modulation of transcriptional and post-transcriptional rates.

6.3 Classification matrix

A Boolean matrix of atypical regulation of genes across samples was obtained as described in Section 5.1.3. Not expressed genes in individual samples are reported as missing observations.

Genes and samples filtering for data visualization For the sake of heatmap visualization (Main Fig. 7), we filtered both samples and genes from the RNA-seq dataset described in the previous section (6.1). In particular, we selected 620 (out of the 669) samples representative of the 26 most abundant cell lines. Additionally, we selected genes with both non-zero premature and mature quantification in at least half of the samples. This filter selected 18'621 protein coding, 3'910 pseudogenes, and 12'420 non-coding genes.

Hierarchical clustering Rows and columns were clustered using hierarchical clustering with euclidean distance, by assigning a value of 0 to not-expressed genes (corresponding to NAs in the classification matrix), 1 to expressed genes (corresponding to FALSE in the classification matrix), and 2 to expressed and atypically regulated genes (corresponding to TRUE in the classification matrix).

Impact of atypical regulations on samples clustering In order to quantify the impact of the atypically regulated genes on samples classification, we repeated the clustering only based on the expression information,

i.e. by setting all entries whose value was 2 to 1. The two dendrograms had a cophenetic correlation coefficient of 0.68 calculated with R package 'dendextend' version 1.63 (Supplemental Fig. 14), meaning that about 30% of clustering information derived from the atypical regulations.
In order to discriminate whether the classification provided by atypical regulations could be generated by chance, we randomly selected 1'824'559 atypical regulations among the expressed entries (the same number of atypical regulations identified by INSPEcT-), clustered the columns and estimated the cophenetic correlation coefficient against the clustering based on the sole expression matrix. After repeating this procedure 100 times, we estimated that the mean correlation coefficients was close to 0.68. Therefore, we concluded that the impact of INSPEcT-'s classification on the samples clustering was compatible with the random null model (Supplemental Fig. 14B).
However, this does not mean that the classification matrix produced by INSPEcT- belongs to the random set. To check this hypothesis, we started performing a correlation analysis between the classification matrix of INSPEcT- and all the random counterparts. Then, we repeated the correlation analysis for each random classification matrix against all the others and we observed a difference between the INSPEcT- correlations distribution and the random ones. We quantified these differences through the Wilcoxon test (W - R stat package) finding significant p-values for the comparison of INSPEcT- both against each random distribution and against the cumulative one ($p - value = 1.2e - 47$). The latter test was repeated for each random distribution and we found values much less significant that the one observed for INSPEcT- (Supplemental Fig. 14C).
Finally, we exploited the samples annotations to compare the clustering on INSPEcT- against the 100 random matrix and the expression matrix too. Specifically, we assigned to each leaf of the dendrogram the corresponding tissue label and we counted the fraction of leafs with N equal neighbours (homogeneity score). The boxplots in Supplemental Fig. 14D show the distributions of the 100 random matrices at four different orders. We see that the distributions are always centred around the score obtained from the analysis of the expression matrix which means that this property of the clustering is mainly driven by the not-expressed vs expressed entries. On the other hand, INSPEcT- provides a clustering always more fragmented than the random one due to the peculiar distribution of atypically regulated conditions.
From this analysis, it is reasonable to conclude that the probability to generate a classification matrix similar to the one provided by INSPEcT- through random sampling is very low.

6.4 Functional enrichment analyses

We performed a functional enrichment analysis on the 1'000 genes with the highest number of atypically regulated samples and on the 1'000 genes with the highest number of expressed and non-atypically regulated samples. Additionally, we investigated functional enrichment of genes with the highest extent of atypical regulations in specific cell lines selected due to their proximity in the clustered matrix (50 Brain samples, 29 Heart samples and 10 T cell samples). The enrichment was evaluated in 86 and 417 genes with at least 80% of regulated samples in Heart and T cell cell lines, respectively, and in 861 genes with at least 90% of regulated samples in Brain cell lines. The enrichment was evaluated using the R-package *rGREAT* (version 1.11.1) using all the available ontologies. Terms were considered significant when both the hypergeometric and binomial tests returned p-values below $1e-2$. Genes, annotations and their functional enrichment are available as Supplemental file.

6.5 Characterization of regulated genes in brain

The 5' and 3' UTR sequences of regulated genes were obtained from the UCSC Genome Browser (hg38 assembly) [14], along with their predicted secondary structure folding. Differences in length, GC content and free energy between 17'786 background and 861 regulated genes in Brain were computed by means of a Wilcoxon test, and plotted with R. De novo sequence motifs were searched by means of DREME [15], using the sequence of UTRs of the same type from non-regulated genes of the same tissue as the background set, an E-value threshold of 0.05 and considering only motifs on the forward strand. Eventually, to identify known binding sites of regulatory factors in those UTRs, we applied the Regulatory Enrichment tool of the AURA2 database [16] on the regulated genes set of each tissue separately, using an enrichment significance threshold of 0.05 (BH-adjusted p-value).

References

[1] Zaghlool, A. et al. *Efficient cellular fractionation improves RNA sequencing analysis of mature and nascent transcripts from human tissues.* BMC Biotechnol 13, 99 (2013).

[2] Uvarovskii, A., Naarmann-de Vries, I. S. & Dieterich, C. *On the optimal design of metabolic RNA labeling experiments.* PLoS Comput Biol 15, e1007252 (2019).

[3] Chechik, G. et al. *Activity motifs reveal principles of timing in transcriptional control of the yeast metabolic network.* Nat Biotechnol 26, 1251–1259 (2008).

[4] Chechik, G. Koller, D. *Timing of Gene Expression Responses to Environmental Changes.* Journal of Computational Biology 16, 279–290 (2009).

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

[6] Rabani, M. et al. *High-Resolution Sequencing and Modeling Identifies Distinct Dynamic RNA Regulatory Strategies.* Cell 159, 1698–1710 (2014).

[7] de Pretis, S. et al. *INSPEcT: a computational tool to infer mRNA synthesis, processing and degradation dynamics from RNA- and 4sU-seq time course experiments.* Bioinformatics 31, 2829–2835 (2015).

[8] Lugowski, A., Nicholson, B. & Rissland, O. S. *DRUID: a pipeline for transcriptome-wide measurements of mRNA stability.* RNA 24, 623–632 (2018).

[9] Pavelka, N. et al. *A power law global error model for the identification.* BMC Bioinformatics 5, 203 (2004).

[10] Schofield, J. A., Duffy, E. E., Kiefer, L., Sullivan, M. C. & Simon, M. D. *TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding.* Nat Methods 15, 221–225 (2018).

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

[12] https://raw.githubusercontent.com/DiseaseOntology/HumanDiseaseOntology/master/src/ontology/doid-non-classified.obo

[13] https://raw.githubusercontent.com/obophenotype/cell-ontology/master/cl.obo

[14] Haeussler M. et al. *The UCSC Genome Browser database: 2019 update* Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019, Pages D853–D858

[15] Timothy L. Bailey *DREME: motif discovery in transcription factor ChIP-seq data* Bioinformatics, Volume 27, Issue 12, 15 June 2011, Pages 1653–1659

[16] Erik Dassi et al. *AURA2: Empowering discovery of post-transcriptional networks* Translation, Volume 2, 2014 - Issue 1, Article: e27738