Multiple latent clustering model for the inference of RNA life-cycle kinetic rates from sequencing data

Gianluca Mastrantonio[1], Enrico Bibbona[1], Mattia Furlan†[2]

[1]*Department of Mathematical Science, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy.*

[2]*Center for Genomic Science, Fondazione Istituto Italiano di Tecnologia (IIT), Via Adamello 16, 20139 Milano, Italy.*

Summary. We here propose a hierarchical Bayesian model to infer RNA synthesis, processing, and degradation rates from sequencing data, based on an ordinary differential equation system that models the RNA life cycle. We parametrize the latent kinetic rates, that rule the system, with a novel functional form, and estimate their parameters through 6 Dirichlet process mixture models. Owing to the complexity of this approach, we are able to simultaneously perform inference, clustering and model selection. We apply our method to investigate transcriptional and post-transcriptional responses of murine fibroblasts to the activation of proto-oncogene Myc. Our approach uncovers simultaneous regulations of the rates, which had not previously been observed in this biological system.

Keywords: Dirichlet Process, kinetic rates, RNA

1. Introduction

RNA is one of the most important actors in the cellular biology context and it is involved, directly or indirectly, in any process that occurs inside a cell. This molecule is a cornerstone of the information-flow that subsists from DNA to proteins, due to both its role as a template for protein assembly and because of the involvement of non-coding RNAs in the regulation of gene expression levels (GELs) (e.g. modulation of transcript stability, protein synthesis and protein localisation) (Marchese et al., 2017; Slack and Chinnaiyan, 2019; Vandevenne et al., 2019). A cell constantly regulates the expression levels of thousands of genes, i.e. the number of the associated transcripts, in order to preserve its homoeostasis and adapt to the environment.

The RNA life-cycle in eukaryotic cells can be summarized in the following three steps: the synthesis of premature RNA molecules in the nucleus, their processing into mature transcripts (which includes exporting it to the cytosol), and mature RNA cytoplasmic degradation. The characterization of these mechanisms, which the cell exploits to modulate the amount of specific transcripts, according to internal and external stimuli, can provide exceptional insights into the biology of these responses. The investigation of the RNA life-cycle requires the experimental quantification of the GELs. The state-of-the-art approach used to perform this task is Next Generation RNA sequencing (RNA-Seq)

†Corresponding Author. E-Mail: `mattia.furlan@iit.it`

(Goodwin et al., 2016). Owing to the relevance of the topic, a remarkable number of public RNA-Seq datasets are currently available and easily accessible (e.g., the *Gene Expression Omnibus project* (Edgar et al., 2002)), and a large amount of literature has been produced on the analysis of RNA-Seq data. In some papers, mixture models are used on the observed data to identify differences in GELs, see, for example, "RNA-Seq by expectation-maximization" (Li and Dewey, 2011), "Cufflinks" (Trapnell et al., 2013), "Casper" (Rossell et al., 2014), or the new approach developed by Papastamoulis and Rattray (2018). Some of these tools have also been used for the identification of genes that are differentially expressed under multiple experimental conditions, and this is one of the most common practices in the field (Trapnell et al., 2013; Papastamoulis and Rattray, 2018). However, the mere quantification of expression levels is not enough to acquire a complete picture of the RNA life cycle, and the study of this datum alone could lead to misleading conclusions. Indeed, a cell can regulate gene expression through different fundamental processes. For example, an increase in the expression level of a gene between two conditions is usually interpreted as an intensification of its transcription, although the same observation could be due to a decreased degradation rate. The recent literature has shown that modeling the fundamental processes of the RNA life cycle and estimating the rates at which they occur can contribute greatly to a better understanding of the regulation mechanisms (Rabani et al., 2011, 2014; Furlan et al., 2019; Tesi et al., 2019).

The RNA life cycle can be modeled as a system of linear ordinary differential equations (ODEs) that arises from a network of chemical reactions (see, for example, de Pretis et al., 2015; Anderson and Kurtz, 2015). The time-dependent coefficients of such ODEs, the so-called kinetic rates (KRs), can be interpreted as the instantaneous rates at which the fundamental synthesis, processing and degradation mechanisms occur. In the last few years, several tools have been proposed to infer KRs from RNA-seq data, and of these, DRiLL (Rabani et al., 2014) and INSPEcT (de Pretis et al., 2015; Furlan et al., 2020) provide a characterization of the full RNA life cycle. Motivated by a real data application, focused on the study of the activation of the proto-oncogene Myc in murine fibroblasts (de Pretis et al., 2017), we propose a novel approach to the inference of the RNA life-cycle, cast in a Bayesian framework, that includes several clusterization steps.

To highlight the elements of novelty of our methodology, we here give a short summary of our contribution. We start by remarking that the experimental data are GELs, while our goal is to infer the three KRs, which are latent time and gene dependent functions. In our experimental setting, the cell regulates the expression of groups of genes to respond to a stimulus. In this common biological scenario, (Allocco et al., 2004), the KRs are expected to take on several different but typical shapes. To infer such shapes, we define a single parametric family of functions that is sufficiently flexible to cover all of them. In this way, we recast our problem into the framework of parametric inference, and therefore do not need any a-posteriori model selection step. As a further layer of complexity, which has not yet been added to similar models, we define two distinct mixture models for each KR, that group genes with similar baselines or temporal responses to Myc activation into clusters. We decided to maintain these clusterizations independent, because genes with different/similar initial rates may respond to the treatment in similar/different ways.

The complexity of the model is such that certain devices are needed to make the implementation reasonably simple and, at the same time, the computational cost acceptable.

The most relevant source of complexity is that the six mixture models are defined over latent quantities; this is rare in the literature, since mixture models are generally defined over the observed data. Moreover, the emission distribution of the mixture models should be defined on the space where the KR parameters are identifiable. Unfortunately, due to the complexity of this space, such a distribution would be difficult to define and handle. Our solution is to define an approximated likelihood that allows us to use a Gaussian emission distribution over a suitable modified set of parameters, and to mimic the spike and slab mechanism for two of the parameters which, from an interpretative point of view, must be allowed to assume the value 0 with positive posterior probability.

Unlike other proposals, we estimate likelihood parameters, variance of the measurement error, the scaling factor, KR functions, and clusterings in a single Bayesian model. Moreover our proposal is able to extract and exploit information shared by the elements of the same cluster, thus resulting in better estimates of the parameters. We demonstrate the inferential gain provided by our approach, by detecting small but significant modulations of post-transcriptional rates (i.e. processing and degradation) which had not been appreciated in previous analyses of the same dataset (de Pretis et al., 2015). Since Myc is a transcription factor, the synthesis rate is the most informative layer of regulation, and clusterization of its parameters provides groups of genes that are similarly involved in basic cellular processes, cancer-related processes, and in the RNA metabolism. Moreover, we manage to identify pervasive modulations of post-transcriptional rates, most likely due to either secondary regulations or the adaptation of processing and degradation to the transcriptional induction following Myc activation.

The paper is organised as follows. We start by describing the experiment used to study Myc activation, and the resulting dataset (Section 2). We then present the mathematical model we use to describe the RNA life cycle (Section 3) and the function we developed to parametrize the KRs (Section 3.3); we also discuss the solutions to some identifiability issues. We proceed by formalising the latent clustering models and their practical application to study Myc activation (Section 3.4). The final section of the paper (Section 4) regards a comparison of our novel Bayesian approach with other methods and a discussion of our results. We conclude with a critical summary of our work and some perspectives (Section 5). The online supplemental material (SM), available on the web page of the journal, contains additional figures.

2. The Data

Our dataset, taken from de Pretis et al. (2017), is organised as illustrated in Figure 1 A. It provides GELs (in Reads Per Kilobase Million, RPKM) of premature, mature, and nascent RNA for more than 10 000 genes, at 11 time-points, for three replicas of the experiment. The experiment is designed to follow the activation of the transcription factor Myc in a murine fibroblast cell-line (3T9) over time. Myc plays a crucial role in the genesis and progression of tumors, and it is involved in the regulation of such basal cellular processes as differentiation, growth and proliferation (Dang, 2012; Chen et al., 2018).

The experiment starts with a population of cells, which is divided into multiple samples, in a stationary biological environment. Each sample is treated to induce Myc activation and, after a different time span, it is sequenced to quantify RNA expression levels. Myc

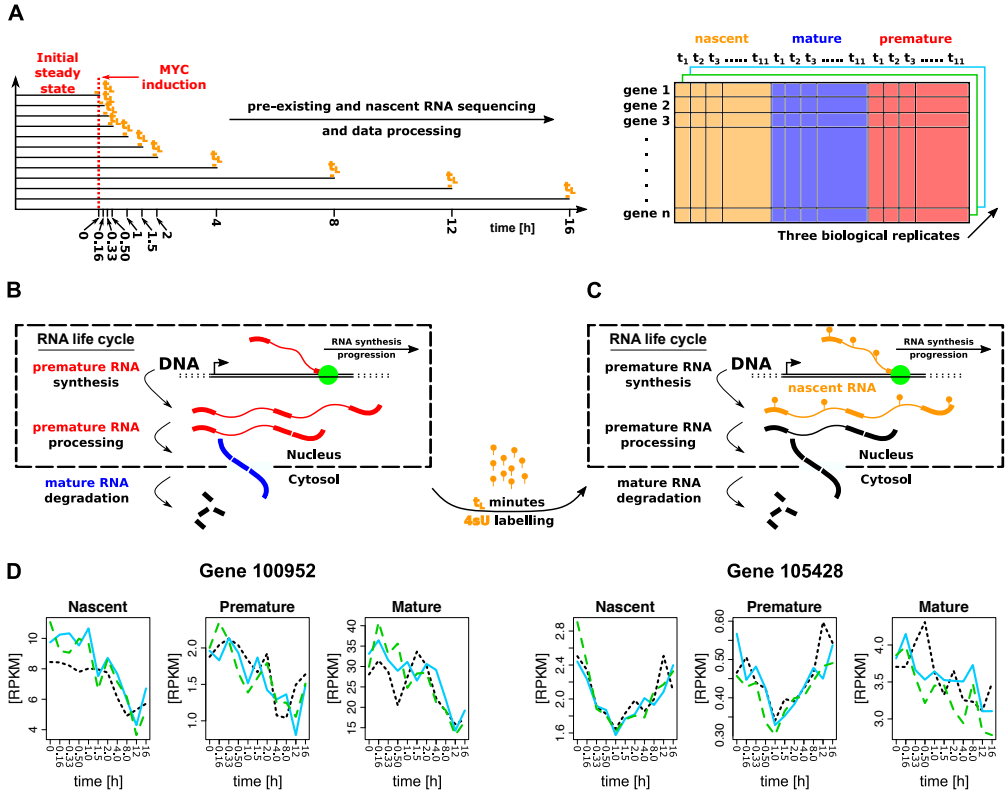


Fig. 1. (A) Experimental design used for the study of Myc activation in 3T9 cells.(B) RNA life cycle in eukaryotic cells and definition of premature and mature RNA. (C) RNA metabolic labeling for nascent RNA quantification. (D) Gene expression profiles for two genes from the dataset; each replicate is represented by a different type of line.

activation is achieved through the expression of an artificial chimera (Littlewood et al., 1995). This protein is natively inactive, and unable to perform any function, but it can rapidly be activated by adding the 4-hydroxytamoxifen (OHT) hormone to the cell culture medium. The authors performed standard (ribo-depleted) RNA-Seq, following Myc activation, through 11 time-points from an OHT treatment: 0h, 1/6h, 2/6h, 1/2h, 1h, 3/2h, 2h, 4h, 8h, 12h, 16h (Figure 1 A). Each experiment, performed on independent samples, was replicated three times, and resulted in expression levels of premature and mature RNA (Figure 1 B). The same experimental design was used to quantify nascent RNA through 4sU-Seq (Figure 1 C). In this case, an exogenous nucleotide (4-thiouridine or 4sU) is provided to the cells before sequencing for a fixed span of time (labeling time). 4sU is incorporated in the transcripts produced during the entire labeling time (nascent RNA) and is later exploited to physically separate them from the pre-existing RNA molecules. This portion of the transcriptome can be sequenced through standard RNA-Seq (Dölken et al., 2008).

We focus on a set of 4909 transcriptional units, classified as Myc targets through a chromatin immunoprecipitation sequencing experiment, and altered in their kinetics.

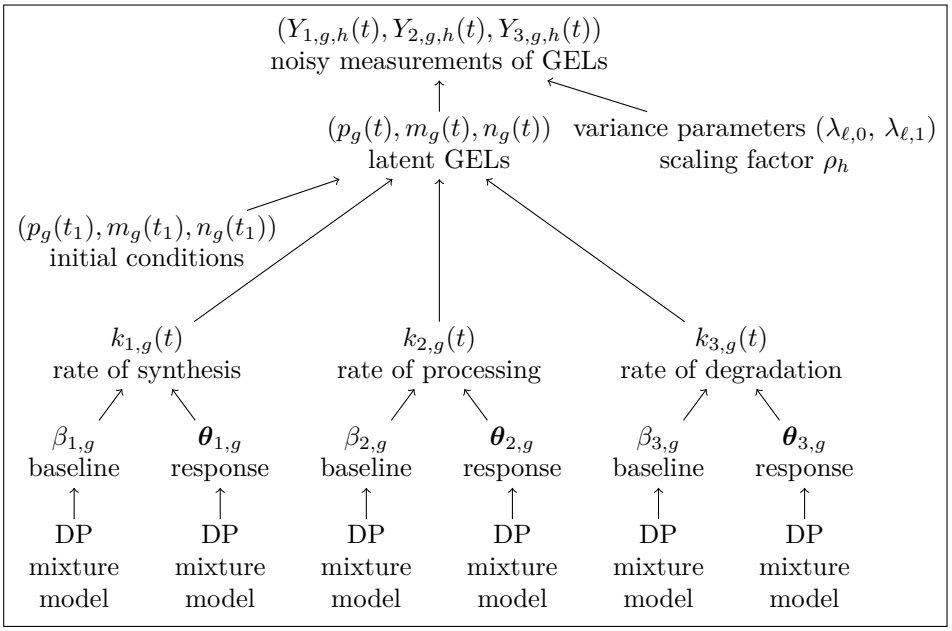


Fig. 2. Graphical representation of the model structure.

However, it was not possible to analyze 12 transcriptional units, because they had negative expression levels. In the end, we retrieved a dataset of premature, mature and nascent RNA expression levels for 4897 genes in 3 replicates and for 11 time-points. Figure 1 D reports two examples from the dataset: the first one represents a typical transcriptional regulation, as can be seen from the adherence of the three profiles, while the second one may be associated with a more complex post-transcriptional scenario.

3. Methods

We provide a schematic representation of our Bayesian hierarchical model in Figure 2, in the form of a tree diagram whose elements are described in detail in the next sections and summarized hereafter.

For each gene $g \in \{1, \cdots, G\}$, time-point $t \in \{t_1, \cdots, t_T\}$ and replica $h \in \{1, \cdots, H\}$, $Y_{\ell,g,h}(t)$ denote the *measured* GEL of premature RNA if $\ell = 1$, mature RNA if $\ell = 2$, and nascent RNA if $\ell = 3$. These are noisy and scaled versions of the true unobserved values, which we indicate with $p_g(t)$, $m_g(t)$ and $n_g(t)$. In our dataset, $G = 4897$, $T = 11$, and $H = 3$. The distribution of the observed GELs depends on the parameters $\rho_h(t)$, $\lambda_{\ell,0}$, and $\lambda_{\ell,1}$, and is described in Section 3.1. The three *latent* GELs, $(p_g(t), m_g(t), n_g(t))$, can be modeled as the solution of a system of ordinary differential equations that is specified by three time-dependent KRs $k_{j,g}(t)$, with $j = 1, 2, 3$. These KRs can be interpreted as the rates at which the fundamental mechanisms underlying the RNA life cycle, namely synthesis, processing and degradation, occur; see Section 3.2. A parametric family is introduced to define these KRs, see Section 3.3, and we show how the parameters can conveniently be subdivided into two sets, $\beta_{j,g}$ and $\boldsymbol{\theta}_{j,g}$. The former is a scalar variable

related to the baseline value of the KR before Myc activation, while the latter is composed of four parameters which characterize the shape of the response to the stimulus. We define for each of these parameters ($\beta_{j,g}$ and $\boldsymbol{\theta}_{j,g}$), in each rate ($k_{j,g}(t)$), a mixture model that groups genes with similar baselines ($\beta_{j,g}$) or temporal responses to Myc activation ($\boldsymbol{\theta}_{j,g}$), into clusters. We decided to keep the clusterization on the three $\beta_{j,g}$ distinct from that of the three $\boldsymbol{\theta}_{j,g}$ because the relation between the basal value of the rates and their response to stimuli is not trivial, and, genes with different (or similar) initial rates may therefore respond to the treatment in similar (or different) ways.

3.1. The noise model

Since each $Y_{\ell,g,h}(t)$ needs to be positive, for the noisy observed GELs we assume

$$\begin{aligned} &Y_{1,g,h}(t) \sim N_{>0}(p_g(t), \tau_{1,g}(t)), \qquad Y_{2,g,h}(t) \sim N_{>0}(m_g(t), \tau_{2,g}(t)), \\ &Y_{3,g,h}(t) \sim N_{>0}(n_g(t)\rho_h(t), \tau_{3,g}(t)), \end{aligned} \tag{1}$$

where $N_{>0}(\cdot,\cdot)$ indicates a truncated normal distribution defined over $\mathbb{R}^+$, and $\rho_h(t)$ is a scaling factor, that is required to normalize the nascent RNA libraries to the pre-existing RNA counterparts (Rabani et al., 2011, 2014; de Pretis et al., 2015). As shown in previous works, see, for example, Subramaniam and Hsiao (2012), the GELs affect both the mean and the variance of the measurement and we therefore assume the following:

$$\begin{aligned} &\log(\tau_{1,g}(t)) = \lambda_{1,0} + \lambda_{1,1}\log(p_g(t)), \qquad \log(\tau_{2,g}(t)) = \lambda_{2,0} + \lambda_{2,1}\log(m_g(t)), \\ &\log(\tau_{3,g}(t)) = \lambda_{3,0} + \lambda_{3,1}\log(n_g(t)). \end{aligned} \tag{2}$$

It should be noted that all the replicates of the expression levels of a gene share the same parameters, and, by introducing replicas we therefore increase the information on the latent GELs.

3.2. A mathematical model of the RNA life cycle

The life cycle of RNA molecules in eukaryotic cells, is divided into three sub-processes (Figure 1 B). The first one is the synthesis of premature RNA from DNA. This portion of the transcriptome is located inside the nucleus and it is not ready to perform its task (e.g. protein translation). Premature RNA requires structural modifications and/or exporting to the cytosol. This second stage of the RNA life cycle is named processing and its product is the mature RNA, which is eventually degraded by the cell (third step). The process may be described by the following network of chemical reactions

$$\emptyset \xrightarrow{k_{1,g}(t)} P_g \xrightarrow{k_{2,g}(t)} M_g \xrightarrow{k_{3,g}(t)} \emptyset, \qquad g = 1, \cdots, G, \tag{3}$$

where P_g and M_g denote premature and mature RNA for gene g, respectively. The empty-set symbols are used to emphasize that premature RNA is synthesized from DNA without consuming any reactant, and mature RNA is subjected to degradation without forming any product. A system of ODEs that translates the reaction network (3) into mathematical terms is

$$\begin{cases} \dot{p}_g(t) = -k_{2,g}(t)p_g(t) + k_{1,g}(t) \\ \dot{m}_g(t) = k_{2,g}(t)p_g(t) - k_{3,g}(t)m_g(t), \end{cases} \tag{4}$$

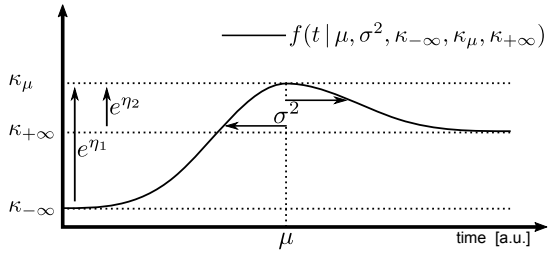


Fig. 3. Graphical representations of the KR parametrization. It should be noted that e^{η_1} and e^{η_2} are shown to indicate the section of the function they determine, but they are not equal to the length of the arrows, see equation (8).

where the dots denote time derivatives. The effect of the processing is to decrease $p_g(t)$ and correspondingly increase $m_g(t)$ at rate $k_{2,g}(t)$. The degradation decreases $m_g(t)$ at rate $k_{3,g}(t)$, while the synthesis increases $p_g(t)$ at rate $k_{1,g}(t)$.

It is well known that, for the model described so far, it is difficult to identify all three KRs. Measurements of another variable, the so-called nascent RNA (Dölken et al., 2008; Rabani et al., 2011, 2014; de Pretis et al., 2015), are usually included to ameliorate the identifiability, cf. Figure 1 C. Nascent RNA is defined as the amount of total RNA, premature plus mature, which is produced by the cell in a short span of *labeling* time t_L, cf. Figure 1. By definition, the nascent RNA is absent at the beginning of the labeling time, and it is produced according to the same dynamics as the pre-existing counterpart during t_L. However, the effect of degradation can be neglected in such a short span. The expression level of the premature ($p_g^\star(t)$) and mature ($m_g^\star(t)$) nascent RNA is therefore ruled by the following equations

$$\begin{cases} \dot{p}_g^\star(t) = -k_{2,g}(t)p_g^\star(t) + k_{1,g}(t), \\ \dot{m}_g^\star(t) = k_{2,g}(t)p_g^\star(t). \end{cases}$$

The sum $n_g(t) = p_g^\star(t) + m_g^\star(t)$ is the *nascent* RNA level. By summing the previous equations, one has that $n_g(t)$ only varies as a result of the effect of the synthesis:

$$\dot{n}_g(t) = k_{1,g}(t). \tag{5}$$

Since the time window for which nascent RNA evolves is short (t_L), this rate can be considered approximately constant, and equation (5) can be integrated to obtain $n_g(t) = k_{1,g}(t)t_L$, which is a third equation that has to be added to model (4) to facilitate the estimation of $k_{1,g}(t)$. It should be noted that the initial conditions $(p_g(t_1), m_g(t_1))$ need to be known to solve the ODE (4), and are here considered as further model parameters.

3.3. KR parametrization

In several biological experiments, a cell culture is perturbed and gene expression levels are repeatedly measured before and after the perturbation in order to understand which genes are involved in the response. In this case, the KRs typically vary over time and the typical shapes that we expect them to take on have the following characteristics: steady at the beginning of the experiment, steady after a long time from the perturbation, and

Table 1. KR shapes as functions of the log-fold-changes and the names we use to describe the shapes.

η_1 sign	η_2 sign	names			shape
0	0	constant			
+	+	peak-like	peak-like+		
-	-		peak-like-		
+	0	monotonic	monotonic+	up-c	
+	-			up-up	
0	-			c-up	
0	+		monotonic-	c-down	
-	+			down-down	
-	0			down-c	

varying (with some regularity) in the transient region between the two steady-states. Some typical shapes are: constant (some rates are not altered at all), monotonic (both increasing and decreasing), and peak-like functions. They have already been successfully applied to describe transcriptional and post-transcriptional responses in several biological systems (see, for example, Chechik and Koller, 2009; Rabani et al., 2011, 2014; de Pretis et al., 2015). More complex patterns would be difficult to identify with the number of time-points and replicates that are usually available.

We introduce a unique parametric family of functions which, for different values of the parameters, can cover all such characteristic shapes. Let $\phi(\cdot|\mu,\sigma^2)$ be a Gaussian density with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}^+$. We define the family of functions $f(t\,|\,\mu,\sigma^2,\kappa_{-\infty},\kappa_\mu,\kappa_{+\infty})$ to which all the KRs belong in the following way:

$$f(t\,|\,\mu,\sigma^2,\kappa_{-\infty},\kappa_\mu,\kappa_{+\infty}) = \begin{cases} \kappa_{-\infty} + \frac{\phi(t|\mu,\sigma^2)}{\phi(\mu|\mu,\sigma^2)}(\kappa_\mu - \kappa_{-\infty}), & \text{if } t < \mu, \\ \kappa_{+\infty} + \frac{\phi(t|\mu,\sigma^2)}{\phi(\mu|\mu,\sigma^2)}(\kappa_\mu - \kappa_{+\infty}), & \text{if } t \geq \mu, \end{cases} \tag{6}$$

where $\kappa_{-\infty}$, κ_μ and $\kappa_{+\infty}$ all belong to $\mathbb{R}^+$. Function f in equation (6) is obtained by applying different scalings and vertical translations of a Gaussian density to its right and left halves, with respect to the mean value μ, taking care to preserve continuity at time-point $t = \mu$ (Figure 3). It is easy to see that:

$$\begin{aligned} \kappa_{-\infty} &= \lim_{t\to-\infty} f(t\,|\,\mu,\sigma^2,\kappa_{-\infty},\kappa_\mu,\kappa_{+\infty}), \\ \kappa_\mu &= f(\mu\,|\,\mu,\sigma^2,\kappa_{-\infty},\kappa_\mu,\kappa_{+\infty}), \\ \kappa_{+\infty} &= \lim_{t\to+\infty} f(t\,|\,\mu,\sigma^2,\kappa_{-\infty},\kappa_\mu,\kappa_{+\infty}). \end{aligned} \tag{7}$$

For easiness of interpretation, we split and rename the parameters as follows. First, we single out $\kappa_{-\infty}$ and we rename it β to simplify the notation. Unlike the other parameters, which are related to the response, β is the baseline level, i.e. the initial steady-state, and

it is analyzed separately. Secondly, we introduce the logarithmic ratios

$$\eta(t', t) = \log \frac{f(t' \,|\, \mu, \sigma^2, \kappa_{-\infty}, \kappa_\mu, \kappa_{+\infty})}{f(t \,|\, \mu, \sigma^2, \kappa_{-\infty}, \kappa_\mu, \kappa_{+\infty})}.$$

These quantities are called *log-fold-changes* in computational biology and are usually used to measure modulations with respect to the baseline level $f(t \,|\, \mu, \sigma^2, \kappa_{-\infty}, \kappa_\mu, \kappa_{+\infty})$, and we therefore define

$$\eta_1 = \eta(\mu, -\infty) = \log \frac{\kappa_\mu}{\kappa_{-\infty}}, \quad \eta_2 = \eta(\mu, +\infty) = \log \frac{\kappa_\mu}{\kappa_{+\infty}}. \tag{8}$$

Parameters μ, σ^2, η_1, η_2 are all related to the characterization of the response to perturbations. In particular, μ and σ^2 characterize the *temporal* location and duration of the response, while η_1 and η_2 determine the typical shape, as highlighted in Table 1. We collect these four parameters in a single vector that we denote $\boldsymbol{\theta}$ to obtain a more compact notation. Examples of the forms that can be obtained with (6), by changing its parameters, are shown in Table 1, where we can see that all the standard shapes (constant, increasing/decreasing, peak-like) are possible. The family of functions (7) can now be re-parametrized as $f(t \,|\, \beta, \boldsymbol{\theta})$, with $\beta = \kappa_{-\infty}$ and $\boldsymbol{\theta} = (\mu, \sigma^2, \eta_1, \eta_2)$.

Although the family of functions (7) is well defined for all real values of μ, η_1, and η_2, and for positive values of σ^2, certain identifiability and interpretability issues may arise if some conditions are not met. For example, if μ is smaller than the first observed time t_1, and σ^2 is small (compared to $|\mu - t_1|$), function f in the interval $[t_1, t_T]$, is indistinguishable for any arbitrary choice of η_1 and η_2 from a constant one, which should instead be given by $\eta_1 = \eta_2 = 0$ (Table 1). For this reason, identifiability constraints are needed. One main requirement is that the value of the function (6) at time-points t_1 and t_T should be "close" to β and κ_∞, respectively, which means that the most relevant part of the function graph lies within the observed time window. Hence, for peak-like shapes ($\eta_1\eta_2 > 0$, cf. Table 1), we require

$$\left|f(t_1 | \, \beta, \boldsymbol{\theta}) - \beta\right| \leq 0.01|\kappa_\mu - \beta|, \quad \text{and} \quad \left|f(t_T | \, \beta, \boldsymbol{\theta}) - \kappa_{+\infty}\right| \leq 0.01|\kappa_\mu - \kappa_{+\infty}|, \tag{9}$$

which implies two conditions:

$$\mu - \sqrt{-2\sigma^2 \log(0.01)} \geq t_1 \qquad \text{and} \qquad \mu + \sqrt{-2\sigma^2 \log(0.01)} \leq t_T. \tag{10}$$

The value 0.01, or any other value with comparable magnitude, ensures that the derivative of our function is almost zero at time t_1 and t_T (it is ≈ 0.01 in (10)) and then $f(t_1 | \, \beta, \boldsymbol{\theta}) \approx \beta$, which lets us interpret β as the steady-state at time t_1, and $f(t_T | \, \beta, \boldsymbol{\theta}) \approx \kappa_\infty$, which ensures a steady-state at the end of the observed window. Considering a value that us too small would restrict the possible values of (μ, σ^2), without any significant improvement in the steady-state approximation, while such an approximation may no longer be acceptable for a value that is too large. It should be noted that one of the log-fold-changes vanishes for monotonic shapes; however, we can still derive an identifiability condition, provided that (9) holds, e.g. when $\eta_1 = 0$, we have a c-up or c-down shape, and $\kappa_\mu = \beta$. Since the function is constant from t_1 to μ, the first equation of (9) holds if, and only if, $t_1 \leq \mu$. The conditions we obtain for c-up and c-down shapes ($\eta_1 = 0$) are therefore

$$\mu \geq t_1 \qquad \text{and} \qquad \mu + \sqrt{-2\sigma^2 \log(0.01)} \leq t_T. \tag{11}$$

Similarly, for up-c and down-c shapes ($\eta_2 = 0$), we obtain

$$\mu - \sqrt{-2\sigma^2 \log(0.01)} \geq t_1 \qquad \text{and} \qquad \mu \leq t_T. \tag{12}$$

The monotonic up-up (down-down) is an intermediate shapes between c-up and up-c (c-down and down-c). Hence, the limits on μ should be close to (11), if $\eta_1 < \eta_2$, and close to (12), if the opposite is true. Therefore, we define the constraints on μ as:

$$\mu - \left|\frac{\eta_1}{\eta_1 - \eta_2}\right| \sqrt{-2\sigma^2 \log(0.01)} \geq t_1, \qquad \mu + \left(1 - \left|\frac{\eta_1}{\eta_1 - \eta_2}\right|\right) \sqrt{-2\sigma^2 \log(0.01)} \leq t_T. \tag{13}$$

The subset of the parameter space where identifiability holds is denoted as $\mathcal{D} \subset \mathbb{R}^5$, and is defined by the condition $\beta > 0$ and one of inequalities (10), (11), (12) or (13), depending on the signs of η_1 and η_2, which are instead unconstrained real numbers. Although identifiability is only granted in $\mathcal{D}$, we prefer not to force the parameter to belong to this set for easiness of implementation, and, as we explain in the next section, we introduce an approximated likelihood that gives very little support to values outside $\mathcal{D}$.

3.4. Latent clustering models

It is biologically reasonable to allow groups of genes that have a similar baseline $\beta_{j,g}$ level to respond differently to a perturbation (therefore with different $\boldsymbol{\theta}_{j,g}$). Moreover, even though two genes have a similar rate of synthesis, it is not necessarily true that the other rates are similar. For this reason, we separately introduce mixture models, based on the DP, to the bottom level of the model hierarchy, for the parameters $\beta_{j,g}$ and $\boldsymbol{\theta}_{j,g}$ for each $j \in \{1, 2, 3\}$, see Figure 2. Making inference on such hidden mixtures is particularly hard and makes the implementation of the MCMC more involved. For this reason, it is necessary to ensure that the sampling of the DP is as straightforward as possible, and the use of Gibbs updates for the mixture parameters is preferable. Moreover, we also want to ensure that the model can discriminate between the possible shapes of $k_{j,g}(t)$, and that our MCMC algorithm does not become trapped in a region of the parameter space where such shapes are not identifiable. To this aim, the most natural choice would be to adopt a distribution with support in $\mathcal{D}$ for $(\beta_{j,g}, \boldsymbol{\theta}_{j,g})$; however, this choice is impractical, since is not possible to find an easy-to-handle distribution over this complicated domain. The solution we propose here is to let $(\beta_{j,g}, \boldsymbol{\theta}_{j,g})$ be defined over the whole $\mathbb{R}^5$ and to modify the data likelihood (equation (1)) with an alternative one that gives an almost null posterior support to all the parameter values outside $\mathcal{D}$. Our proposal is:

$$\begin{cases} Y_{\ell,g,h}(t) \sim N_{>0}(x_{\ell,g}(t), \tau_{\ell,g}(t)), & \text{if } (\beta_{j,g}, \boldsymbol{\theta}_{j,g}) \in \mathcal{D} \, \forall j = 1, 2, 3, \\ Y_{\ell,g,h}(t) \sim N(\mathbf{0}, 10^{10000}), & \text{otherwise,} \end{cases}$$

where $x_{1,g}(t) = p_g(t)$, $x_{\ell,g}(t) = m_g(t)$ and $x_{3,g}(t) = n_g(t)\rho_h(t)$. By doing so, we enforce the posterior distribution to have a very small mass outside $\mathcal{D}$, which means that the parameters in the unidentifiable region are almost never sampled in the MCMC algorithm. A second issue is that if the marginal distribution of the log-fold-change $\eta_{1,j,g}$ is continuous, the posterior probability to be equal to 0 vanishes, which means that we are not able to estimate a constant $k_{j,g}(t)$ (or c-up, c-down, up-c, and down-c shape). One possible

solution is to use a distribution that is continuous over $(-\infty, 0) \cup (0, \infty)$ and has a point mass at 0. We can do this by introducing the new variables $\eta^*_{1,j,g}$ and $\eta^*_{2,j,g}$, related to $\eta_{1,j,g}$ and $\eta_{2,j,g}$ through the following:

$$\eta_{i,j,g} = \begin{cases} \max(0, \eta^*_{i,j,g} - \xi), & \text{if } \eta^*_{i,j,g} > 0, \\ \min(0, \eta^*_{i,j,g} + \xi), & \text{otherwise,} \end{cases}, \quad i = 1, 2. \tag{14}$$

If we assume a continuous distribution for $\eta^*_{1,j,g}$, as a result of (14), the distribution over $\eta_{1,j,g}$ is continuous over $(-\infty, 0) \cup (0, \infty)$ and has a point mass at 0 equal to the cumulative distribution of $\eta^*_{1,j,g}$ between $-\xi$ and ξ. A similar result holds for $\eta_{2,j,g}$. Equation (14), therefore defines a spike-and-slab distribution for $\eta_{1,j,g}$ and $\eta_{2,j,g}$ (Ishwaran and Rao, 2005), which has a spike (point mass) at 0 and a slab (continuous distribution) over $\mathbb{R}$. However, our approach allows us to work with the continuous "latent" variables $\eta^*_{1,j,g}$ and $\eta^*_{2,j,g}$, which makes the implementation of the MCMC easier. It should also be noted that the posterior probability mass on $\eta_{i,j,g} = 0$ and $\eta_{2,j,g} = 0$ depends on their marginal distributions, e.g., if the mean is zero and the variance is small/large, the mass is close to 1/negligible. As shown in equation (15), the means and variances of these distributions are parameters that are inferred from the model fitting and, more importantly, they are cluster dependent. Therefore, also the posterior probability mass on $\eta_{1,j,g} = 0$ and $\eta_{2,j,g} = 0$ is cluster dependent.

We can now work with the parameters $\beta_{j,g}$ and $\boldsymbol{\theta}^*_{j,g} = (\mu_{j,g}, \sigma^2_{j,g}, \eta^*_{1,j,g}, \eta^*_{2,j,g})$, which are all defined over $\mathbb{R}$. We then define 6 mixture models, based on Gaussian densities, 3 of them over $\beta_{1,g}$, $\beta_{2,g}$ and $\beta_{3,g}$ and the others over $\boldsymbol{\theta}^*_{1,g}$, $\boldsymbol{\theta}^*_{2,g}$ and $\boldsymbol{\theta}^*_{3,g}$. In other words, the models are DP Gaussian mixtures (Neal, 2000) formalized as:

$$\begin{aligned} \beta_{j,g} \,|\, \zeta^{\beta}_{j,z^{\beta}_{j,g}}, \omega^{\beta}_{j,z^{\beta}_{j,g}}, z^{\beta}_{j,g} &\sim N(\zeta^{\beta}_{j,z^{\beta}_{j,g}}, \omega^{\beta}_{j,z^{\beta}_{j,g}}), & \boldsymbol{\theta}^*_{j,g} | \boldsymbol{\zeta}^{\boldsymbol{\theta}}_{j,z^{\boldsymbol{\theta}}_{j,g}}, \boldsymbol{\Omega}^{\boldsymbol{\theta}}_{j,z^{\boldsymbol{\theta}}_{j,g}}, z^{\boldsymbol{\theta}}_{j,g} &\sim N(\boldsymbol{\zeta}^{\boldsymbol{\theta}}_{j,z^{\boldsymbol{\theta}}_{j,g}}, \boldsymbol{\Omega}^{\boldsymbol{\theta}}_{j,z^{\boldsymbol{\theta}}_{j,g}}), \\ z^{\beta}_{j,g} \,|\, \boldsymbol{\pi}^{\beta}_j &\sim \text{Discrete}(\boldsymbol{\pi}^{\beta}_j), & z^{\boldsymbol{\theta}}_{j,g} | \boldsymbol{\pi}^{\boldsymbol{\theta}}_j &\sim \text{Discrete}(\boldsymbol{\pi}^{\boldsymbol{\theta}}_j), \\ \boldsymbol{\pi}^{\beta}_j \,|\, \alpha^{\beta}_j &\sim \text{GEM}(\alpha^{\beta}_j), \quad , & \boldsymbol{\pi}^{\boldsymbol{\theta}}_j | \alpha^{\boldsymbol{\theta}}_j &\sim \text{GEM}(\alpha^{\boldsymbol{\theta}}_j), \\ \zeta^{\beta}_{j,k} &\sim N(M_0, V_0), & \boldsymbol{\zeta}^{\boldsymbol{\theta}}_{j,k} &\sim N(\mathbf{M}, \mathbf{V}), \\ \omega^{\beta}_{j,k} &\sim IW(\nu_0, \psi_0), & \boldsymbol{\Omega}^{\boldsymbol{\theta}}_{j,k} &\sim IW(\nu, \boldsymbol{\Psi}), \end{aligned} \tag{15}$$

where $k \in \mathbb{N}$. In models (15), $z^{\beta}_{j,g}$ and $z^{\boldsymbol{\theta}}_{j,g}$ are the discrete random variables which represent the labels which identify the component of the mixture to which the parameters belong. These variables are assumed to come from a discrete distribution, whose probabilities follow a Dirichlet Processes defined by the GEM (or stick-breaking) distribution (Gnedin et al., 2001). Given the allocation variables $z^{\beta}_{j,g}$ and $z^{\boldsymbol{\theta}}_{j,g}$, parameters $\beta_{j,g}$ and $\boldsymbol{\theta}^*_{j,g}$ are normally distributed. All the random quantities in model (15) can easily be updated in the MCMC algorithm using only Gibbs steps, thereby facilitating the implementation of the model.

4. Real data application

The results obtained by means of our model on the real data application are presented in this section. Before discussing these results, we offer some implementation details to

explain how we process the model output to obtain the statistics used to interpret the results.

4.1. Implementation details

Priors. We choose weakly informative priors for all the model parameters. We set $\mathbf{M} = \mathbf{0}$, $\mathbf{V} = 100\mathbf{I}$, $M_0 = 0$, $V_0 = 100$, $\mathbf{\Psi} = \mathbf{I}$, $\nu = 5$, $\nu_0 = 2$, and $\alpha_j^{\beta}, \alpha_j^{\boldsymbol{\theta}} \sim G(1,1)$ for the DP components of our proposal (equation (15)). We choose a normal distribution $N(0,100)$ for the variance parameters $\lambda_{\ell,0}$ and $\lambda_{\ell,1}$ (equation (2)), and we select a Gamma $G(1,1)$ for the scaling factor $\rho_h(t)$ (equation (1)). A uniform $U(0,10000)$ is used as a prior for the initial conditions $n_g(t)$ and $m_g(t)$ (equation (4)) of the ODE system, where 10000 is the maximum level we can expect to see. We also assume $\xi = 10$ but, through preliminary simulation experiments, we verified that if $\xi < 20$, and considering the given priors, its value has almost no influence on the posterior inference.

Implementation. The model is implemented in R/C++ (R Core Team, 2020) and uses OpenMP (OpenMP Architecture Review Board, 2008) for parallel computing. We use the sampling scheme proposed by Escobar and West (1995) for the DP hyper-parameters α_j^{β} and $\alpha_j^{\boldsymbol{\theta}}$. We marginalize with respect to the state probabilities $\boldsymbol{\pi}_j^{\beta}$ and $\boldsymbol{\pi}_j^{\boldsymbol{\theta}}$, and we use the algorithm number 8 of Neal (2000) (with parameter $m = 1$) at each MCMC iteration and for each mixture model, to perform a global update of the latent allocation variables $z_{j,g}^{\beta}$ and $z_{j,g}^{\boldsymbol{\theta}}$, and a single step of the "restricted Gibbs sampling split-merge" of Jain and Neal (2007) (with parameters (0,1,0,0)), to split and merge existing states. Both updates are easily implemented since, given the chosen priors, the means and variances of the normal densities can easily be sampled. Parameters $\boldsymbol{\theta}_{j,g}^*$, $\beta_{j,g}$, and sets $\{\lambda_{1,0}, \lambda_{1,1}, \lambda_{2,0}, \lambda_{2,1}, \lambda_{3,0}, \lambda_{3,1}\}$, $\{\rho_1(t_i), \rho_2(t_i), \rho_3(t_i)\}_{i=1}^T$, and $\{p_g(t_1), m_g(t_1)\}$ are all sampled separately using Metropolis steps with the adaptive proposals of Andrieu and Thoms (2008); algorithm number 4. The ODE solutions are computed with the *odeint* C++ library (Ahnert and Mulansky, 2011). The code that implements the methodology is available at `https://github.com/GianlucaMastrantonio/Multiple_latent_clusterisation_model_for_the_inference_of_RNA_life_cycle_kinetic_rates`.

Posterior samples. We estimate the model on a computer cluster with 32 cores and using a stop-and-go strategy to check whether and when convergence has been achieved. In detail, we run the model using batches of 250 000 iterations, burnin 175 000, and thin 30. We check convergence after each batch using split$-\hat{R}$ statistics (Gelman et al., 2013) computed on parameters $\boldsymbol{\theta}_{j,g}^*$, $\beta_{j,g}^*$, α_j^{β}, $\alpha_j^{\boldsymbol{\theta}}$, $\lambda_{\ell,0}$, $\lambda_{\ell,1}$, $n_g(t)$, $m_g(t)$, and $\rho_h(t)$ and, if not all of them have a lower split$-\hat{R}$ than 1.05 (our selected threshold), we run another batch with the last simulated parameters of the previous MCMC batch as the starting values. We use random values for the first batch. After 9 batches of simulations, all the parameters were within the split$-\hat{R}$ threshold and only then, did we use a 10-th batch of simulations to obtain the posterior samples we use to describe the results. Hence, the model required $250\,000 \times 10 = 2\,500\,000$ iterations, with a total burnin of $250\,000 \times 9 + 175\,000 = 2\,425\,000$, thin 30, having than 2500 posterior samples; the computations took 20 days.

Output processing. Using the algorithm of Wade and Ghahramani (2018), we find a representative point estimate of the cluster membership variables $z_{j,g}^{\beta}$ and $z_{j,g}^{\boldsymbol{\theta}}$, which we indicate with $\hat{z}_{j,g}^{\beta}$ and $\hat{z}_{j,g}^{\boldsymbol{\theta}}$, while we use the posterior means $\hat{\beta}_{j,g}$, $\hat{\mu}_{j,g}$, $\hat{\sigma}_{j,g}^2$, $\hat{\eta}_{1,j,g}$, $\hat{\eta}_{2,j,g}$ as

Table 2. Fraction of the CRPS values which satisfy the condition reported at the top of the grid.

	CRPS M1≤CRPS M2	CRPS M1≤CRPS M3	CRPS M2≤CRPS M3
premature	**0.650**	0.517	0.489
mature	**0.653**	0.517	0.501
nascent	**0.643**	0.528	0.508

Table 3. For each model and function k, we indicate the fraction of constant functions, decreasing monotonic or peak-like functions (monotonic- and peak-), and increasing monotonic or peak-like functions (monotonic+ and peak+)

		constant	peak+ or monotonic+	peak- or monotonic-
M1	$k_{1,\cdot}(\cdot)$	0.019	0.415	0.566
	$k_{2,\cdot}(\cdot)$	0.515	0.342	0.143
	$k_{3,\cdot}(\cdot)$	0.460	0.063	0.476
M2	$k_{1,\cdot}(\cdot)$	0.100	0.408	0.492
	$k_{2,\cdot}(\cdot)$	0.641	0.198	0.161
	$k_{3,\cdot}(\cdot)$	0.703	0.150	0.147
M3	$k_{1,\cdot}(\cdot)$	0.013	0.357	0.630
	$k_{2,\cdot}(\cdot)$	0.712	0.081	0.207
	$k_{3,\cdot}(\cdot)$	0.801	0.118	0.081

representative values of the KR parameters. Since the posterior probability that a KR has one of the 5 shapes defined in Table 1 depends on whether $\eta_{1,j,g}$, $\eta_{2,j,g}$ are positive, negative or null, we can associate the representative shape as the one with the highest posterior probability to each KR.

4.2. Comparison of the multiple inference method

We compare the performance of our model (M1) with a simplified version (M2), which assumes $z^{\beta}_{j,g} = z^{\boldsymbol{\theta}}_{j,g} = 1$ for all j and g (i.e. the mixture has a single component). We use the same priors as M1 for the estimation of M2, and the same strategy to check for the MCMC convergence. We use the Deviance Information Criterion DIC $= -2f(\mathbf{y}|\hat{\boldsymbol{\chi}}) + 4\mathrm{Var}_{\mathrm{post}}(f(\mathbf{y}|\boldsymbol{\chi}))$ (Gelman et al., 2014) to compare the models, where $f(\mathbf{y}|\cdot)$ is the data likelihood given by the distributions in (1), $\boldsymbol{\chi}$ is the vector of the likelihood parameters computed using the posterior means of our model, and $\mathrm{Var}_{\mathrm{post}}(\cdot)$ is the variance with respect to the posterior distribution. The value of DIC is 209400 for M1 and 257268 for M2, thus showing that our proposal is preferable. Since our approach can be though of as a Bayesian extension of the frequentist framework implemented in INSPEcT (M3), a comparison between these two methods would also be interesting, despite the absence of obvious ways of comparing Bayesian and frequentist approaches in the literature. One possible solution is to use cross-validation indices, but this would require fitting the models multiple times which, given the significant computational cost of our proposal, would be prohibitive. We approximate the cross-validation predictive distribution for a given gene/replica/time-point with the standard predictive distribution, and we use the continuous ranked probability score (CRPS) (Gneiting and Raftery, 2007) for the comparison. CPRS is often used to assess the accuracy of probabilistic models, and it is defined as the

mean square error between the predicted and empirical cumulative distribution functions:

$$\mathrm{CRPS}(y) = \int_{-\infty}^{\infty} \left[F(x) - \mathbb{1}(y - x)\right]^2 dx,$$

where $F(\cdot)$ is the predictive distribution and $\mathbb{1}$ is the Heaviside step function. We estimate the CRPS for each of our observations and, for comparison purposes, we compute the fraction of data points with a lower index for one model than another one; we decided not to average across all the data points because they are not identically distributed. These comparisons are recapitulated in Table 4.2. Further insight can be gained by computing pairwise Pearson's correlations between the CRPS indexes estimated for premature, mature and nascent RNA for each model. The correlations are between 0.34 and 0.62, and interestingly, they are always higher for M1 and M2, than for M3 (0.52, 0.52 and 0.38 on average, respectively, Figure 37-SM). It follows that the Bayesian approaches tend to fit all the RNA species profiles with a similar goodness, while the frequentist one generally recapitulates the profile of one experimental quantity better than the others. This is not desirable in the current application scenario, since any RNA species is potentially equally informative about the underlying regulations of the RNA metabolism.

As a result of the way we compute the CRPS, i.e. without using holdout observations, its value may be influenced by overfitting. Hence, we also decided to compare M1 and M3 focusing on the different features and interpretability of the estimated KRs, showing that our proposal (M1) gives a better and clearer description of the data. Therefore we report the proportions of constant and variable rates in Table 3, both increasing (monotonic+ and peak+) and decreasing (monotonic- and peak-) for each model. Since Myc is a transcription factor, we expect a primary response at the synthesis level. This is in particular the case for M1 and M3, which have less than 2% of the genes constant in synthesis. The percentage is higher for M2 (10%). Nevertheless, k_1 is still by far the most variable rate. We can also observe an agreement about the direction of the modulations for all the inference methods which predict a higher percentage of repressed genes than induced ones ($0.49 - 0.63$ versus $0.36 - 0.42$). The inferred regulations are more different for the post-transcriptional rates; indeed, M3 predicts a higher percentages of constant k_2 and k_3 (71% and 80%, respectively) than the Bayesian approaches (between 46% and 70%). This means that M1 and M2 tend to explain the expression more as a simultaneous regulation of all the three kinetic rates than M3 (Figure 4 and 38-SM). This choral regulation, which was not observed to the same extent in past analyses (M3), is biologically reasonable. Indeed, the role of synthesis remains prevalent in shaping the Myc response, but indirect and less intense post-transcriptional regulations can also be expected (see, for example, McManus et al., 2015). We conclude that INSPEcT fails to capture the more complex regulatory scenarios, which are instead inferred by means of our novel Bayesian method.

4.3. Myc response analysis

In this section, we analyze the results inferred by means of our model. Each gene and KR is associated with a cluster, and we comment on all the sets of genes returned by our procedure for the processing and degradation rates. We instead focus on the clusters with at least 50 elements for the synthesis rate. This constraint is introduced to filter small gene sets (less than 200 genes in total) with modest responses and/or heterogeneity, for

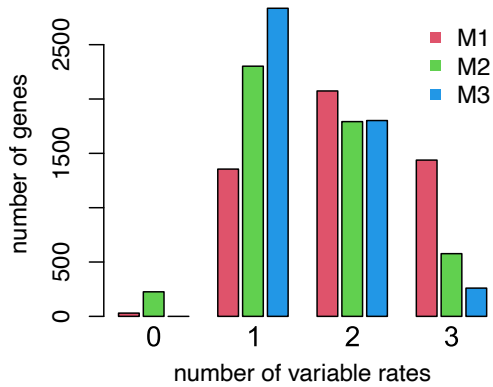


Fig. 4. Number of genes (y-axis) with a given number of non-constant rates (x axis) for M1, M2 and M3.

the sake of clarity and space, even though the number of elements is not necessarily a proxy of the biological relevance of a cluster. The missing clusters are depicted in Figure 3-SM and Figure 4-SM. In the first column of Figures 5 and 6, we show heatmaps of the temporal behavior of the KR log-fold-changes that belong to each cluster, computed using the posterior mean values $\hat{\mu}_{j,g}$, $\hat{\sigma}^2_{j,g}$, $\hat{\eta}_{1,j,g}$, $\hat{\eta}_{2,j,g}$ of the KR parameters. In the second column, for each time-point, we show the boxplots that represent the distributions of the KR log-fold-changes computed in the first column. The third and fourth columns present the $\hat{\mu}$ versus $\hat{\sigma}^2$ and $\hat{\eta}_1$ versus $\hat{\eta}_2$ smoothed color density representations of the scatterplots. Both of these graphs provide valuable information about the shape of the responses in the cluster of interest. Figure 6 contains analogues graphs for the processing and the degradation rates.

In order to support the discussion, we perform enrichment analyses, based on GO annotations (Dessimoz and Škunca, 2017; The Gene Ontology Consortium, 2019), for each rate. The graphical results of these analyses are shown in the supplemental material, with the exception of Figure 7, which here is reported as an example of a full output and refers to a specific case which has been selected both because of its interest and its graphical clarity. The GO enrichment analysis exploits a set of terms which, in our case, are pertinent to biological processes (e.g. "regulation of mRNA stability" or "cellular response to stress"), associated by means of relations (the edges in Figure 7, such as "is a" or "regulates"). Each term is also matched with a set of relevant genes by means of a curated annotation, which is constantly updated according to the literature, in order to reflect the knowledge of the scientific community on the biological domain. Given a set of genes, it is possible to search for those terms that are over-represented (enriched) in the group of interest, compared to a background (a number of associated genes that is proportional to the node size in Figure 7), which is a larger set that potentially accounts for all the annotated transcriptional units. A hypergeometric test is usually performed for any possible term to statistically test the enrichments, and a threshold is then imposed on the corrected p-values or q-values, selecting the most significant results. These terms (labels in Figure 7) provide a broad overview of the biological processes involved with the selected genes. We perform these analyses using the Bioconductor R-package *clusterProfiler* (Yu

et al., 2012).

Our analyses, detailed in the next sections, return a picture of complex gene expression programs involving all three KRs. The clusters obtained for the rate of synthesis show a plethora of diverse up- and down- regulations, which differ according to their magnitude, timing and shape. Several clusters are supported by a significant enrichment, associated with basic biological processes as well as interesting terms in the context of cancer biology and RNA metabolism. As expected, the clustering derived from the temporal response of the synthesis rate only partially recapitulates the steady-state counterpart which, however, reveals that the sets of genes involved in the regulation of the RNA metabolism are also among the most rapidly transcribed. As expected, the signals emerging from our analyses are less defined for the post-transcriptional rates, since the regulation of RNA synthesis is predominant in the Myc induction context. Nevertheless, our analyses provide interesting insights into the biology of this system and show that large sets of genes are potentially involve indirectly in the cellular response.

Synthesis rate. The graphical description of the clusters associated with the synthesis rate is shown in Figure 5. *Cluster 1* accounts for 1293 genes characterized by peak-type functions ($\hat{\eta}_1$ and $\hat{\eta}_2 < 0$), with a minimum between 1 and 3 hours and a small variance, which indicates quick responses. This behavior can be explained as a side effect of Myc activation, which causes the polarization of the resources required for the proficient transcription of the induced target genes (discussed in the following clusters). This response is globally compensated for in the last fraction of the time-course, thus indicating that these transcriptional units are required by the cell for its homoeostasis; this is confirmed by the enrichment analysis and pointed out in the literature (Dang, 2012; Chen et al., 2018) (see Figure 7-SM). *Cluster 2* accounts for 1103 genes, characterized by a monotonic decrease of the synthesis rate ($\hat{\eta}_1 < 0$ and $\hat{\eta}_2 > 0$). The temporal response is more heterogeneous, with $\hat{\mu}$ and $\hat{\sigma}^2$ spanning large domains. This cluster contains genes that are involved in cell growth and development, cell adhesion, migration, and apoptotic signaling regulation (Figure 8-SM); these are interesting clues that point to the role played by Myc in cancer biology (Dang, 2012; Chen et al., 2018). *Cluster 3*, which accounts for 904 genes, is composed of monotonic increasing responses ($\hat{\eta}_1 > 0$ and $\hat{\eta}_2 < 0$) concentrated in the first 5 hours of the time-course. This behavior indicates a potential direct Myc regulation. These genes are involved in coding and non-coding RNA metabolisms (Figure 7). They affect *RNA localization*, e.g. exporting from the nucleus, *RNA splicing* and *non-coding RNA processing*, and also RNA stability, e.g. the *RNA catabolic process*. Moreover, these genes are related to *RNA modification*, an emerging dynamic regulatory layer of the transcript metabolism (Roundtree et al., 2017). For example, $N^6 - methyladenosine$ is a methylated nucleotide (*methylation*, *RNA methylation*, *macromolecule methylation*) that is pervasive in the transcriptome of various species, e.g. mice, with a well established role in the control of transcript stability (Wang et al., 2014) and translation (Wang et al., 2015). Increasing evidence also links this RNA modification to synthesis and splicing (see, for example, Louloupi et al., 2018; Furlan et al., 2019). The analysis of these terms provides a coherent picture that relates Myc activation to several regulatory layers of the RNA metabolism and translation (e.g., the *cellular amino acid metabolic process*, *regulation of translation*). This evidence supports the post-transcriptional rate modulations predicted by our approach and, partially, by INSPEcT. *Cluster 4* accounts for 392 monotonic +

and - ($\hat{\eta}_1 \approx 0$ and $\hat{\eta}_2 < 0$ or $\hat{\eta}_1 < 0$ and $\hat{\eta}_2 \approx 0$, respectively) late responding genes ($\hat{\mu}$ generally larger than 4 and up to 14). These transcriptional regulations are probably secondary responses and are clustered together due to their temporal features. *Clusters 5, 6 and 8* are composed of 299, 246 and 148 monotonically induced genes, respectively, in response to Myc activation. These clusters were split due to their temporal responses which are slower and more heterogeneous in *Cluster 6*, and less sharp in *Cluster 8* (up-c shape: $\hat{\eta}_1 < 0$ and $\hat{\eta}_2 \approx 0$). *Cluster 6* is partially related to non-coding RNA processing (see Figure 10-SM). *Cluster 7* is composed of 182 genes characterized by weak peak-like induction ($\hat{\eta}_1$ and $\hat{\eta}_2 > 0$), which occurs earlier in the time-course and quicker than those belonging to Cluster 1, as can be seen from the values of $\hat{\mu}$ and $\hat{\sigma}^2$. Finally, *Cluster 9* is composed of 142 genes characterized by a monotonic decrease of the synthesis rate ($\hat{\eta}_1 \approx 0$ and $\hat{\eta}_2 > 0$). The response has a similar timing compared to *Cluster 2*, but is sharper and stronger; noticeably, these genes are involved in epithelial cell proliferation and migration, in agreement with the terms related to the second cluster. Our model also provides a clustering of genes according to their steady-state values of the synthesis rate (β_1), specifically 6 clusters enumerated with a progression of letters. *Cluster A* is related to cell-cycle regulation, cell growth, apoptosis and RNA processing. The latter process is also enriched in *Clusters C and D* (Figures 20-SM, 22-SM and 23-SM), while, the genes in *Cluster F* are involved in epithelium cell migration and translation (Figure 24-SM). Interestingly, a remarkable percentage ($\approx 30\%$) of the genes in *Clusters C and D* is also part of *Cluster 3* (Figure 35-SM). Moreover, these are clusters that are characterized by fast kinetics, which is a required condition, even though not sufficient, to quickly regulate the expression level of a gene, and is a clue of the fundamental regulatory role played by these transcriptional units (Figures 36-SM).

Processing rate. A graphical description of the clusters associated with the processing rate is given in Figure 6 (a). *Cluster 1* is composed of 3537 elements, which respond, to a great extent, within 3 hours from the stimulus with a moderate and sharp monotonic increase of k_2. Because of the timing of the response and the extension of this cluster, this behavior could be due to the general feedback mechanisms which link RNA synthesis and processing. On the other hand, *Cluster 2* is composed of 1212 genes characterized by small and late, positive and negative monotonic responses. These are probably secondary responses that are mediated by the remarkable number of genes involved in the RNA processing regulation perturbed by Myc activation. The weakness of these modulations is puzzling but, since they take place in the most coarse-grained part of the time-course, our approach may only detect a reflex of the real biological regulation that could occur between the experimental observations. Finally, *Cluster 3* is composed of 148 genes characterized by a fast and transitory induction of the processing rate taking place in the first 2 hours of the time-course, and resulting in a transitory increase in the amount of mature RNAs. The Gene Ontology enrichment analysis of *Cluster 2* and *Cluster 3* do not provide any relevant terms, while several biological processes, already mentioned while discussing the synthesis rate response, can be found for *Cluster 1*. However, the enrichments are less significant and they disappear when the 4897 differentially expressed genes are used as the background instead of all the annotated ones (Figure 12-SM and Figure 13-SM). The same is true for all of the 5 clusters which our method returns for the processing of the steady-state rate (Figure 25-SM, Figure 26-SM and Figure 27-SM).

Degradation rate. The modulations of the degradation rate are divided into four clusters, see Figure 6 (b). *Cluster 1* is composed of 3094 elements characterized by very quick peak-type responses and with a minimum between 1 and 2 hours. As we have seen for k_2, this behavior may be due to the coupling with the synthesis rate. *Cluster 2* and *Cluster 3* account for 1279 and 438 genes, respectively, characterized by late, positive and negative, monotonic responses, which are probably secondary. Like the processing rate, this could point to indirect regulations that were under-sampled due to the temporal design of the experiment. *Cluster 4* is composed of 86 genes characterized by a quick peak-like induction. A similar increase in the rate of degradation would result in a more rapid turnover of mature RNA molecules, while a similar behavior could be compatible with either deleterious transcripts for the cellular response, or those whose fast modulation is needed in the Myc induction context. The Gene Ontology enrichment analysis results are analogues with those discussed for the processing rate (Figure 14-SM and Figure 15-SM) for the 5 steady-state clusters (Figure 28-SM, Figure 29-SM and Figure 34-SM).

5. Final remarks

Motivated by a real data application, we here proposed a Bayesian approach to the analysis of RNA expression levels. In our proposal, the experimental data are hypothesized to be noisy observations of a true process, which is a solution to a system of ODEs. We assume that the ODEs depend on KRs, that are time- and gene-dependent. The KRs are the main object of inference, since they characterize the RNA life cycle and provide important insights into the analysis of gene expression levels. The temporal evolution of KRs is encoded with a new single family of functions, defined by only 5 parameters that can easily be interpreted from the biological perspective (i.e. initial value, relative log-fold-changes, and temporal location and duration of the response). The parameters are divided into two groups, according to their role in defining either the initial value of the KR or its temporal modulation. A mixture model, based on the DP, is defined for both of them, and for each KR. This allows us to find sets of genes with similar KR shapes or steady-state values to guide the inference. This approach is conceptually based on the well-established co-regulation of genes, which a cell often exploits to coordinate the expression level of the multiple transcripts required to operate a specific task. Therefore, the idea of including a clustering step in the inference process is not only biologically robust, but also provides a valuable information.

The results obtained with the new method are biologically relevant. The enrichment analysis of the clusters results in sets of terms that are meaningful in the Myc biology context, and which are in conceptual agreement with the shape of the responses. This is particularly true for the synthesis rate, which is the most informative regulatory layer in this system. Moreover, our method manages to identify a remarkable fraction of genes as post-transcriptionally regulated, thus pointing to weak indirect secondary responses.

A limitation of our method is the independence of the synthesis, processing and degradation rate clusters. In principle, this could be overcome by defining a mixture model on the parameters that shape the response of multiple rates. However, this inference would take place over a much larger space, which would be difficult to handle.

Acknowledgements

The work of the first two authors has partially been developed under the MIUR grant Dipartimenti di Eccellenza 2018 - 2022 (E11G18000350001), conferred to the Dipartimento di Scienze Matematiche - DISMA, Politecnico di Torino. The computational resources were provided by the Dipartimento di Scienze Matematiche - DISMA, Politecnico di Torino. The authors would like to thank Mattia Pelizzola (CGS@SEMM - IIT), the Editor-in-Chief, the Associate Editor and the two anonymous reviewers for their comments that have greatly improved the manuscript.

References

Ahnert, K. and Mulansky, M. (2011). Odeint - solving ordinary differential equations in c++. AIP Conference Proceedings, **1389**(1), 1586–1589.

Allocco, D. J., Kohane, I. S., and Butte, A. J. (2004). Quantifying the relationship between co-expression, co-regulation and gene function. BMC Bioinformatics, **5**(1), 18.

Anderson, D. F. and Kurtz, T. G. (2015). Stochastic analysis of biochemical systems, volume 1 of Mathematical Biosciences Institute Lecture Series. Stochastics in Biological Systems. Springer, Cham; MBI Mathematical Biosciences Institute, Ohio State University, Columbus, OH.

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive mcmc. Statistics and Computing, **18**(4), 343–373.

Chechik, G. and Koller, D. (2009). Timing of gene expression responses to environmental changes. Journal of Computational Biology, **16**(2), 279–290.

Chen, H., Liu, H., and Qing, G. (2018). Targeting oncogenic myc as a strategy for cancer treatment. Signal Transduction and Targeted Therapy, **3**(1), 5.

Dang, C. V. (2012). Myc on the path to cancer. Cell, **149**(1), 22–35.

de Pretis, S., Kress, T., Morelli, M. J., Melloni, G. E. M., Riva, L., Amati, B., and Pelizzola, M. (2015). INSPEcT: a computational tool to infer mRNA synthesis, processing and degradation dynamics from RNA- and 4sU-seq time course experiments. Bioinformatics, **31**(17), 2829–2835.

de Pretis, S., Kress, T. R., Morelli, M. J., Sabó, A., Locarno, C., Verrecchia, A., Doni, M., Campaner, S., Amati, B., and Pelizzola, M. (2017). Integrative analysis of rna polymerase ii and transcriptional dynamics upon myc activation. Genome Research.

Dessimoz, C. and Škunca, N., editors (2017). The gene ontology handbook. Number volume 1446 in Methods in molecular biology. Humana Press ; Springer Open, New York. OCLC: ocn959227666.

Dölken, L., Ruzsics, Z., Rädle, B., Friedel, C. C., Zimmer, R., Mages, J., Hoffmann, R., Dickinson, P., Forster, T., Ghazal, P., and Koszinowski, U. H. (2008). High-resolution gene expression profiling for simultaneous kinetic parameter analysis of rna synthesis and decay. RNA, **14**(9), 1959–1972.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research, **30**(1), 207–210.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association, **90**(430), 577–588.

Furlan, M., Galeota, E., De Pretis, S., Caselle, M., and Pelizzola, M. (2019). m6A-Dependent RNA Dynamics in T Cell Differentiation. Genes, **10**(1), 28.

Furlan, M., Galeota, E., Gaudio, N. D., Dassi, E., Caselle, M., de Pretis, S., and Pelizzola, M. (2020). Genome-wide dynamics of RNA synthesis, processing, and degradation without RNA metabolic labeling. Genome Research.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2013). Bayesian Data Analysis. Chapman and Hall/CRC, third edition edition.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. Statistics and Computing, **24**(6), 997–1016.

Gnedin, A., Gnedin, E., and Kerov, S. (2001). A characterization of gem distributions. Combin. Probab. Comp, **10**, 213–217.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, **102**(477), 359–378.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics, **17**(6), 333–351.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. The Annals of Statistics, **33**(2), 730 – 773.

Jain, S. and Neal, R. M. (2007). Splitting and merging components of a nonconjugate Dirichlet process mixturemodel. Bayesian Analysis, **2**(3), 445 – 472.

Li, B. and Dewey, C. N. (2011). Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. BMC Bioinformatics, **12**(1), 323.

Littlewood, T. D., Hancock, D. C., Danielian, P. S., Parker, M. G., and Evan, G. I. (1995). A modified oestrogen receptor ligand-binding domain as an improved switch for the regulation of heterologous proteins. Nucleic Acids Research, **23**(10), 1686–1690.

Louloupi, A., Ntini, E., Conrad, T., and Ørom, U. A. V. (2018). Transient N-6-Methyladenosine Transcriptome Sequencing Reveals a Regulatory Role of m6A in Splicing Efficiency. Cell Reports, **23**(12), 3429–3437.

Marchese, F. P., Raimondi, I., and Huarte, M. (2017). The multidimensional mechanisms of long noncoding rna function. Genome Biology, **18**(1), 206.

McManus, J., Cheng, Z., and Vogel, C. (2015). Next-generation analysis of gene expression regulation - comparing the roles of synthesis and degradation. Mol. BioSyst., **11**, 2680–2689.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics, **9**(2), 249–265.

OpenMP Architecture Review Board (2008). OpenMP application program interface version 3.0.

Papastamoulis, P. and Rattray, M. (2018). A bayesian model selection approach for identifying differentially expressed transcripts from rna sequencing data. Journal of the Royal Statistical Society: Series C (Applied Statistics), **67**(1), 3–23.

R Core Team (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

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

Rabani, M., Raychowdhury, R., Jovanovic, M., Rooney, M., Stumpo, D. J., Pauli, A., Hacohen, N., Schier, A. F., Blackshear, P. J., Friedman, N., Amit, I., and Regev, A. (2014). High-resolution sequencing and modeling identifies distinct dynamic rna regulatory strategies. Cell, **159**(7), 1698–1710.

Rossell, D., Stephan-Otto Attolini, C., Kroiss, M., and Stöcker, A. (2014). Quantifying alternative splicing from paired-end rna-sequencing data. Ann. Appl. Stat., **8**(1), 309–330.

Roundtree, I. A., Evans, M. E., Pan, T., and He, C. (2017). Dynamic RNA Modifications in Gene Expression Regulation. Cell, **169**(7), 1187–1200.

Slack, F. J. and Chinnaiyan, A. M. (2019). The role of non-coding rnas in oncology. Cell, **179**(5), 1033 – 1055.

Subramaniam, S. and Hsiao, G. (2012). Gene-expression measurement: variance-modeling considerations for robust data analysis. Nature Immunology, **13**(3), 199–203.

Tesi, A., Pretis, S., Furlan, M., Filipuzzi, M., Morelli, M. J., Andronache, A., Doni, M., Verrecchia, A., Pelizzola, M., Amati, B., and Sabò, A. (2019). An early Myc-dependent transcriptional program orchestrates cell growth during B-cell activation. EMBO reports, **20**(9).

The Gene Ontology Consortium (2019). The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Research, **47**(D1), D330–D338.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with rna-seq. Nature Biotechnology, **31**(1), 46–53.

Vandevenne, M., Delmarcelle, M., and Galleni, M. (2019). Rna regulatory networks as a control of stochasticity in biological systems. Frontiers in Genetics, **10**, 403.

Wade, S. and Ghahramani, Z. (2018). Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion). Bayesian Analysis, **13**(2), 559 – 626.

Wang, X., Lu, Z., Gomez, A., Hon, G. C., Yue, Y., Han, D., Fu, Y., Parisien, M., Dai, Q., Jia, G., Ren, B., Pan, T., and He, C. (2014). N6-methyladenosine-dependent regulation of messenger RNA stability. Nature, **505**(7481), 117–120.

Wang, X., Zhao, B. S., Roundtree, I. A., Lu, Z., Han, D., Ma, H., Weng, X., Chen, K., Shi, H., and He, C. (2015). N6-methyladenosine Modulates Messenger RNA Translation Efficiency. Cell, **161**(6), 1388–1399.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS: A Journal of Integrative Biology, **16**(5), 284–287.

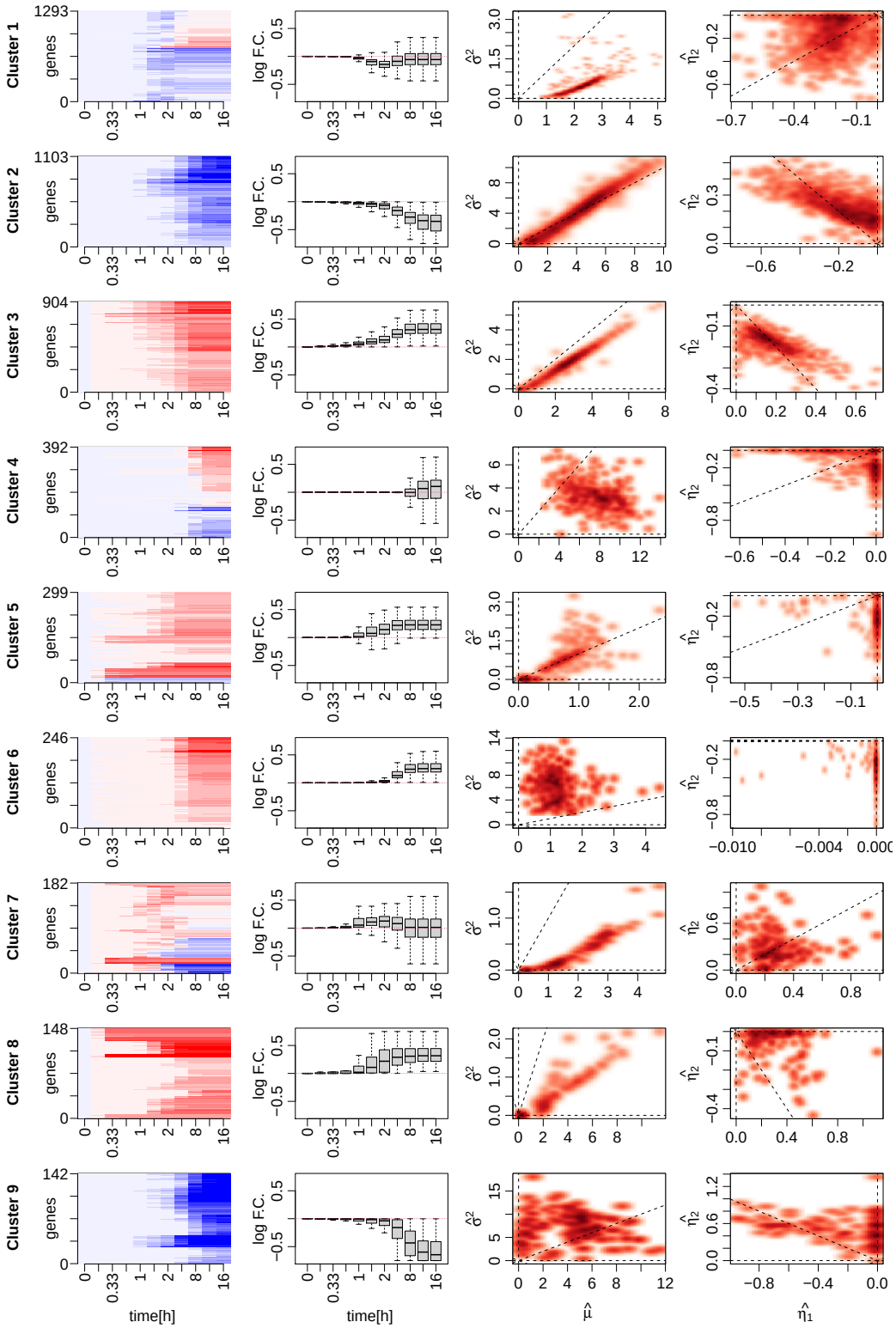
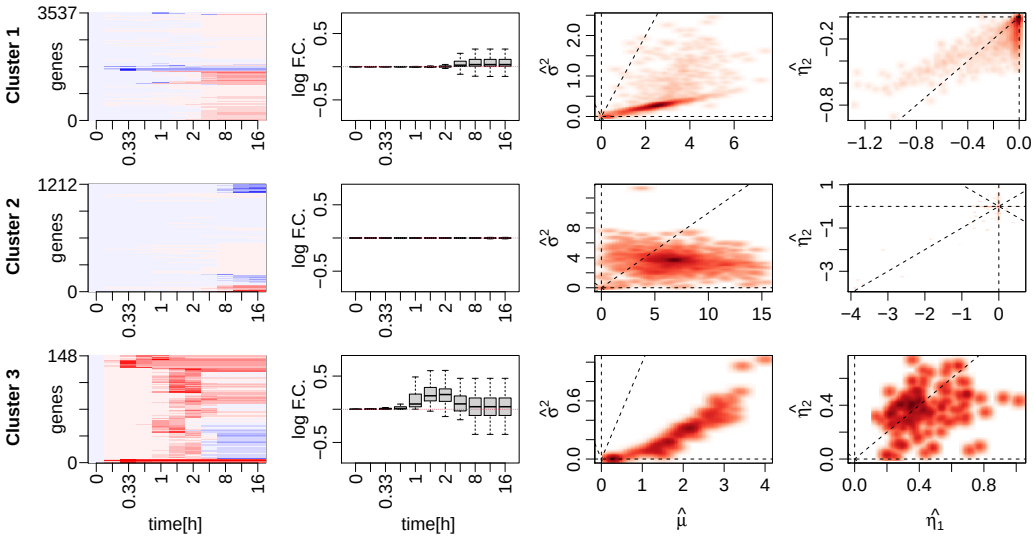
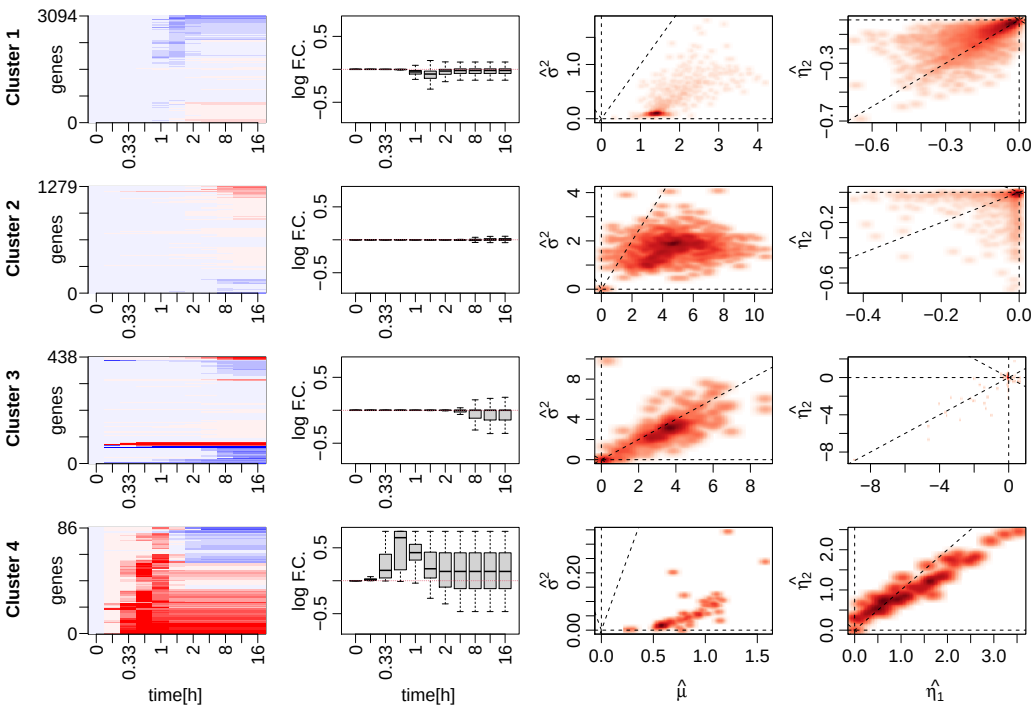


Fig. 5. Synthesis rate modulations in response to Myc activation for 7 clusters that are composed of at least 50 genes. (First column) Heatmaps showing the log-fold-changes of the rate $\hat{\eta}(t, 0)$, compared to the first time-point, for each gene in the cluster (values saturated at ± 0.75). (Second column) Boxplots showing the distribution of the synthesis rate log-fold-changes, compared to the first time-point (values saturated at ± 0.75). (Third column) $\hat{\mu}$ versus $\hat{\sigma}^2$ smooth-scatter plot. (Fourth column) $\hat{\eta}_1$ versus $\hat{\eta}_2$ smooth-scatter plot. The dashed lines represent the horizontal and vertical axes and the bisector of the first and third quadrants.



(a) Processing



(b) Degradation

Fig. 6. Processing (a) and Degradation (b) rate modulations in response to Myc activation. (First column) Heatmaps showing the log-fold-changes of the rate $\hat{\eta}(t, 0)$, compared to the first time-point, for each gene in the cluster (values saturated at ± 0.75). (Second column) Boxplots showing the distribution of the rate log-fold-changes, compared to the first time-point (values saturated at ± 0.75). (Third column) $\hat{\mu}$ versus $\hat{\sigma}^2$ smooth-scatter plot. (Fourth column) $\hat{\eta}_1$ versus $\hat{\eta}_2$ smooth-scatter plot. The dashed lines represent the horizontal and vertical axes and the bisector of the first and third quadrants.

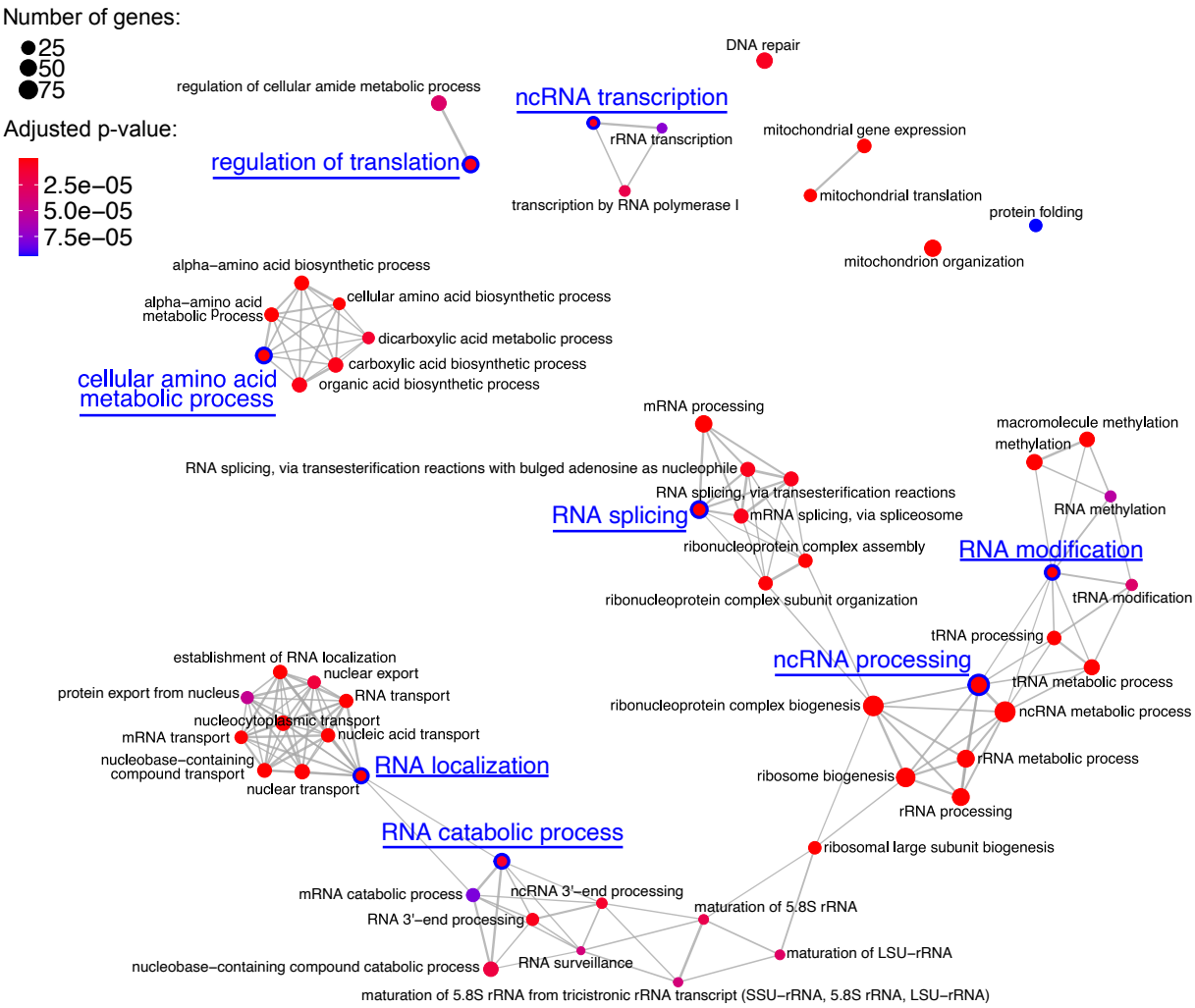


Fig. 7. Synthesis rate Cluster 3 Gene Ontology analysis. Biological processes (nodes of the network) associated with the genes belonging to the third cluster identified from the analysis of the synthesis rate modulations. The network structure is indicative of the semantic similarity of the terms; i.e., linked and adjacent terms are close to each other in the reference ontology. The size of each dot is proportional to the number of genes identified in the cluster, while the color is a proxy of the significance of the enrichment that takes into account the total number of genes associated with the specific term in the background. The nodes that are easier to interpret and which are characteristic of different communities of the network are highlighted.