

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

*Original*

Multiple latent clustering model for the inference of RNA life-cycle kinetic rates from sequencing data / Mastrantonio, Gianluca; Bibbona, Enrico; Furlan, Mattia. - In: THE ANNALS OF APPLIED STATISTICS. - ISSN 1932-6157. - 18:4(2024). [10.1214/24-aos1945]

*Availability:*

This version is available at: 11583/2994574 since: 2024-11-19T18:11:38Z

*Publisher:*

Institute of Mathematical Statistics

*Published*

DOI:10.1214/24-aos1945

*Terms of use:*

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

*Publisher copyright*

(Article begins on next page)

# MULTIPLE LATENT CLUSTERING MODEL FOR THE INFERENCE OF RNA LIFE-CYCLE KINETIC RATES FROM SEQUENCING DATA

BY GIANLUCA MASTRANTONIO<sup>1,a</sup> , ENRICO BIBBONA<sup>1,b</sup>  AND MATTIA FURLAN<sup>2,c</sup> 

<sup>1</sup>Department of Mathematical Sciences, Politecnico di Torino, <sup>a</sup>[gianluca.mastrantonio@polito.it](mailto:gianluca.mastrantonio@polito.it), <sup>b</sup>[enrico.bibbona@polito.it](mailto:enrico.bibbona@polito.it)

<sup>2</sup>Center for Genomic Science of IIT@SEMM, Fondazione Istituto Italiano di Tecnologia, Milan, Italy, <sup>c</sup>[mattia.furlan@iit.it](mailto:mattia.furlan@iit.it)

We propose a hierarchical Bayesian model to infer RNA synthesis, processing, and degradation rates from time-course RNA sequencing data, based on an ordinary differential equation system that models the RNA life cycle. We parametrize the latent kinetic rates, which rule the system, with a novel functional form and estimate their parameters through three 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 the proto-oncogene *Myc*. Our approach uncovers simultaneous regulations of the rates, which had been largely missed in previous analyses of this biological system.

**1. Introduction.** RNA is one of the most important actors in cell biology: it 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 noncoding RNAs in the regulation of gene expression levels (GELs) (e.g., modulation of transcripts stability, proteins synthesis, and proteins localization) (Marchese, Raimondi and Huarte (2017), Slack and Chinnaiyan (2019), Vandevenne, Delmarcelle and Galleni (2019)). A cell constantly regulates the expression levels of thousands of genes, that is, the number of the associated transcripts, in order to preserve its homeostasis and adapt to the environment. The state-of-the-art approach used to measure GELs is the “Next Generation RNA sequencing” (RNA-Seq) (Goodwin, McPherson and McCombie (2016)). Owing to the relevance of the topic, a remarkable number of public RNA-Seq datasets are currently available and easily accessible, for example, through the Gene Expression Omnibus project (Edgar, Domrachev and Lash (2002)), and a large amount of literature has been produced on the analysis of RNA-Seq data. In the literature, mixture models had been used on the observed data to estimate GELs (Huang and Sanguinetti (2016), Tuerk, Wiktorin and Güler (2017)) and also to identify genes differentially expressed under multiple experimental conditions (Love, Huber and Anders (2014), Sun et al. (2017), Papastamoulis and Rattray (2018), Tiberi and Robinson (2020)), which is one of the most common practices in the field.

Time course data represent a peculiar application of the paradigm of differential expression. They consist in the measurements of GELs of portions of a common cell culture measured at different times (Allocco, Kohane and Butte (2004)) to observe alterations occurring in response to a given perturbation. In the analysis of differential expression, or time-course data, the mere quantification of expression levels alone could lead to misleading conclusions. Indeed, a cell can regulate gene expression through different fundamental processes, like transcript synthesis and degradation, and while an increase in the expression level of a gene between two conditions is often interpreted as an intensification of its transcription, this could also result from enhanced stability of its transcripts.

---

Received January 2024.

*Key words and phrases.* Dirichlet process, kinetic rates, RNA, Gene expression.

In the literature, several models have been proposed to describe the RNA life-cycle, ranging from the accounting of only RNA synthesis and decay (Farina et al. (2008), Schwab et al. (2012), Uvarovskii and Dieterich (2017), Jürges, Dölken and Erhard (2018)) to more complex versions involving RNA export (Chen and van Steensel (2017)), association with polysomes (Li (2015), Fang et al. (2018)), or RNA processing (Zeisel et al. (2011), Rabani et al. (2011, 2014), de Pretis et al. (2015), Alkallas et al. (2017), La Manno et al. (2018), Bergen et al. (2020), Liu et al. (2023)). The latter is the most comprehensive one which does not involve the explicit modelling of RNA localization and, consequently, the requirement of specific experiments to isolate and sequence RNA molecules in specific cellular compartments (i.e., cells nucleus or polysomes). For this reason this model found a large application and is also used in the present work.

The RNA life-cycle model we adopted can be expressed as a system of linear ordinary differential equations (ODEs) whose (possibly time-dependent) coefficients, the so-called kinetic rates (KRs), can be interpreted as the instantaneous rates at which the mechanisms of synthesis, processing, and degradation occur. The recent literature has shown that KRs estimation can contribute greatly to a better understanding of the regulation mechanisms. Few works have also explored this direction in Bayesian inference frameworks (Rummel, Sakellaridi and Erhard (2023), Jürges, Dölken and Erhard (2018), Schofield et al. (2018)).

Motivated by the study of the activation of the proto-oncogene Myc in murine fibroblasts (de Pretis et al. (2017)), we propose a novel Bayesian approach to the inference of the KRs of the RNA life-cycle from time-course RNA-seq data, that includes a clusterization step for each KR, in order to gather the so-called co-regulated genes (Allocco, Kohane and Butte (2004), Farina et al. (2008)) that response to the stimulus with a similar modulation of one of the basic processes.

To highlight the novelties of our methodology, we first remark 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 KRs are expected to take on several 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 a mixture model for each KR that groups genes with similar responses to Myc activation into clusters. The approach is somehow similar to what has been earlier done in pharmacokinetics (PK) (Tatarinova et al. (2013), Muller and Rosner (1997), Walker and Wakefield (1998)), except that in PK the clustering is applied directly to the ODE coefficients, while in our case each KR is a different function of time that belongs to the same parametric family, and the clustering step is applied to the parameters of this family, separately for each KR, increasing the complexity of the hierarchical model (see Figure 1 and Figure 2).

The complexity is such that certain devices are needed to make the implementation reasonably simple and, at the same time, the computational cost acceptable. The three mixture models are indeed defined over latent quantities, not over the observed data. Moreover, the emission distribution of the mixture models should be defined on the subset of the parameter space where the KR parameters are identifiable. Our solution is to define a set of unconstrained working parameters, which are then suitably transformed into the natural parameters over the identifiable domain. Moreover, we need 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 zero value with positive probability.

Unlike other proposals, we estimate likelihood parameters, the variance of the measurement error, the scaling factor, KR functions, and clusterings in a single Bayesian model. Moreover, our approach is able to extract and exploit information shared by the elements of the same cluster, thus resulting in better estimates of the parameters.

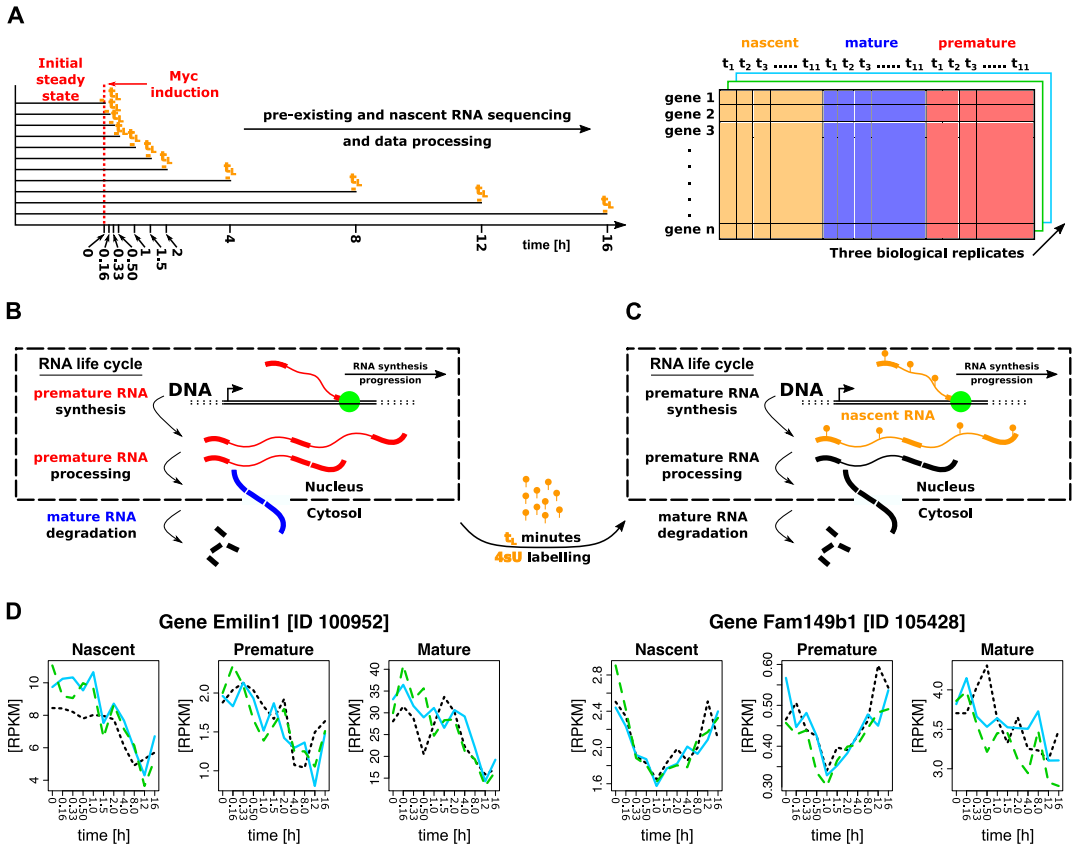


FIG. 1. (A) Experimental design used for the study of Myc activation in 3T9 cells. (B) RNA life-cycle in eukaryotic cells and definition of premature and mature RNA. (C) RNA metabolic labelling for nascent RNA quantification. (D) Gene expression profiles for two genes from the dataset; each replicate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

The capability of our method to capture the correct values of the parameters and their clusters is shown by a careful simulation study, both in the case when data are generated from the same model used for the inference and, in the case the model, is subject to mild misspecification. We demonstrate the inferential gain provided by our approach on real data, using cross-validation, and we show that our method is able to detect small but significant modulations of post-transcriptional rates (i.e., processing and degradation), which had been largely missed by previous analyses of the same dataset (de Pretis et al. (2017)).

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 3.2); 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 3.3). The real data application is described in Section 4. We conclude with a critical summary of our work and some perspectives in Section 5.

**2. Dataset overview.** Our dataset, taken from de Pretis et al. (2017), is organised as illustrated in Figure 1A. It provides GELs of premature, mature, and nascent RNA for more than 10,000 genes, at 11 time points, for three replicates of the experiment. These values are reported as reads per kilobase million (RPKM), which is a common unit used in the field, as an alternative to raw counts, to prevent biases due to heterogeneous gene lengths and/or

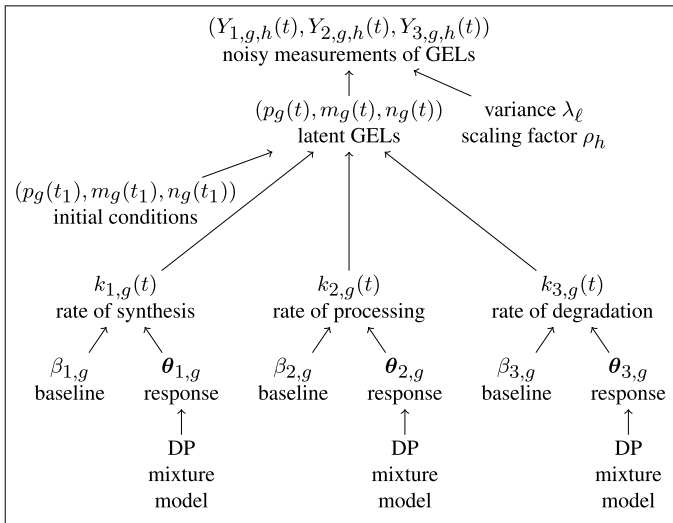


FIG. 2. Graphical representation of the model structure.

sequencing depths (i.e., the total number of reads collected for a sample); see [Conesa et al. \(2016\)](#). 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 tumours, and it is involved in the regulation of such basal cellular processes as differentiation, growth, and proliferation ([Dang \(2012\)](#), [Chen, Liu and Qing \(2018\)](#)).

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, GELs are profiled through RNA sequencing. Myc 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: 0 h,  $\frac{1}{6}$  h,  $\frac{2}{6}$  h,  $\frac{1}{2}$  h, 1 h,  $\frac{3}{2}$  h, 2 h, 4 h, 8 h, 12 h, 16 h (Figure 1A). Each experiment, performed on independent samples, was replicated three times and resulted in expression levels of premature and mature RNA (Figure 1B). The first quantity is estimated based on the number of reads exclusively overlapping the intronic regions of a gene (i.e., never annotated as exonic in any gene isoform), normalized by both the cumulative introns length and the library size (PRKM). Conversely, mature RNA is defined as the difference between the number of RPKM normalized reads overlapping the exonic regions of a gene (i.e., annotated as exonic in at least one isoform) and premature RNA.

The same experimental design was used to quantify nascent RNA through 4sU-Seq (Figure 1C). In this case, an exogenous nucleoside (4-thiouridine or 4sU) is provided to the cells before sequencing for a fixed span of time (labelling time). 4sU is incorporated in the transcripts produced during the entire labelling 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. However, it was not possible to analyse 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 three replicates and 11 time points. Figure 1D reports two examples from the dataset: the first one (gene *Emilin1*) represents a typical transcriptional regulation, as

can be seen from the adherence of the three profiles, while the second one (gene Fam149b1) may be associated with a more complex post-transcriptional scenario.

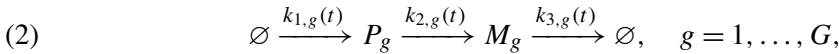
**3. Inference framework.** For each gene  $g \in \{1, \dots, G\}$ , time point  $t \in \{t_1, \dots, t_T\}$ , and replica  $h \in \{1, \dots, H\}$ ,  $Y_{\ell,g,h}(t)$  denotes the *measured* GEL of premature RNA if  $\ell = 1$ , mature RNA if  $\ell = 2$ , and nascent RNA if  $\ell = 3$ . Variables  $Y_{\ell,g,h}(t)$  are noisy and scaled versions of the true unobserved GELs, 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  $\lambda_{\ell,1}$  in the following way:

$$(1) \quad \begin{aligned} \log Y_{1,g,h}(t) &\sim N(\log p_g(t), \lambda_1(t)), \\ \log Y_{2,g,h}(t) &\sim N(\log m_g(t), \lambda_2(t)), \\ \log Y_{3,g,h}(t) &\sim N(\rho_h(t) \log n_g(t), \lambda_3(t)). \end{aligned}$$

The parameters  $\rho_h(t)$  are scaling factors that are required to normalize the nascent RNA libraries to the pre-existing RNA counterparts (Miller et al. (2011), Rabani et al. (2011), de Pretis et al. (2015)). In Section 3.1 the three latent GELs  $\{p_g(t), m_g(t), n_g(t)\}$  are modelled as the solution of a system of ordinary differential equations, that is, specified by three time-dependent coefficients  $k_{j,g}(t)$ , named KRs. These KRs can be interpreted as the rates at which the fundamental mechanisms underlying the RNA life cycle, namely, synthesis, processing, and degradation, occur. In Section 3.2 we introduce a parametric family of functions to which we assume that these KRs belong, and we show how the parameters can conveniently be subdivided into two sets,  $\beta_{j,g}$  and  $\theta_{j,g}$ . 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 that characterize the shape of the response to the stimulus. For the parameters  $\theta_{j,g}$ , in each rate  $k_{j,g}(t)$ , we define a mixture model that groups genes with similar temporal responses to Myc activation.

3.1. *A mathematical model of the RNA life-cycle.* According to the model we use to describe transcripts life-cycle, RNA molecules undergo three processes (Figure 1B). The first one is the synthesis of premature RNA from DNA. Premature transcripts are not suitable to perform their canonical task (e.g., protein translation) and require structural modifications (e.g. introns splicing). This second step of the RNA life cycle is called processing, and its product is mature RNA, which is eventually degraded by the cell in the final stage of the RNA life cycle. The process may be described by the following network of chemical reactions:



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 (2) into mathematical terms is

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

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 (Figure 1C). Nascent RNA is defined as the amount of total RNA, premature plus mature, which is produced by the cell in a short span of *labelling* time  $t_L$  (Figure 1). By definition, the nascent RNA is absent at the beginning of the labelling 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 ( $p_g^*(t)$ ) and mature ( $m_g^*(t)$ ) nascent RNA is, therefore, ruled by the following equations:

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

The sum  $n_g(t) = p_g^*(t) + m_g^*(t)$  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,

$$(4) \quad \dot{n}_g(t) = k_{1,g}(t).$$

Since the time window for which nascent RNA evolves is short ( $t_L$ ), this rate can be considered approximately constant, and equation (4) 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 (3) 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 (3) and are here considered as further model parameters.

**3.2. KR parametrization.** A routine experimental approach to investigate transcriptional programs consists of the perturbation of a cell culture followed by the repeated measurement of GELs to identify transcriptional units involved in the response. The KRs of modulated genes vary over time with typical shapes that share the following characteristics: steady at the beginning of the experiment, steady after a long time from the perturbation, and varying (with some regularity) in the transient region between the two steady-states. Some typical shapes are: (i) constant (some rates are not altered at all), (ii) monotonic (both increasing and decreasing), (iii) and peak-like functions. They have already been successfully applied to describe transcriptional and post-transcriptional responses in several biological systems (see, e.g., Chechik and Koller (2009), Rabani et al. (2011, 2014), de Pretis et al. (2014)). 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:

$$(5) \quad f(t|\mu, \sigma^2, \kappa_{-\infty}, \kappa_\mu, \kappa_{+\infty}) = \begin{cases} \kappa_{-\infty} + \frac{\phi(t|\mu, \sigma^2)}{\phi(\mu|\mu, \sigma^2)}(\kappa_\mu - \kappa_{-\infty}) & \text{if } t < \mu, \\ \kappa_{+\infty} + \frac{\phi(t|\mu, \sigma^2)}{\phi(\mu|\mu, \sigma^2)}(\kappa_\mu - \kappa_{+\infty}) & \text{if } t \geq \mu, \end{cases}$$

where  $\kappa_{-\infty}$ ,  $\kappa_\mu$  and  $\kappa_{+\infty}$  belong to  $\mathbb{R}^+$ . The function  $f(\cdot)$  in equation (5) 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  $\mu$ , taking care to preserve continuity at time point  $t = \mu$  (Figure 3). It is easy to see that

$$\kappa_{-\infty} = \lim_{t \rightarrow -\infty} f(t|\mu, \sigma^2, \kappa_{-\infty}, \kappa_\mu, \kappa_{+\infty}),$$



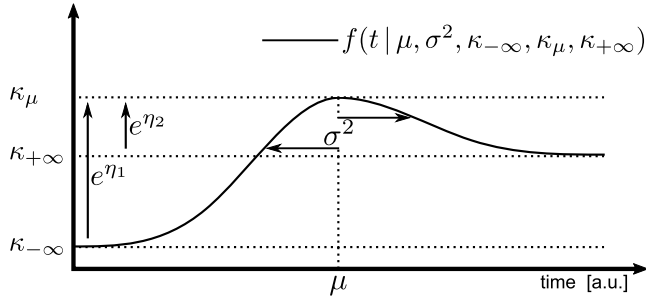


FIG. 3. Graphical representations of the KR parametrization. It should be noted that  $e^{\eta_1}$  and  $e^{\eta_2}$  are shown to indicate the section of the function they determine, but they are not equal to the length of the arrows; see equation (7).

$$(6) \quad \begin{aligned} \kappa_\mu &= f(\mu \mid \mu, \sigma^2, \kappa_{-\infty}, \kappa_\mu, \kappa_{+\infty}), \\ \kappa_{+\infty} &= \lim_{t \rightarrow +\infty} f(t \mid \mu, \sigma^2, \kappa_{-\infty}, \kappa_\mu, \kappa_{+\infty}). \end{aligned}$$

For easiness of interpretation, we split and rename the parameters as follows. First, we single-out  $\kappa_{-\infty}$ , and we rename it  $\beta$  to simplify the notation. Unlike the other parameters, which are related to the response,  $\beta$  is the baseline level, that is, the initial steady-state, and it is analysed separately. Second, we introduce the logarithmic ratios

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

These quantities are called *log-fold changes* in computational biology and are usually used to measure modulations with respect to the baseline level  $\beta$ , and we, therefore, define

$$(7) \quad \eta_1 = \eta(\mu, -\infty) = \log \frac{\kappa_\mu}{\kappa_{-\infty}}, \quad \eta_2 = \eta(\mu, +\infty) = \log \frac{\kappa_\mu}{\kappa_{+\infty}}.$$

Parameters  $\mu, \sigma^2, \eta_1, \eta_2$  are all related to the characterization of the response to perturbations. In particular,  $\mu$  and  $\sigma^2$  characterize the *temporal* location and duration of the response, while  $\eta_1$  and  $\eta_2$  determine the typical shape, as highlighted in Table 1. We collect these four parameters in a single vector that we denote  $\theta$  to obtain a more compact notation. Examples of the forms that can be obtained with (5), 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 (6) can now be reparametrized as  $f(t \mid \beta, \theta)$ , with  $\beta = \kappa_{-\infty} \in \mathbb{R}^+$  and  $\theta = (\mu, \sigma, \eta_1, \eta_2)$ .

Although the family of functions (6) is well defined for all real values of  $\mu, \eta_1$ , and  $\eta_2$ , and for all positive values of  $\sigma^2$ , certain identifiability and interpretability issues may arise if some conditions are not met. For example, if  $\mu$  is smaller than the first observed time  $t_1$ , and  $\sigma^2$  is small (compared to  $|\mu - t_1|$ ), function  $f$  in the interval  $[t_1, t_T]$  is indistinguishable for any arbitrary choice of  $\eta_1$  and  $\eta_2$  from a constant one, which should instead be given by  $\eta_1 = \eta_2 = 0$  (Table 1). For this reason, identifiability constraints are needed. The main requirement is that the value of the function (5) at time points  $t_1$  and  $t_T$  should be close in value to the steady state, that is,  $\beta$  and  $\kappa_\infty$ , respectively, which means that the most relevant part of the function graph lies within the observed time window. Hence, we require

$$|f(t_1 \mid \beta, \theta) - \beta| < 0.01|\kappa_\mu - \beta|, \quad \text{and} \quad |f(t_T \mid \beta, \theta) - \kappa_{+\infty}| < 0.01|\kappa_\mu - \kappa_{+\infty}|,$$

which implies two conditions,

$$(8) \quad \mu - \sigma\sqrt{-2\log(0.01)} > t_1, \quad \text{and} \quad \mu + \sigma\sqrt{-2\log(0.01)} < t_T.$$



TABLE 1  
*KR shapes as functions of the log-fold changes, and the names we use to describe these shapes*

$\eta_1$ sign	$\eta_2$ sign	Names		Shape
0	0	constant		—
+	+	peak-like	peak-like+	
-	-		peak-like-	
+	0	monotonic	monotonic+	
+	-		up-up	
0	-		c-up	
0	+		monotonic-	
-	+		down-down	
-	0		down-c	

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  $\approx 0.01$  in (8)) and then  $f(t_1 | \beta, \theta) \approx \beta$ , which lets us interpret  $\beta$  as the steady-state at time  $t_1$ , and  $f(t_T | \beta, \theta) \approx \kappa_\infty$ , which ensures a steady-state at the end of the observed window.

The subset of the parameter space, where  $\theta$  is identifiable and interpretable, is denoted as  $\mathcal{D} \subset \mathbb{R}^4$  and is defined by the conditions

$$(9) \quad \mu \in (t_1, t_T), \quad 0 < \sigma < \frac{\min(\mu - t_1, t_T - \mu)}{\sqrt{-2 \log(0.01)}}.$$

3.3. *Latent clustering models.* For each  $j \in \{1, 2, 3\}$ , we introduce a mixture model, based on the Dirichlet Process (DP), for the parameter  $\theta_{j,g}$ . We prefer this option with respect to making a single mixture model over the parameters  $\{\theta_{j,g}\}_{j \in \{1,2,3\}}$  since, for example, even though two genes may have a similar rate of synthesis, it is not necessarily true that the other rates are similar.

For each mixture model, the most natural choice would be to adopt an emission distribution with support in  $\mathcal{D}$ . However, this choice is impractical since it is hard to find an easy-to-handle distribution that fulfils the constraints in (9). Moreover, we are interested in the quantification of how likely is to have a constant  $k_{j,g}(t)$ , and we then need that the probability of  $\eta_{1,j,g} = 0$  is positive and so is that of  $\eta_{2,j,g} = 0$ .

To solve the problem with the domain of the emission distribution, we propose to reparametrize the model with the working parameters  $\beta_{j,g}^* \in \mathbb{R}$ , and  $\theta_{j,g}^* \in \mathbb{R}^4$ , which can be transformed into the natural parameters  $(\beta_{j,g}, \theta_{j,g}) \in \mathbb{R}^+ \times \mathcal{D}$  by

$$\begin{aligned} \beta_{j,g} &= \exp \beta_{j,g}^*, \\ \mu_{j,g} &= \frac{t_1 + t_2 \exp \mu_{j,g}^*}{1 + \exp \mu_{j,g}^*}, \\ \sigma_{j,g} &= \frac{\sigma_{\max,j,g} \exp \sigma_{j,g}^*}{1 + \exp \sigma_{j,g}^*}, \quad \sigma_{\max,j,g} = \frac{\min(\mu_{j,g} - t_1, t_T - \mu_{j,g})}{\sqrt{-2 \log(0.01)}}. \end{aligned}$$

Moreover, for the second issue, we have to ensure that a continuous distribution over  $(\beta_{j,g}^*, \theta_{j,g}^*)$  induces a mixed type probability distribution for  $\eta_{1,j,g}$  and  $\eta_{2,j,g}$ , with a continuous component over  $\mathbb{R}$  and a point mass at 0. Hence, we define

$$(10) \quad \eta_{i,j,g} = \begin{cases} \max(0, \eta_{i,j,g}^* - \xi) & \text{if } \eta_{i,j,g}^* > 0, \\ \min(0, \eta_{i,j,g}^* + \xi) & \text{otherwise,} \end{cases} \quad i = 1, 2.$$

It should be noted that, if we assume a continuous distribution for  $\eta_{1,j,g}^*$ , as a result of (10), the distribution over  $\eta_{1,j,g}$  is continuous over  $(-\infty, 0)$  and  $(0, \infty)$  and has a point mass at 0 equal to the cumulative distribution of  $\eta_{1,j,g}^*$  between  $-\xi$  and  $\xi$ . A similar result holds for  $\eta_{2,j,g}$ . Therefore, equation (10) 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, for example, if the mean is zero and the variance is small (large), the mass is close to 1 (negligible). As shown in equation (11), the means and variances of these distributions are parameters 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  $\theta_{j,g}^*$  and define three mixture models, based on Gaussian densities over  $\theta_{1,g}^*$ ,  $\theta_{2,g}^*$  and  $\theta_{3,g}^*$ ,

$$(11) \quad \begin{aligned} \theta_{j,g}^* | \zeta_{j,z_{j,g}}, \Omega_{j,z_{j,g}}, z_{j,g} &\sim N(\zeta_{j,z_{j,g}}, \Omega_{j,z_{j,g}}), \\ z_{j,g} | \pi_j &\sim \text{Discrete}(\pi_j), \\ \pi_j | \alpha_j &\sim \text{GEM}(\alpha_j), \\ \zeta_{j,k}, \Omega_{j,k} &\sim \text{NIW}(\mathbf{M}, \tau, \nu, \Psi), \end{aligned}$$

where  $k \in \mathbb{N}$  and NIW() stands for the normal inverse-Wishart distribution, where its first parameter is the mean of the normal distribution, the second is the scaling factor of the variance, while the third and fourth parameters are the ones of the inverse-Wisharts. Variable  $z_{j,g}$  is the discrete random variable that represents the label that identifies the component of the mixture to which the parameters belong. These variables are assumed to come from a discrete distribution, whose probabilities follow a DP defined by the GEM (or stick-breaking) distribution (Gnedin and Kerov (2001)). Given the allocation variables  $z_{j,g}$ , parameters  $\theta_{j,g}^*$  are normally distributed.

3.4. *Implementation details.* We choose weakly informative priors for all the model parameters. We set  $\mathbf{M} = \mathbf{0}$ ,  $\tau = 10^{-10}$ ,  $\nu = 6$ ,  $\Psi = \mathbf{I}$ , and  $\alpha_j \sim G(0.01, 0.01)$  for the DP components of our proposal (11). We choose a gamma distribution  $G(0.01, 0.01)$  for the variance parameter  $\lambda_\ell$  and a Normal  $N(0, 10e5)$  for the scaling factor  $\rho_h(t)$  (equation (1)). A  $G(0.01, 0.01)$  is also used as a prior for the initial conditions  $p_g(t_1)$  and  $m_g(t_1)$  (equation (3)) of the ODE system. We also assume  $\xi = 10$ . We use the sampling scheme proposed by Escobar and West (1995) for the DP hyperparameters  $\alpha_j$ . To simplify the implementation, we marginalize with respect to the state probabilities  $\pi_j$  and the likelihood parameters  $(\zeta_{j,k}, \Omega_{j,k})$ . The latter is possible due to the normal inverse-Wishart prior. The marginalization allows us the use of the algorithm number 3 of Neal (2000) to perform a global update of the latent allocation variables  $z_{j,g}$  and a single step of the “restricted Gibbs sampling split–merge” of Jain and Neal (2007) (with parameters (15, 1, 0)) to split and merge existing states. Parameters  $\{\theta_{1,g}^*, \beta_{1,g}, p_g(t_1)\}$ ,  $\{\theta_{2,g}^*, \beta_{2,g}, m_g(t_1)\}$ , and  $\{\theta_{3,g}^*, \beta_{3,g}\}$  are all sampled separately

using Metropolis steps with the adaptive proposals of [Andrieu and Thoms \(2008\)](#) (algorithm number 4), while  $\lambda_\ell$  and  $\rho_h(t)$  are sampled from their full conditionals that are normal distributions. The model is implemented in Julia ([Bezanson et al. \(2017\)](#)). We estimate the model on a computer cluster of 32 cores, with  $1e5$  iterations, burn-in  $6e4$ , and thin 20. We check convergence using split- $\hat{R}$  statistics ([Gelman et al. \(2013\)](#)) computed on parameters  $\theta_{j,g}^*$ ,  $\alpha_j$ ,  $\lambda_\ell$ ,  $n_g(t)$ ,  $m_g(t)$ ,  $p_g(t)$  and  $\rho_h(t)$ , with 1.05 as threshold. The computations took three days. In terms of output processing, using the algorithm of [Wade and Ghahramani \(2018\)](#), we find a representative point estimate of the cluster membership variables  $z_{j,g}$ , which we indicate with  $\hat{z}_{j,g}$ . From the model output, it is easy to determine the posterior distribution of the random variables  $W_{\theta,j}$ , which are the number of clusters for the  $j$ th mixture model. The  $b$ th posterior samples of  $W_{\theta,j}$  is defined as the number of unique values assumed by  $z_{j,g}$ , across all  $g = 1, \dots, G$  at iteration  $b$ . After model fitting, a posterior sample of the cluster-specific parameters ( $\zeta_{j,k}$ ,  $\Omega_{j,k}$ ) are obtained by simulation from their joint full-conditional distribution, and the algorithm of pivotal-reordering is used to deal with the label-switching problem ([Marin, Mengersen and Robert \(2005\)](#), [Papaspiliopoulos and Roberts \(2008\)](#)).

**3.5. Model validation over simulated datasets.** In this section we give a summary of the primary outcomes of our method when it is applied to three simulated datasets. More details and several illustrative figures are included in the Supplementary Material ([Mastrantonio, Bibbona and Furlan \(2024\)](#)). The first two datasets are simulated based on the correct model with distinct parameter values, and they mainly differ in the amount of variability within each cluster. On the other hand, the third dataset is created using a different model, and the data generation process utilizes different KR functions (wherein the parameters maintain the same interpretation). Moreover, it does not involve any clustering and employs a different likelihood for the observations. The purpose of this final numerical experiment is to assess whether our model can still accurately capture the values of the parameters despite a moderate degree of misspecification. The MCMCs are implemented using the same priors and iterations applied for the real data application.

The simulation study revealed that our model slightly overestimates the number of clusters predicting additional sets of genes with very small sizes. Indeed, most of the Maximum A Posteriori (MAP) cluster labels in the first two datasets were accurately assigned; this is irrelevant for the third dataset, which was generated without any clustering. Concerning the parameters of the three kinetic rates, the estimates for synthesis and processing exhibit high accuracy while the degradation rates counterparts result slightly more challenging. Nevertheless, the overall performance remains commendable. Finally, the analyses on the third dataset, where the interpretation of the parameters remains consistent despite a misspecified data-generating model, show the remarkable ability of our model to accurately capture the values of the parameters  $\eta_{i,j,g}$ ,  $\mu_{j,g}$ , and  $\beta_{j,g}$ . However, the efficiency of estimating the  $\sigma_{j,g}$  parameters is reduced.

**4. Real data application.** In this section we present the results of the application of our method to the Myc induction dataset. We explain how the model output can offer useful statistics to interpret the results, and we discuss the main biological insights provided by our approach.

**4.1. Output description and interpretation.** To assess the ability of our model to describe the data, we randomly selected 450 genes that were not used in the model fitting process. After model fitting we drew samples from their predictive posterior distributions using standard MCMC machinery.

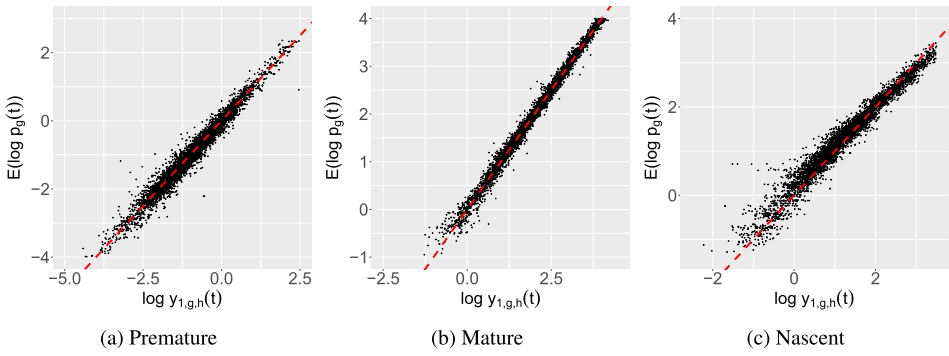


FIG. 4. Scatterplot that shows the observed  $\log(y_{\ell,g,h}(t))$  (x-axis) in the hold-out set and the posterior means (y-axis) of the associated posterior samples obtained in a model where  $\log(y_{\ell,g,h}(t))$  was not used in the model fitting.

Figure 4 shows a set of scatterplots comparing the logarithm of the observed data of the 450 hold-out samples and the associated posterior means of  $\log Y_{\ell,j,h}(t)$  (y-axis). The figure indicates that our proposal effectively summarizes the experimental data. All subsequent analyses in this section are based on an MCMC algorithm implemented using the complete dataset.

Table 2 and Figures 5 to 8 facilitate the discussion of the results in the following paragraphs. Figure 5 depicts the posterior distribution of the variable  $W_{\theta,j}$  for each mixture model. By examining the unique values of  $\hat{z}_{j,g}$ , we have identified 14 clusters for  $\theta_{1,\cdot}^*$ , nine clusters for  $\theta_{2,\cdot}^*$ , and 11 clusters for  $\theta_{3,\cdot}^*$ . The numbers of associated genes for each cluster are (1252, 1064, 953, 546, 285, 186, 160, 147, 100, 92, 53, 31, 24, 4) for  $\theta_{1,\cdot}^*$ , (3824, 528, 187, 113, 107, 97, 21, 15, 5) for  $\theta_{2,\cdot}^*$ , and (3291, 1024, 303, 84, 47, 42, 41, 26, 20, 14, 5) for  $\theta_{3,\cdot}^*$ ; clusters are ordered in decreasing value of associated genes.

To facilitate the discussion of the results, we grouped the possible shapes of the KRs into three categories: constant, peak+ or monotonic+, and peak− or monotonic−. Then we assigned to each KR a characteristic shape, which is the shape that has the largest posterior probability. This was easily done since the possible shapes are defined by the signs of  $\eta_{1,j,g}$  and  $\eta_{2,j,g}$ , and whether they are zeros. The proportions of the characteristic shapes for each KR are shown in Table 2. We have also aimed to describe the KRs associated with each cluster in the form of log-fold changes. To achieve this, we sampled from the predictive distributions of new sets of parameters ( $\theta_{1,\cdot}^*$ ,  $\theta_{2,\cdot}^*$ ,  $\theta_{3,\cdot}^*$ ) for a given cluster value  $w$ . Each sample was then used to compute the log-fold changes at each observed time point, and the posterior means along with 95% credible intervals (CI) are depicted in Figures 6, 7, and 8. To facilitate the discussion, we decided to present this information only for clusters with at least 50 associated genes, based on the unique value of  $\hat{z}_{j,g}$ .

TABLE 2

For each function  $k_{j,\cdot}(\cdot)$ , the table shows the fraction of “constant” functions, decreasing monotonic, or peak-like functions (“monotonic− or peak−”), and increasing monotonic or peak-like functions (“monotonic+ or peak+”) estimated by the model

	Constant	Peak+ or monotonic+	Peak− or monotonic−
$k_{1,\cdot}(\cdot)$	0.320	0.297	0.382
$k_{2,\cdot}(\cdot)$	0.88	0.0161	0.100
$k_{3,\cdot}(\cdot)$	0.605	0.109	0.285

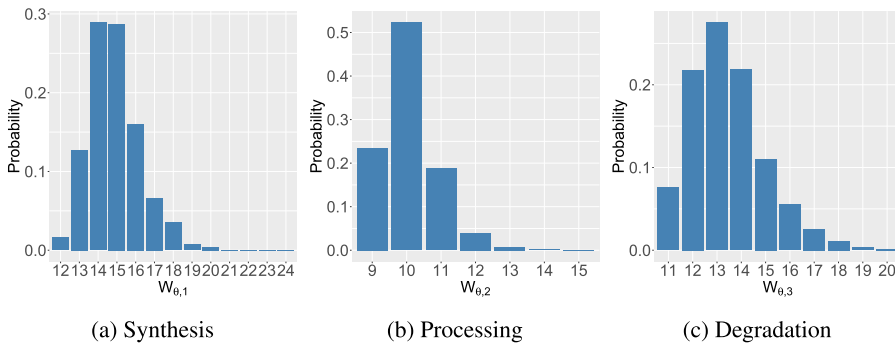


FIG. 5. Posterior distribution of the number of clusters for each of the three mixture models.

*Transcriptional responses.* Our results reveal a widespread modulation of the rate of synthesis in response to Myc activation involving 68% of the genes (Table 2). This regulation is characterized by a remarkable complexity, with both transcriptionally activated and repressed genes (30% and 38% of the regulated genes, respectively—Table 2) variable in terms of response timing (clusters averages between 10 minutes and eight hours—Figure 6) and shape (both monotonic and peak-like functions involved—Figure 6). This heterogeneity results in a set of four clusters accounting for more than 500 elements each.

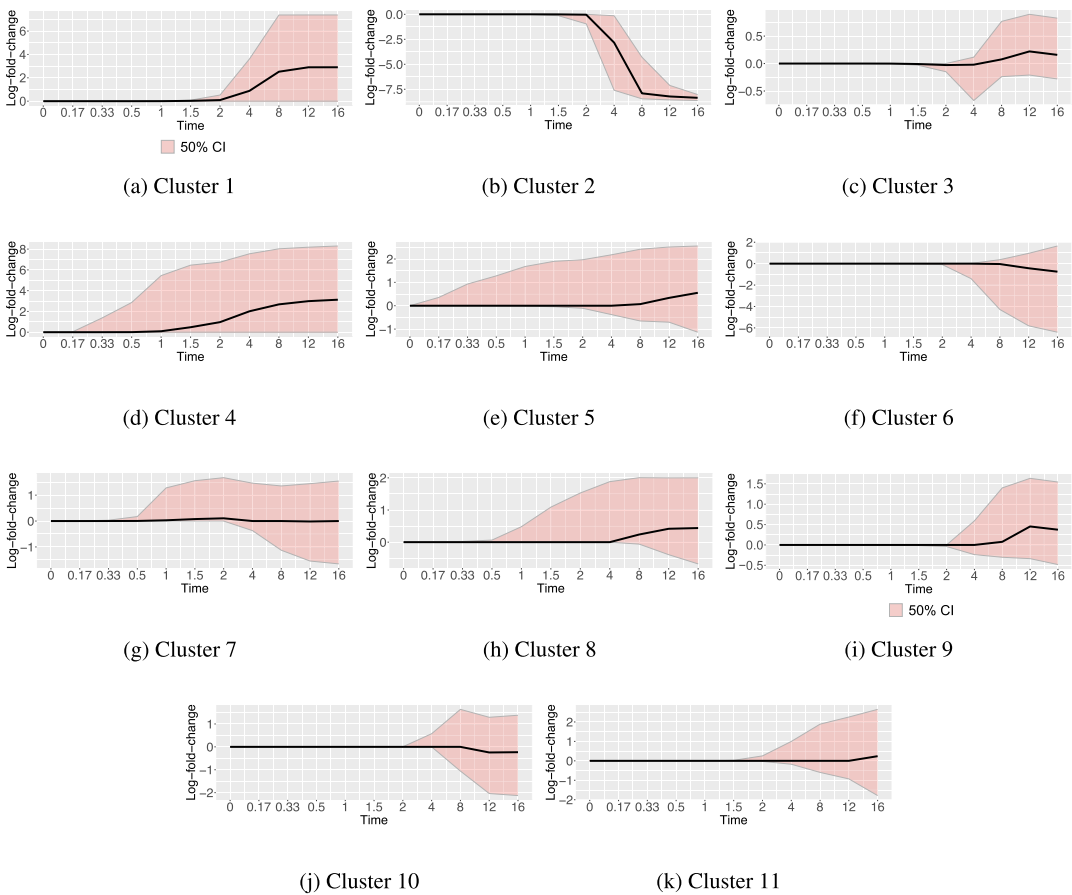


FIG. 6. Synthesis-plot of the posterior distribution of the mean log-fold changes, that is, log-fold changes computed with  $\zeta_{j,\cdot}$ . The shaded area represents the 95% CI, while the solid line is the median.

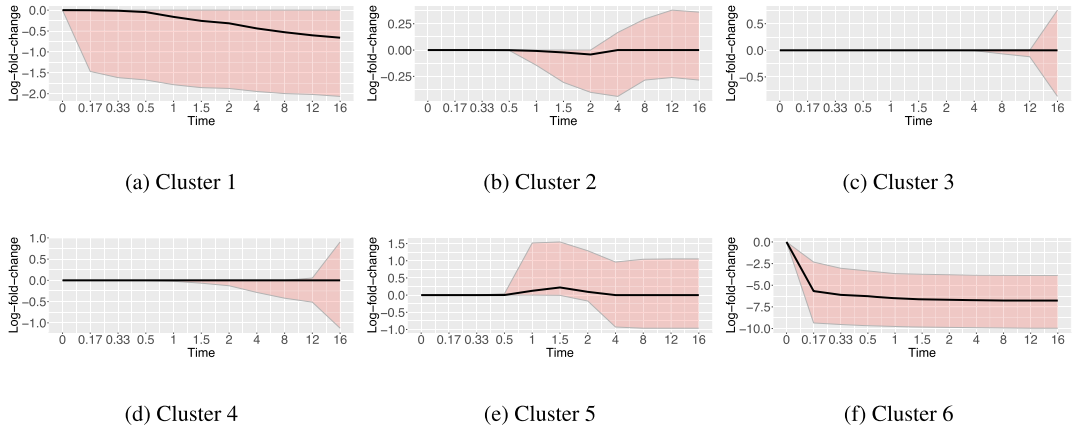


FIG. 7. Processing-plot of the posterior distribution of the mean log-fold changes, that is, log-fold changes computed with  $\xi_{j,\cdot}$ . The shaded area represents the 95% CI, while the solid line is the median.

In order to support the discussion, we performed enrichment analyses based on Gene Ontology (GO) annotations (Ashburner et al. (2000), Dessimoz and Škunca (2017), The Gene Ontology Consortium (2019)) for each set of genes. Briefly, Gene Ontology is a powerful framework in bioinformatics that provides a curated annotation of genes based on their involvement in various biological activities using a standardized and controlled vocabulary. The enrichment analysis is a statistical method, usually based on a hypergeometric test, which leverages Gene Ontology annotations to uncover terms significantly overrepresented within a given set of genes compared to a larger background. These analyses were performed using the Bioconductor R-package *clusterProfiler* (Yu et al. (2012)).

Cluster 1 is composed of 1252 induced genes which could be direct Myc targets responding to the activity of the transcription factor. Interestingly, they are enriched for transcriptional units involved in splicing regulation and ribosome biology (Figure S1). This observation is reasonable since Myc transcriptional activity requires adequate support from the splicing and translational machinery (Hsu et al. (2015), Ruggero (2009), Stine et al. (2015)). In contrast, cluster 2 accounts for 1064 down-regulated genes, which may be indirectly modulated by Myc, either through precise transcriptional feedback loops or, more in general, due to the concentration of cellular resources required for transcription toward Myc’s targets. Notably,

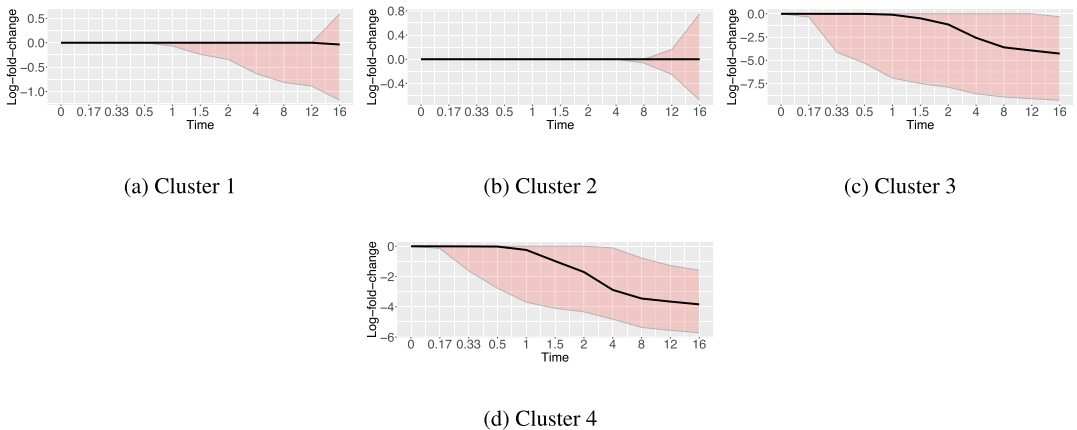


FIG. 8. Degradation-plot of the posterior distribution of the mean log-fold changes, that is, log-fold changes computed with  $\xi_{j,\cdot}$ . The shaded area represents the 95% CI, while the solid line is the median.

this set of genes is enriched in several cancer-related hallmarks, such as cell adhesion, migration, and differentiation (Figure S2). Cluster 3 is composed of 954 genes which display a mild slowdown of their synthesis around hours followed by a small increase in the rate. Again, this peak-like modulation could represent an indirect consequence of the transcriptional activation of Myc targets which negatively impacts the synthesis of other genes. Cluster 4, while not showing any specific enrichment, is intriguing since it comprises 546 early-induced genes, potentially further involved in direct Myc responses. Finally, all the other clusters are generally smaller (22% of the genes) and characterized by weaker and later responses, which probably represent indirect adaptations of the cell.

*Post-transcriptional responses.* The post-transcriptional rates of processing and degradation show weaker and simpler responses involving smaller fractions of the genes, 12% and 39.5%, respectively. For both these rates, we observe a global down-regulation which suggests a decrease of the splicing efficiency, in line with the aforementioned Myc-driven transcriptional stress, and a stabilization of the transcripts, which is reasonable to induce gene expression levels. Interestingly, the processing rate cluster 2 accounts for 528 genes initially downregulated and later restored in the second part of the time course. These transcriptional units are enriched in the same terms pointing to ribosome biology found in the first synthesis rate cluster, further suggesting the relevance of this process in Myc biology (Figure S4) (Ruggero (2009), Stine et al. (2015)). Noticeably, this diffused modulation of post-transcriptional rates in support of the primary transcriptional response is largely missed by INSPECT which models the expression profiles through a stronger regulation of the rate of synthesis (Figure S5).

*Clustering a posteriori.* We also applied, independently for each KR, a hierarchical clustering method with Ward-type distance (Murtagh and Legendre (2014)) to INSPECT's log-fold changes, and we selected the optimum cluster number through silhouette index optimization between 1 and 150 (Dudek (2020)). For the rate of synthesis the procedure resulted in the subdivision of the genes as down- and up-regulated with two macro-clusters of 3078 and 1819 elements, respectively (see Supplementary Material section *Clustering on the INSPECT KRs*). On the other hand, for both the post-transcriptional rates, the procedure grouped the large majority of the genes in a single cluster (71% and 80% for processing and degradation rates, respectively), while the remaining elements were spread across 149 clusters always smaller than 50 elements (median size of seven and five genes for processing and degradation rates, respectively). These results, which provide a poorer description of KRs modulations compared to the one obtained through our approach, suggest that clustering kinetic rates a posteriori represents a nontrivial task and further demonstrate the relevance of our inference framework, which natively incorporates this feature.

4.2. *Comparison of the multiple inference methods.* To validate the inference performance of our method (M1), we compared it against both INSPECT (M3) and a simplified Bayesian framework (M2) that assumes  $z_{j,g} = 1$  for all  $j$  and  $g$  (i.e., the mixture has a single component). We evaluated the approaches by means of the CRPS, which is the mean square error between the predicted and empirical cumulative distribution functions,

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

where  $F(\cdot)$  is the data distribution and  $\mathbb{1}$  is the Heaviside step-function. The CRPS is a widely used metric that can be applied to compare both Bayesian and frequentist models. To compute the CRPSs, we used the same dataset with the 450 hold-out samples of Section 4.1, and the posterior samples from the predictive distribution of the two Bayesian models are used to estimate the CRPS, while a closed form expression can be used for INSPECT, due



TABLE 3

The table shows the mean CRPS computed on the 450 hold-out data for the proposed model (M1), the proposed model without clustering (M2) and INSPEcT (M3)

	Premature	Mature	Nascent
M1	0.150	0.056	0.090
M2	0.151	0.070	0.120
M3	0.153	1.036	0.375

to the assumption on normality of the data in logarithm scale (Gneiting et al. (2005)). The CRPS results are presented in Table 3. The table shows that the CRPS for the premature RNA are quite similar, while the other two groups, especially the mature RNA, show a large difference between INSPEcT and the two Bayesian models. Moreover, for all three variables, our proposal is preferable to M2.

Noticeably, this increase in performance results also in a rise of the computational time required to fit the model. Indeed, our procedure requires three days to be completed, while, according to the data presented in the Supplementary Figure 1 of Furlan et al. (2020), INSPEcT could process the same dataset in less than nine hours. However, we believe this additional computational time represents an acceptable drawback, given the higher inference quality, especially because this type of analysis is meant to be performed once for each collected dataset.

**5. Final remarks.** Motivated by a real data application, we propose a Bayesian approach for the analysis of RNA expression levels. In our framework the experimental data are hypothesized to be noisy observations of a true process, which is the solution of a system of ODEs. We assume that the ODEs are a function of time- and gene-dependent KRs, which 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 five 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 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 valuable information.

The results obtained with the proposed 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 framework manages to identify a remarkable fraction of genes as post-transcriptionally regulated, thus pointing to weak indirect secondary responses.

Our approach represents a new asset for the community, which we foresee could have extensive application in the coming years for analysing both published (Rabani et al. (2011, 2014), Davari et al. (2017), Li et al. (2017), Michel et al. (2017), Wachutka et al. (2019), Tan et al. (2020), Choi et al. (2021), Bandiera et al. (2021)) and novel datasets. In the next few years, we expect a raise in the number of studies relying on the temporal profiling of nascent

and total RNA to investigate transcriptional programs. Indeed, this approach is currently the sole method to obtain a comprehensive genome-wide view of temporal modulations in RNA life-cycle kinetic rates, and its application has been limited mainly by sequencing expense (metabolic labelling can be performed with commercial kits and protocols with a reasonable effort). Fortunately, these costs have been steadily decreasing since 2001, suggesting that this issue will be eventually overcome allowing a larger application of this approach and, consequently, of our inference framework.

**Acknowledgments.** The authors would like to thank Mattia Pelizzola (IIT and UNIMIB) for his comments that have greatly improved the manuscript.

**Funding.** The work of GM has been partially carried out within the FAIR—Future Artificial Intelligence Research Foundation and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR)—MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3—D.D.1555 11/10/2022, PE00000013).

EB has received support under the National Plan for Complementary Investments to the NRRP, project “D34H—Digital Driven Diagnostics, prognostics and therapeutics for sustainable Health care” (project code: PNC0000001), Spoke 4, funded by the Italian Ministry of University and Research.

The work of MF has been supported by the Giorgio Boglio fellowship from the Italian Association for Cancer Research (AIRC-ID 26611).

## SUPPLEMENTARY MATERIAL

**Supplementary Material** (DOI: [10.1214/24-AOAS1945SUPP](https://doi.org/10.1214/24-AOAS1945SUPP); .pdf). The online Supplementary Material (Mastrantonio, Bibbona and Furlan (2024)), available on the web page of the journal, contains additional figures, some results of the clustering procedure applied to the INSPECT estimated KR, and three simulations results.

## REFERENCES

- ALKALLAS, R., FISH, L., GOODARZI, H. and NAJAFABADI, H. S. (2017). Inference of rna decay rate from transcriptional profiling highlights the regulatory programs of Alzheimer’s disease. *Nat. Commun.* **8**.
- ALLOCCO, D. J., KOHANE, I. S. and BUTTE, A. J. (2004). Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinform.* **5** 18.
- ANDRIEU, C. and THOMS, J. (2008). A tutorial on adaptive MCMC. *Stat. Comput.* **18** 343–373. [MR2461882 https://doi.org/10.1007/s11222-008-9110-y](https://doi.org/10.1007/s11222-008-9110-y)
- ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S. et al. (2000). Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25** 25–29.
- BANDIERA, R., WAGNER, R. E., BRITTO-BORGES, T., DIETERICH, C., DIETMANN, S., BORNELÖV, S. and FRYE, M. (2021). Rn7sk small nuclear rna controls bidirectional transcription of highly expressed gene pairs in skin. *Nat. Commun.* **12**.
- BERGEN, V., LANGE, M., PEIDL, S., WOLF, F. A. and THEIS, F. J. (2020). Generalizing rna velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38** 1408–1414.
- BEZANSON, J., EDELMAN, A., KARPINSKI, S. and SHAH, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Rev.* **59** 65–98. [MR3605826 https://doi.org/10.1137/141000671](https://doi.org/10.1137/141000671)
- CHECHIK, G. and KOLLER, D. (2009). Timing of gene expression responses to environmental changes. *J. Comput. Biol.* **16** 279–290.
- CHEN, H., LIU, H. and QING, G. (2018). Targeting oncogenic myc as a strategy for cancer treatment. *Signal Transduct. Targeted Ther.* **3** 5.
- CHEN, T. and VAN STEENSEL, B. (2017). Comprehensive analysis of nucleocytoplasmic dynamics of mrna in drosophila cells. *PLoS Genet.* **13** e1006929.
- CHOI, J., LYSAKOVSKAIA, K., STIK, G., DEMEL, C., SÖDING, J., TIAN, T. V., GRAF, T. and CRAMER, P. (2021). Evidence for additive and synergistic action of mammalian enhancers during cell fate determination. *eLife* **10**. <https://doi.org/10.7554/eLife.65381>

- CONESA, A., MADRIGAL, P., TARAZONA, S., GOMEZ-CABRERO, D., CERVERA, A., MCPHERSON, A., SZCZEŚNIAK, M., GAFFNEY, D. J., ELO, L. L. et al. (2016). A survey of best practices for rna-seq data analysis. *Genome Biol.* **17** 13.
- DANG, C. V. (2012). Myc on the path to cancer. *Cell* **149** 22–35.
- DAVARI, K., LICHTI, J., GALLUS, C., GREULICH, F., UHLENHAUT, N. H., HEINIG, M., FRIEDEL, C. C. and GLASMACHER, E. (2017). Rapid genome-wide recruitment of rna polymerase ii drives transcription, splicing, and translation events during t cell responses. *Cell Rep.* **19** 643–654.
- 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** 2829–2835.
- DE PRETIS, S., KRESS, T. R., MORELLI, M. J., SABO, A., LOCARNO, C., VERRECCHIA, A., DONI, M., CAMPANER, S., AMATI, B. et al. (2017). Integrative analysis of RNA polymerase II and transcriptional dynamics upon MYC activation. *Genome Res.* **27** 1658–1664. <https://doi.org/10.1101/gr.226035.117>
- DESSIMOZ, C. and ŠKUNCA, N., eds. (2017). *The Gene Ontology Handbook. Methods in Molecular Biology* **1446**. Humana Press, New York. OCLC: ocn959227666.
- DÖLKEN, L., RUZSICS, Z., RÄDLE, B., FRIEDEL, C. C., ZIMMER, R., MAGES, J., HOFFMANN, R., DICKINSON, P., FORSTER, T. et al. (2008). High-resolution gene expression profiling for simultaneous kinetic parameter analysis of rna synthesis and decay. *RNA* **14** 1959–1972.
- DUDEK, A. (2020). Silhouette index as clustering evaluation tool. In *Classification and Data Analysis* (K. Jajuga, J. Batóg and M. Walesiak, eds.) 19–33. Springer, Cham.
- EDGAR, R., DOMRACHEV, M. and LASH, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30** 207–210.
- ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. [MR1340510](https://doi.org/10.1080/01621459.1995.10477134)
- FANG, H., HUANG, Y.-F., RADHAKRISHNAN, A., SIEPEL, A., LYON, G. J. and SCHATZ, M. C. (2018). Scikit-ribo enables accurate estimation and robust modeling of translation dynamics at codon resolution. *Cell Syst.* **6** 180–191.e4.
- FARINA, L., DE SANTIS, A., SALVUCCI, S., MORELLI, G. and RUBERTI, I. (2008). Embedding mrna stability in correlation analysis of time-series gene expression data. *PLoS Comput. Biol.* **4** 1–12.
- 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 Res.* **30** 1492–1507. <https://doi.org/10.1101/gr.260984.120>
- GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2013). *Bayesian Data Analysis*, 3rd ed. Chapman and Hall/CRC, Boca Raton, FL.
- GNEDIN, A. and KEROV, S. (2001). A characterization of GEM distributions. *Combin. Probab. Comput.* **10** 213–217. [MR1841641 https://doi.org/10.1017/S0963548301004692](https://doi.org/10.1017/S0963548301004692)
- GNEITING, T., RAFTERY, A. E., WESTVELD, A. H. and GOLDMAN, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Mon. Weather Rev.* **133** 1098–1118.
- GOODWIN, S., MCPHERSON, J. D. and MCCOMBIE, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17** 333–351.
- HSU, T. Y.-T., SIMON, L. M., NEILL, N. J., MARCOTTE, R., SAYAD, A., BLAND, C. S., ECHEVERRIA, G. V., SUN, T., KURLEY, S. J. et al. (2015). The spliceosome is a therapeutic vulnerability in myc-driven cancer. *Nature* **525** 384–388.
- HUANG, Y. and SANGUINETTI, G. (2016). Statistical modeling of isoform splicing dynamics from RNA-seq time series data. *Bioinformatics* **32** 2965–2972.
- ISHWARAN, H. and RAO, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. [MR2163158 https://doi.org/10.1214/009053604000001147](https://doi.org/10.1214/009053604000001147)
- JAIN, S. and NEAL, R. M. (2007). Splitting and merging components of a nonconjugate Dirichlet process mixture model. *Bayesian Anal.* **2** 445–472. [MR2342168 https://doi.org/10.1214/07-BA219](https://doi.org/10.1214/07-BA219)
- JÜRGES, C., DÖLKEN, L. and ERHARD, F. (2018). Dissecting newly transcribed and old rna using grand-slam. *Bioinformatics* **34** i218–i226.
- LA MANNO, G., SOLDATOV, R., ZEISEL, A., BRAUN, E., HOCHGERNER, H., PETUKHOV, V., LIDSCHREIBER, K., KASTRITI, M. E., LÖNNERBERG, P. et al. (2018). Rna velocity of single cells. *Nature* **560** 494–498.
- LI, G.-W. (2015). How do bacteria tune translation efficiency? *Curr. Opin. Microbiol.* **24** 66–71. <https://doi.org/10.1016/j.mib.2015.01.001>
- LI, H.-B., TONG, J., ZHU, S., BATISTA, P. J., DUFFY, E. E., ZHAO, J., BAILIS, W., CAO, G., KROEHLING, L. et al. (2017). m6a mrna methylation controls t cell homeostasis by targeting the il-7/stat5/socs pathways. *Nature* **548** 338–342.

- 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 Res.* **23** 1686–1690.
- LIU, H., ARSIÈ, R., SCHWABE, D., SCHILLING, M., MINIA, I., ALLES, J., BOLTENGAGEN, A., KOCKS, C., FALCKE, M. et al. (2023). SLAM-drop-seq reveals mRNA kinetic rates throughout the cell cycle. *Mol. Syst. Biol.* **19** (10).
- LOVE, M. I., HUBER, W. and ANDERS, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with *deseq2*. *Genome Biol.* **15** 550.
- MARCHESE, F. P., RAIMONDI, I. and HUARTE, M. (2017). The multidimensional mechanisms of long noncoding rna function. *Genome Biol.* **18** 206.
- MARIN, J.-M., MENGERSEN, K. and ROBERT, C. P. (2005). Bayesian modelling and inference on mixtures of distributions. In *Bayesian Thinking: Modeling and Computation. Handbook of Statist.* **25** 459–507. Elsevier, Amsterdam. MR2490536 [https://doi.org/10.1016/S0169-7161\(05\)25016-2](https://doi.org/10.1016/S0169-7161(05)25016-2)
- MASTRANTONIO, G., BIBBONA, E. and FURLAN, M. (2024). Supplement to “Multiple latent clustering model for the inference of RNA life-cycle kinetic rates from sequencing data.” <https://doi.org/10.1214/24-AOAS1945SUPP>
- MICHEL, M., DEMEL, C., ZACHER, B., SCHWALB, B., KREBS, S., BLUM, H., GAGNEUR, J. and CRAMER, P. (2017). Tt-seq captures enhancer landscapes immediately after t-cell stimulation. *Mol. Syst. Biol.* **13**.
- MILLER, C., SCHWALB, B., MAIER, K., SCHULZ, D., DÜMCKE, S., ZACHER, B., MAYER, A., SYDOW, J., MARCINOWSKI, L. et al. (2011). Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Mol. Syst. Biol.* **7** 458.
- MULLER, P. and ROSNER, G. L. (1997). A Bayesian population model with hierarchical mixture priors applied to blood count data. *J. Amer. Statist. Assoc.* **92** 1279–1292.
- MURTAGH, F. and LEGENDRE, P. (2014). Ward’s hierarchical agglomerative clustering method: Which algorithms implement Ward’s criterion? *J. Classification* **31** 274–295. MR3277707 <https://doi.org/10.1007/s00357-014-9161-z>
- NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804 <https://doi.org/10.2307/1390653>
- PAPASPILIOPOULOS, O. and ROBERTS, G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika* **95** 169–186. MR2409721 <https://doi.org/10.1093/biomet/asm086>
- PAPASTAMOULIS, P. and RATTRAY, M. (2018). A Bayesian model selection approach for identifying differentially expressed transcripts from RNA sequencing data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **67** 3–23. MR3758753 <https://doi.org/10.1111/rssc.12213>
- RABANI, M., LEVIN, J. Z., FAN, L., ADICONIS, X., RAYCHOWDHURY, R., GARBER, M., GNIKKE, A., NUSBAUM, C., HACOHEEN, N. et al. (2011). Metabolic labeling of rna uncovers principles of rna production and degradation dynamics in mammalian cells. *Nat. Biotechnol.* **29** 436–442.
- RABANI, M., RAYCHOWDHURY, R., JOVANOVIC, M., ROONEY, M., STUMPO, D. J., PAULI, A., HACOHEEN, N., SCHIER, A. F., BLACKSHEAR, P. J. et al. (2014). High-resolution sequencing and modeling identifies distinct dynamic rna regulatory strategies. *Cell* **159** 1698–1710.
- RUGGERO, D. (2009). The role of myc-induced protein synthesis in cancer. *Cancer Res.* **69** 8839–8843.
- RUMMEL, T., SAKELLARIDI, L. and ERHARD, F. (2023). grandr: A comprehensive package for nucleotide conversion rna-seq data analysis. *Nat. Commun.* **14** 3559.
- SCHOFIELD, J. A., DUFFY, E. E., KIEFER, L., SULLIVAN, M. C. and SIMON, M. D. (2018). Timelapse-seq: Adding a temporal dimension to rna sequencing through nucleoside recoding. *Nat. Methods* **15** 221–225.
- SCHWALB, B., SCHULZ, D., SUN, M., ZACHER, B., DÜMCKE, S., MARTIN, D. E., CRAMER, P. and TRESCH, A. (2012). Measurement of genome-wide rna synthesis and decay rates with dynamic transcriptome analysis (dta). *Bioinformatics* **28** 884–885.
- SLACK, F. J. and CHINNAIAN, A. M. (2019). The role of non-coding rnas in oncology. *Cell* **179** 1033–1055.
- STINE, Z. E., WALTON, Z. E., ALTMAN, B. J., HSIEH, A. L. and DANG, C. V. (2015). Myc, metabolism, and cancer. *Cancer Discov.* **5** 1024–1039.
- SUN, S., HOOD, M., SCOTT, L., PENG, Q., MUKHERJEE, S., TUNG, J. and ZHOU, X. (2017). Differential expression analysis for RNAseq using Poisson mixed models. *Nucleic Acids Res.* **45** e106–e106.
- TAN, J. Y., BIASINI, A., YOUNG, R. S. and MARQUES, A. C. (2020). Splicing of enhancer-associated lincnas contributes to enhancer activity. *Life Sci. Alliance* **3** e202000663.
- TATARINOVA, T., NEELY, M., BARTROFF, J., VAN GUILDER, M., YAMADA, W., BAYARD, D., JELLIFFE, R., LEARY, R., CHUBATIUK, A. et al. (2013). Two general methods for population pharmacokinetic modeling: Non-parametric adaptive grid and non-parametric Bayesian. *J. Pharmacokinet. Pharmacodyn.* **40** 189–199.
- THE GENE ONTOLOGY CONSORTIUM (2019). The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47** D330–D338. <https://doi.org/10.1093/nar/gky1055>

- TIBERI, S. and ROBINSON, M. D. (2020). Bandits: Bayesian differential splicing accounting for sample-to-sample variability and mapping uncertainty. *Genome Biol.* **21**.
- TUERK, A., WIKTORIN, G. and GÜLER, S. (2017). Mixture models reveal multiple positional bias types in rna-seq data and lead to accurate transcript concentration estimates. *PLoS Comput. Biol.* **13** 1–25.
- UVAROVSKII, A. and DIETERICH, C. (2017). pulser: Versatile computational analysis of rna turnover from metabolic labeling experiments. *Bioinformatics* **33** 3305–3307.
- VANDEVENNE, M., DELMARCELLE, M. and GALLEN, M. (2019). RNA regulatory networks as a control of stochasticity in biological systems. *Front. Genet.* **10** 403. <https://doi.org/10.3389/fgene.2019.00403>
- WACHUTKA, L., CAZZI, L., GAGNEUR, J. and CRAMER, P. (2019). Global donor and acceptor splicing site kinetics in human cells. *eLife* **8**. <https://doi.org/10.7554/eLife.45056>
- WADE, S. and GHAHRAMANI, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Anal.* **13** 559–626. [MR3807860 https://doi.org/10.1214/17-BA1073](https://doi.org/10.1214/17-BA1073)
- WALKER, S. and WAKEFIELD, J. (1998). Population models with a nonparametric random coefficient distribution. *Sankhya, Ser. B* **60** 196–214. [MR1717082](https://doi.org/10.2307/2346122)
- YU, G., WANG, L.-G., HAN, Y. and HE, Q.-Y. (2012). clusterProfiler: An R package for comparing biological themes among gene clusters. *Omics. J. Integr. Biol.* **16** 284–287.
- ZEISEL, A., KÖSTLER, W. J., MOLOTSKI, N., TSAI, J. M., KRAUTHGAMER, R., JACOB-HIRSCH, J., RECHAVI, G., SOEN, Y., JUNG, S. et al. (2011). Coupled pre-mrna and mrna dynamics unveil operational strategies underlying transcriptional responses to stimuli. *Mol. Syst. Biol.* **7**.