Multiple latent clusterisation 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 propose a hierarchical Bayesian approach to infer the RNA synthesis, processing, and degradation rates from sequencing data. We parametrise kinetic rates with novel functional forms and estimate the parameters through a Dirichlet process defined at a low level of hierarchy. Despite the complexity of this approach, we manage to perform inference, clusterisation and model selection simultaneously. We apply our method to investigate transcriptional and post-transcriptional responses of murine fibroblasts to the activation of proto-oncogene MYC. We uncover a widespread choral regulation of the three rates, which was not previously observed in this biological system.

Keywords: Dirichlet Process, kinetic rates, RNA

1. Introduction

RNA is one of the most important actors in the context of cellular biology and it is involved, directly or indirectly, in any process that occurs inside a cell. This molecule is a cornerstone of the information-flow which 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 (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 associated transcripts, in order to preserve its homoeostasis and adapt to the environment. This regulation is mediated by the choral action of several biological processes that affect the life cycle of the RNA molecules produced by these genes.

The RNA life-cycle in eukaryotic cells can be simplified the following three steps: the synthesis of premature RNA molecules in the nucleus, their processing into mature transcripts (which includes exporting the cytosol), and mature RNA cytoplasmic degradation. The characterisation 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. A RNA life-cycle investigation requires the experimental quantification of the gene expression levels. The state-of-the-art approach

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

to perform this task is *Next Generation RNA sequencing* (RNA-Seq). In a typical experiment, this technique simultaneously provides the average expression level per cell for thousands of genes, at a low cost and with a limited experimental effort (Goodwin et al., 2016). For these reasons, a remarkable number of public RNA-Seq datasets are currently available, and easily accessible, through such open repositories as the *Gene Expression Omnibus project* (Edgar et al., 2002).

A large amount of literature is available on the analysis of RNA-Seq data. In some papers, mixture models are used on the observed data to identify differences in gene expression levels, see, for example, "RNA-Seq by expectation-maximization" (Li and Dewey, 2011), "Cufflinks" (Trapnell et al., 2013), "Casper" (Rossell et al., 2014), the new approach developed by Papastamoulis and Rattray (2018), and the works of de Souto et al. (2008), Oyelade et al. (2016), or Saelens et al. (2018). Some of these tools have also been used for the identification of genes which are differentially expressed under multiple experimental conditions, this being 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 full 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 instance, an expressed gene is usually assumed to be actively transcribed in the biological condition under analysis, but this is not alway the case for very stable transcripts that remain in the cell long after their synthesis. Moreover, 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 modulation of the transcripts stability.

The mathematical modelling of the RNA life cycle can help to deconvolve the experimental data and characterise the different stages of theRNA metabolism. This can be formalised in terms of a network of chemical reactions (see for example de Pretis et al., 2015; Feinberg, 2019; Anderson and Kurtz, 2015), and can be modelled either deterministically or stochastically. The former approach is usually preferred, since it is compatible with standard RNA-Seq datasets originating from cell populations and because it leads to a system of linear ordinary differential equations (ODEs) whose time-dependent coefficients, 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 experimental data, and of these, DRiLL (Rabani et al., 2014) and INSPEcT (de Pretis et al., 2015; Furlan et al., 2020) provide a characterisation of all the crucial steps of the RNA life cycle from sequencing data. These tools are based on a least-squared estimation, and each gene is assumed to be independent of the others.

Motivated by a real data application, we here propose a novel approach to the inference of RNA life-cycle KRs from gene expression levels, cast in a Bayesian framework, and characterised by the application of mixture models defined using the Dirichlet process (DP) (Ferguson, 1973) on rate parameters. KRs are not directly observable, and the data-level of the mixture models are therefor latent quantities. This introduces difficulties in the model estimation, and we need to introduce a new parametrisation of the KR functions and to impose suitable identifiability constraints to overcome such difficulties. At

the same time we tailor an MCMC algorithm for the mixture models, based only on Gibbs steps (Casella and George, 1992), to avoid an increase in computational burden. This is made possible thanks to the adoption of a suitably modified likelihood function. Unlike other proposals, we estimate likelihood parameters, data standard deviations, KR functions, and latent clusterisations in a single Bayesian model, thereby allowing a coherent evaluation of the uncertainty. Moreover, by providing a clusterisation without the need of any post-processing, our proposal is able to gather genes in homogeneous groups and to extract and exploit the shared information. The inclusion of this clustering step in the inference procedure results in the estimation of parameters, at the gene level, even though if the number of experimental observations is limited; a standard situation in biology. Our new method is particularly suitable for investigating the common biological scenario, in which the cell synergically regulates the expression of groups of genes to respond to a modification of its environmental conditions (Allocco et al., 2004).

We apply the model to our motivating example, where data are collected to study the activation of the proto-oncogene MYC in murine fibroblasts (de Pretis et al., 2017). Moreover, we demonstrate the inferential gain provided by our approach, which results in the detection of smaller, but significant, modulations of post-transcriptional rates. From an interpretative point of view, the classification identifies groups of genes modulated by MYC activation, both directly and indirectly, and characterised by specific features, such as the steady-state value of the rate or the timing of its modulation in response to the stimulus. Since MYC is a transcription factor, the synthesis rate is the most informative layer of regulation and provides clusters of genes involved in basic cellular processes, cancer-related processes, and in the RNA metabolism. Nevertheless, we manage also to identify pervasive modulations of post-transcriptional rates, most likely due to either secondary regulations or the adaptation of the entire RNA life-cycle kinetics in response to MYC induced transcriptional stress.

The paper is organised as follows. We start by describing the experiment performed by de Pretis and colleagues 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 parametrise the KRs (Section 4); 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 5). The final section of the paper (Section 6) 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 7). The online supplemental material (SM), available on the web page of the journal, contains additional figures that may be useful to discuss the results but which are not essential for the comprehension of the paper.

2. Data description

Our dataset, taken from de Pretis et al. (2015), is organised as illustrated in Figure 1 A. It provides expression levels (in Reads Per Kilobase Million, RPKM) of premature, mature, and nascent RNA for more than 10.000 genes, at 11 time-points, and 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. This transcription factor

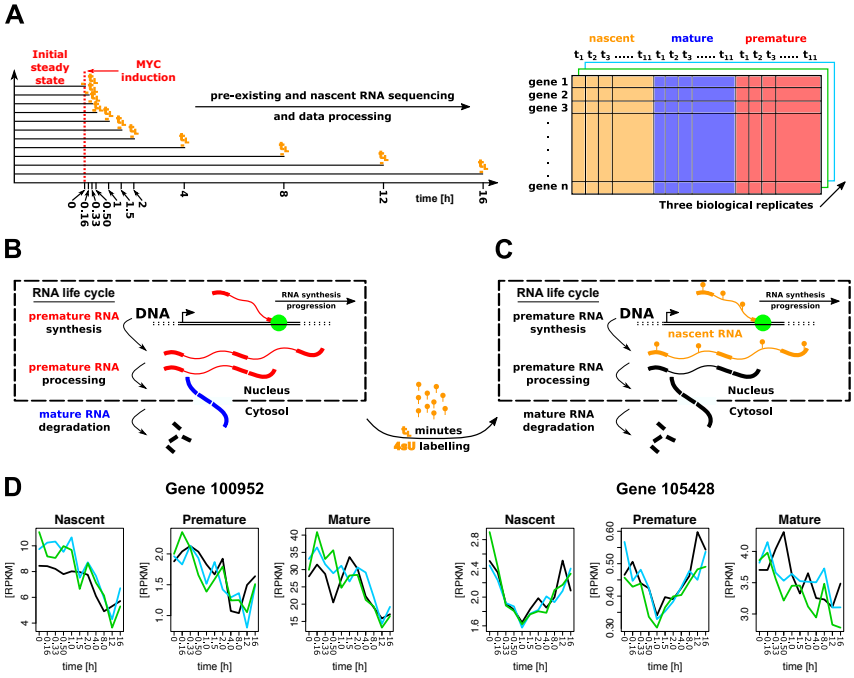

Fig. 1. (A) Experimental design used for the study of MYC activation in 3T9 cells and the associated dataset released by de Pretis et al. (2015). (B) RNA life cycle in eukaryotic cells and definition of premature (red) and mature (blue) RNA. (C) RNA metabolic labelling for nascent RNA (orange) quantification. (D) Gene expression profiles for two genes from the dataset; each replicate is represented by a specific colour.

plays a crucial role in the genesis and progression of tumours, and it is involved to a great extent 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 span of time, it is sequenced to quantify RNA expression levels. MYC activation is achieved through the expression of an artificial chimaera (Littlewood et al., 1995). This protein is natively inactive, i.e. it is unable to perform any function, but it can be rapidly 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, $^1/_6$h, $^2/_6$h, $^1/_2$h, 1h, $^3/_2$h, 2h, 4h, 8h, 12h, 16h (Figure 1 A). Each experiment, which was performed on independent samples, was replicated three times, and gave 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 (labelling time). 4sU is incorporated in the transcripts produced from that moment for the entire labelling time (nascent RNA) and is later exploited to physically separate them from the other RNA molecules (pre-

existing RNA). This portion of the transcriptome can be sequenced through standard RNA-Seq (Dölken et al., 2008).

de Pretis et al. (2015) focused on a set of 4909 transcriptional units, classified as MYC targets through a chromatin immunoprecipitation sequencing experiment, and altered in their kinetics. We decided to restrict our study to the same group of genes so that our results could be compared with the aforementioned authors' result, which are obtained with the INSPEcT tool. However, it was not possible to analyse 12 transcriptional units, because they had negative expression levels. At 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 is probably associated with a more complex post-transcriptional scenario.

3. The likelihood

For each gene $g \in \{1, \cdots, G\}$, time-point $t \in \{t_1, \cdots, t_T\}$ and replica $h \in \{1, \cdots, H\}$. Let $\mathbf{Y}_{g,h}(t) = (Y_{1,g,h}(t), Y_{2,g,h}(t), Y_{3,g,h}(t))'$ denote the measured expression levels of premature, mature and nascent RNA.

We assume that the observations $\mathbf{Y}_{g,h}(t)$ are noisy versions of the true unobserved values denoted with $\mathbf{x}_g(t) = (p_g(t), m_g(t), n_g(t))'$. Since $\mathbf{Y}_{g,h}(t)$ needs to have positive components, we model it as follows:

$$\mathbf{Y}_{g,h}(t) \sim N_{>0}(\mathbf{x}_g(t)\boldsymbol{\rho}_h(t), \mathrm{diag}(\boldsymbol{\tau}_g(t)), \tag{1}$$

where $N_{>0}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ indicates a (truncated) normal distribution with mean vector $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$, and with the components restricted to $\mathbb{R}^+$. In our case, the covariance matrix is diagonal, with diagonal elements collected in the vector $\boldsymbol{\tau}_g(t) \in (\mathbb{R}^+)^3$. The vector $\boldsymbol{\rho}_h(t) = (1, 1, \rho_h(t))$ is a scaling factor that is required to normalise the nascent RNA libraries to the pre-existing RNA counterparts (Miller et al., 2011; Rabani et al., 2011, 2014; de Pretis et al., 2015)

As shown in previous works, see, for example, Pavelka et al. (2004) and Subramaniam and Hsiao (2012), $\mathbf{x}_g(t)$ affects both the mean and the variance of $\mathbf{Y}_{g,h}(t)$. The effect of $\mathbf{x}_g(t)$ on the variance of $\mathbf{Y}_{g,h}(t)$ is modelled through a linear relation between the logarithms of their components:

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

The subject of the next subsection is the mathematical model of the latent gene expression levels $\mathbf{x}_g(t)$.

3.1. A mathematical model of the RNA life cycle

At the current state-of-the-art the life cycle of RNA molecules, in eukaryotic cells (e.g. mammals and plants), is divided into three sub-processes (Figure 1 B). The first 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 original task (e.g. protein translation). Premature RNA requires structural modifications and/or exporting to the cytosol. These steps constitute the second stage of the RNA life cycle, which is named processing. The product of premature RNA processing is mature RNA, which can eventually be degraded by the cell that concludes the RNA life cycle. The process may be described by the following network of chemical reactions

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

where P_g and M_g denote premature and mature RNA for gene g, respectively. The empty-set symbols are used to emphasise that premature RNA is synthesised from DNA without consuming resources, and mature RNA is subject to degradation. The symbols $k_{1,g}(t), k_{2,g}(t)$ and $k_{3,g}(t)$ are the KRs of the synthesis, processing and degradation respectively; they are both time and gene dependent. A system of ODEs that translates the reaction network (2) in mathematical terms is the following

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

Indeed, the effect of the processing of premature into mature RNA at rate $k_{2,g}(t)$ is to decrease $p_g(t)$ and correspondingly increase $m_g(t)$. The degradation of mature RNA 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 very difficult to identify all three KRs. Anotherr variable, the so-called nascent RNA, is usually included to ameliorate the identifiability,(Dölken et al., 2008; Miller et al., 2011; Rabani et al., 2011, 2014; de Pretis et al., 2015). Nascent RNA is the amount of total RNA (both premature and mature) synthesised by the cell in a short span of time and it can be quantified by 4sU-Seq (Figure 1 C). Nascent RNA is, by definition and according to the experimental set-up, absent at the beginning of the experiment. It is produced during the brief labelling time (t_L), according to the same dynamics as the pre-existing counterpart. However, the effect of degradation, in such a short time, can be neglected. The expression level of the premature ($p^\star_g(t)$) and mature ($m^\star_g(t)$) nascent RNA is therefor ruled by the following equations

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

The sum $n_g(t) = p^\star_g(t) + m^\star_g(t)$ is the *nascent* RNA level. By summing the two previous equations, one obtains that the amount $n_g(t)$ of nascent RNA only varies according to the effect of the synthesis rate:

$$\dot{n}_g(t) = \sigma_g(t). \tag{4}$$

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:

$$n_g(t) = k_{1,g}(t)t_L, \tag{5}$$

which is the third equation that is added to model (3). Equation (5) facilitates the estimate of the synthesis rate.

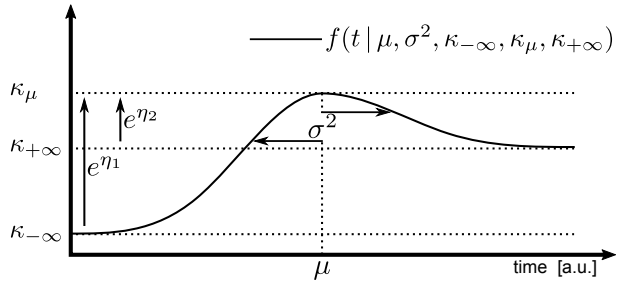

Fig. 2. Graphical representations of the KR parametrisation. It should be noted that e^{η_1} and e^{η_2} are shown to indicate the section of the function they determine, but they are not equal to the length of the arrows, see equation (8).

4. KRs parametrisation

In several biological experiments, a cell culture is perturbed and gene expression levels are repeatedly measured after the perturbation in order to understand which genes are involved in the response. By adopting the model described above, it is possible to shed light on the fundamental mechanisms that a cell uses to regulate gene expression levels by modulating the of synthesis, processing, and degradation rates.

The typical shapes that we expect the rates to take on in response to a certain perturbation of the environment are generally assumed to be constant (some rates are not altered), monotonic (both increasing and decreasing), and peak-like functions. These shapes account for both permanent and transient modulations of the KRs and have already been applied successfully 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).

The first novelty of our proposal is that we introduce a unique parametric family of functions which, for different values of the parameters, can cover all such characteristic shapes. Others approaches use different functional forms and, for each gene, select the best one with external criteria, e.g. the log-likelihood ratio test.

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 $f(t \mid \mu, \sigma^2, \kappa_{-\infty}, \kappa_\mu, \kappa_{+\infty})$ to which all the KRs belong in the following way:

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

where $\kappa_{-\infty}$, κ_μ and $\kappa_{+\infty}$ all belong to $\mathbb{R}^+$. The 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 2). It is easy to see that:

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

Examples of the forms that can be obtained with (6), by changing its parameters, are shown in Table 4. As we can be seen, all the standard shapes (constant, increasing/decreasing, peak-like) are possible.

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. steady-state, and it is analysed 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-change*s in computational biology and are usually used to measure modulations with respect to the baseline level $f(t \,|\, \mu, \sigma^2, \kappa_{-\infty}, \kappa_\mu, \kappa_{+\infty})$. We define

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

The parameters μ, σ^2, η_1, η_2 are all related to the characterisation of the response to perturbations. In particular μ and σ^2 characterise the *temporal* location and duration of the response, while η_1 and η_2 determine the typical shape, as highlighted in Table 4. We collect these four parameters into a single vector that we denote with $\boldsymbol{\theta}$, to obtain a more compact notation.

The family of functions (7) can now be re-parametrised as $f(t \,|\, \beta, \boldsymbol{\theta})$ with

$$\beta = \kappa_{-\infty} \qquad \text{and} \qquad \boldsymbol{\theta} = (\mu, \sigma^2, \eta_1, \eta_2).$$

Identification constraints Although the family of functions (7) is well defined for all real values of μ and positive values of σ^2, identifiability and interpretability issues can 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, the function f in the interval $[t_1, t_T]$, with any arbitrary choice of η_1 and η_2, is indistinguishable from a constant one, which should instead be given by $\eta_1 = \eta_2 = 0$ (Table 4). For this reason, identifiability constraints are needed.

A main requirement is that the value of the function (6) at time-points t_1 and t_T should be "close" to $\kappa_{-\infty}$ and κ_∞, respectively, which means that the most relevant part of the function graph lies within the observed time window. For peak-like shapes ($\eta_1\eta_2 > 0$, cf. Table 4) we ask that

$$\begin{aligned} |f(t_1 | \, \beta, \boldsymbol{\theta}) - \kappa_{-\infty}| &\leq 0.01|\kappa_\mu - \kappa_{-\infty}|, \\ |f(t_n | \, \beta, \boldsymbol{\theta}) - \kappa_{+\infty}| &\leq 0.01|\kappa_\mu - \kappa_{+\infty}|, \end{aligned} \tag{9}$$

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	

which implies two conditions:

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

Notice that, one of the log-fold-changes vanishes for monotonic shapes, however, we can still derive an identifiability condition by requiring that (9) holds, e.g. when $\eta_1 = 0$, we have a c-up or c-down shape, and $\kappa_\mu = \kappa_{-\infty}$. 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 therefor

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

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

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

We should take into consideration that monotonic up-up and down-down are intermediate states between, respectively, a c-up and an up-c, and a c-down and a down-c, respectively. 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:

$$\mu - \xi_1 \sqrt{-2\sigma^2 \log(0.01)} \geq t_1 \qquad \text{and} \qquad \mu + \xi_T \sqrt{-2\sigma^2 \log(0.01)} \leq t_T \tag{13}$$

where $\xi_1 = \left| \frac{\eta_1}{\eta_1 - \eta_2} \right|$ and $\xi_T = 1 - \xi_1$.

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

5. The latent clustering models

As mentioned in Section 4, $\beta_{j,g}$ is interpreted as the baseline level of the rate $k_{j,g}(t)$, while $\boldsymbol{\theta}_{j,g} = (\mu_{j,g}, \sigma_{j,g}, \eta_{1,j,g}, \eta_{2,j,g})$ characterises the modulation of the rate in response to a perturbation of the environment. It is biologically reasonable to allow groups of genes that have a similar baseline level to respond differently to a perturbation. For this reason, we introduce mixture models, based on the DP, at the bottom level of the model hierarchy, for the parameters $\beta_{j,g}$ and $\boldsymbol{\theta}_{j,g}$, for each $j \in \{1, 2, 3\}$.

Unlike most of the DP applications, where the mixture is at the first level of the hierarchy i.e. on the observed data level, we introduce mixture models over the parameters of the non-observable KRs which define the time-varying coefficients of an ODE system. For this reason, it is necessary to ensure that the sampling of the DP is easy, with as many Gibbs updates as possible for the mixture parameters.

Moreover, we also want to ensure that the model can discriminate between the possible shapes of $k_{j,g}(t)$. It is not trivial to define a distribution over the domain $\mathcal{D}$, i.e. the space where the parameters $(\beta_{j,g}, \boldsymbol{\theta}_{j,g})$ are identifiable. What we propose here is to let $(\beta_{j,g}, \boldsymbol{\theta}_{j,g})$ be defined over the whole $\mathbb{R}^5$ but, each time they are outside $\mathcal{D}$, the posterior distribution must have a very small density, which means that parameters outside $\mathcal{D}$ are almost never sampled in the MCMC algorithm. We do this by changing the data likelihood (equation (1)) with the following one:

$$\begin{cases} \mathbf{Y}_{g,h}(t) \sim N_{>0}(\mathbf{x}_g(t)\boldsymbol{\rho}_{g,h}(t), \mathrm{diag}(\tau_g^p(t), \tau_g^m(t), \tau_g^n(t))), & \text{if } (\beta_{j,g}, \boldsymbol{\theta}_{j,g}) \in \mathcal{D} \\ Y_{g,h}(t) \sim N(\mathbf{0}, 10^{10000}\mathbf{I}), & \text{otherwise,} \end{cases}$$

which gives an almost null posterior support to all the parameter values outside $\mathcal{D}$.

A second issue that has to be solved is that if the marginal prior distribution of the log-fold-changes $\eta_{1,g,j}$ and $\eta_{2,g,j}$ is continuous, then the posterior probability that at least one of the two is exactly 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 $\eta^*_{1,g,j}$ and $\eta^*_{2,g,j}$, related to $\eta_{1,g,j}$ and $\eta_{2,g,j}$ through the following:

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

If we assume a continuous distribution for $\eta^*_{1,g,j}$, as a result of (14), the distribution over $\eta_{1,g,j}$ is continuous over $(-\infty, 0) \cup (0, \infty)$ and has a point mass at 0 equal to the cumulative distribution of $\eta^*_{1,g,j}$ between $-\xi$ and ξ. A similar result holds for $\eta_{2,g,j}$.

We can now work with the parameters $\beta_{j,g}$ and $\boldsymbol{\theta}^*_{j,g} = (\mu_{j,g}, \sigma^2_{j,g}, \eta^*_{1,g,j}, \eta^*_{2,g,j})$, which are all defined over $\mathbb{R}$, and we obtain continuous distributions. 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)

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
$Y^p_{\cdot,\cdot}(\cdot)$	0.650	0.517	0.489
$Y^m_{\cdot,\cdot}(\cdot)$	0.653	0.517	0.501
$Y^n_{\cdot,\cdot}(\cdot)$	0.643	0.528	0.508

formalised as:

$$\begin{aligned} \beta_{j,g} \,|\, \zeta^{\beta}_{z^{\beta}_{j,g}}, \omega^{\beta}_{j,z^{\beta}_{j,g}}, z^{\beta}_{j,g} &\sim N(\zeta^{\beta}_{j,z^{\beta}_{j,g}}, \omega^{\beta}_{j,z^{\beta}_{j,g}}), \\ z^{\beta}_{j,g} \,|\, \pi^{\beta}_{j,g} &\sim \text{Discrete}(\pi^{\beta}_{j,g}), \\ \pi^{\beta}_{j,g} \,|\, \alpha^{\beta}_{j} &\sim \text{GEM}(\alpha^{\beta}_{j}), \\ \zeta^{\beta}_{j,k} &\sim N(M_0, V_0), \\ \omega^{\beta}_{j,k} &\sim IW(\nu_0, \psi_0), \end{aligned} \tag{15}$$

and

$$\begin{aligned} \boldsymbol{\theta}^{*}_{j,g} | \boldsymbol{\zeta}^{\boldsymbol{\theta}}_{j,z^{\boldsymbol{\theta}}_{j,g}}, \boldsymbol{\Omega}^{\boldsymbol{\theta}}_{j,z^{\boldsymbol{\theta}}_{j,g}}, z^{\boldsymbol{\theta}}_{j,g} &\sim N(\boldsymbol{\zeta}^{\boldsymbol{\theta}}_{j,z^{\boldsymbol{\theta}}_{j,g}}, \boldsymbol{\Omega}^{\boldsymbol{\theta}}_{j,z^{\boldsymbol{\theta}}_{j,g}}), \\ z^{\boldsymbol{\theta}}_{j,g} | \boldsymbol{\pi}^{\boldsymbol{\theta}}_{j,g} &\sim \text{Discrete}(\boldsymbol{\pi}^{\boldsymbol{\theta}}_{j,g}), \\ \boldsymbol{\pi}^{\boldsymbol{\theta}}_{j,g} | \alpha_j &\sim \text{GEM}(\alpha^{\boldsymbol{\theta}}_{j}), \\ \boldsymbol{\zeta}^{\boldsymbol{\theta}}_{j,k} &\sim N(\mathbf{M}, \mathbf{V}), \\ \boldsymbol{\Omega}^{\boldsymbol{\theta}}_{j,k} &\sim IW(\nu, \boldsymbol{\Psi}). \end{aligned} \tag{16}$$

In models (15) and (16), $z^{\lambda}_{j,g}$ and $z^{\boldsymbol{\theta}}_{j,g}$ are the discrete random variables that 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 that the allocation variable, $\lambda_{j,g}$ and $\boldsymbol{\theta}^{*}_{j,g}$ are normally distributed with parameters that have standard priors. All random quantities in model (15) can easily be updated in the MCMC algorithm using only Gibbs steps, thereby facilitating the implementation of the model.

6. Real data application

Before the discussion of the results obtained from the motivating dataset, we first show how our model performs with respect to some competitive approaches.

6.1. Multiple inference method comparison

We compare the performance of our model (M1) with a simplified version of our proposal, which assumes $z^{\beta}_{j,g} = z^{\boldsymbol{\theta}}_{j,g} = 1$ for all j and g (i.e. no mixture models, a single distribution for each parameter - M2), and the frequentist approach implemented in INSPEcT (M3). Comparisons are conducted in terms of predictive power and interpretation.

The model is implemented in R/C++ and uses OpenMP (OpenMP Architecture Review Board, 2008) for parallel computing. Our method is estimated using 32 cores,

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

		constant	peak+	peak-	monotonic+	monotonic-
M1	$k_{1,\cdot}(\cdot)$	0.019	0.041	0.223	0.373	0.343
	$k_{2,\cdot}(\cdot)$	0.515	0.029	0.102	0.312	0.041
	$k_{3,\cdot}(\cdot)$	0.460	0.017	0.408	0.046	0.069
M2	$k_{1,\cdot}(\cdot)$	0.100	0.050	0.147	0.357	0.345
	$k_{2,\cdot}(\cdot)$	0.641	0.027	0.117	0.171	0.044
	$k_{3,\cdot}(\cdot)$	0.703	0.031	0.060	0.119	0.087
M3	$k_{1,\cdot}(\cdot)$	0.013	0.128	0.365	0.229	0.265
	$k_{2,\cdot}(\cdot)$	0.712	0.071	0.192	0.010	0.016
	$k_{3,\cdot}(\cdot)$	0.801	0.109	0.066	0.009	0.015

2500000 iterations, burnin 2425000, thin 30, therefor with 2500 posterior samples, with $\xi = 10$; the computations take 20 days. We set $\mathbf{M} = \mathbf{0}$, $\mathbf{V} = 100\mathbf{I}$, $M_0 = 0$, $V_0 = 100$, $\mathbf{\Psi} = \mathbf{I}$, $\nu = 5$ and $\nu_0 = 2$. We use a $N(0, 100)$ for $\beta_{j,0}$, $\beta_{j,1}$ as prior distribution, while each $\mathrm{SF}_h(t)$ has a $G(1, 1)$. Since the number of latent mixtures in the DP depends on parameter α_j^ℓ, we view it as a random parameter with $G(1, 1)$ prior. We use the same priors and iterations for M2, while, for M3, we retrieved the modelled KRs for the 4897 genes of interest from the INSPEcT object, released as supplementary material by de Pretis et al. (2017).

First, we want to assess a goodness-of-fit for the three models for each Y_j, to evaluate wheter our proposal can be considered better than the others. For this purpose, we use the continuous ranked probability score (CRPS) (Gneiting and Raftery, 2007), since this index is suitable for comparing the predictive ability of models based on different data distributional assumptions. We compute this index for each modelling approach and experimental condition (i.e. RNA species, replicate, and time-point) and, in Table 6, we report the fraction of the times the first model has a lower CRPS index than the second one, for any pair-wise comparison. The results show that M1 is better able to recapitulate the data.

We extend the analysis 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, compared to M3 (0.52, 0.52 and 0.38 on average, respectively, Figure 28-SM). It follows that the Bayesian approaches tend to fit all the RNA species profiles with 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.

The fraction of rates modeled as constant, monotonic, or peak-like functions are reported In Table 3, for each inference method. In the case of modulation, we also distinguish between increasing (+) or decreasing (-) kinetics (for peak-like responses, $\eta_1 > 0$ and $\eta_2 > 0$ or $\eta_1 < 0$ and $\eta_2 < 0$, respectively). MYC is a transcription factor, and we can therefore 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

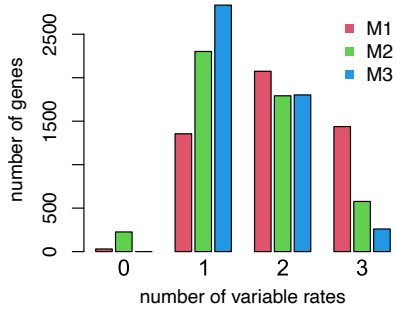

Fig. 3. Number of genes (y-axis) with a given number of non-constant rates (x axis) for M1, M2 and M3.

percentage is higher for M2 (10%). Nevertheless, k_1 is still by far the most variable rate.

Focusing on the genes modulated in synthesis, we can observe that all the inference methods predict a higher fraction of repressed genes (monotonic- or peak-) than the induced counterpart ($0.49-0.63$ versus $0.36-0.42$). However, the complexity of the resulting models is different, with 49% of the modulations in synthesis described by M3 as peak-like functions against 20% and 26% for M1 and M2, respectively.

The higher complexity of the synthesis rate profiles predicted by M3, is accompanied by a lower fraction of post-transcriptional modulations. M3 in fact predicts 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 choral action of the three kinetic rates than M3 (Figure 3 and 29-SM).

Despite the prevalent role of synthesis in shaping the MYC response, indirect and less intense post-transcriptional regulations are expected due to, for example, the known feedbacks that link the three layers of the RNA life cycle (Sun et al., 2012; Eser et al., 2014; McManus et al., 2015).

In light of the higher goodness of fit of M1 than M3, we can conclude that INSPEcT fails to capture the more complex regulatory scenarios inferred by our novel Bayesian method.

6.2. MYC response analysis

In this section, we analyse the model inferred by means of our approach in response to MYC activation in more detail. For simplicity, we compute the MAP clusterisation for each gene and variable, and we only discuss clusters that have at least 150 associated genes. A heatmap and a set of boxplots that recapitulate the temporal behaviour of the synthesis rate of log-fold-changes, compared to the initial time-point, are reported in Figures 4, for each cluster, ordered by the number of elements. We also display the μ versus σ^2 and η_1 versus η_2 plots. Both of these graphs provide valuable information about the shape of the responses in the cluster of interest (Table 4). Figures 6 and 7 are analogues of processing and degradation rates, respectively.

In order to support the discussion, we perform enrichment analyses, based on Gene Ontology (GO) annotations (Ashburner et al., 2000; Dessimoz and Škunca, 2017; The Gene Ontology Consortium, 2019), for each rate. For the sake of simplicity, the results

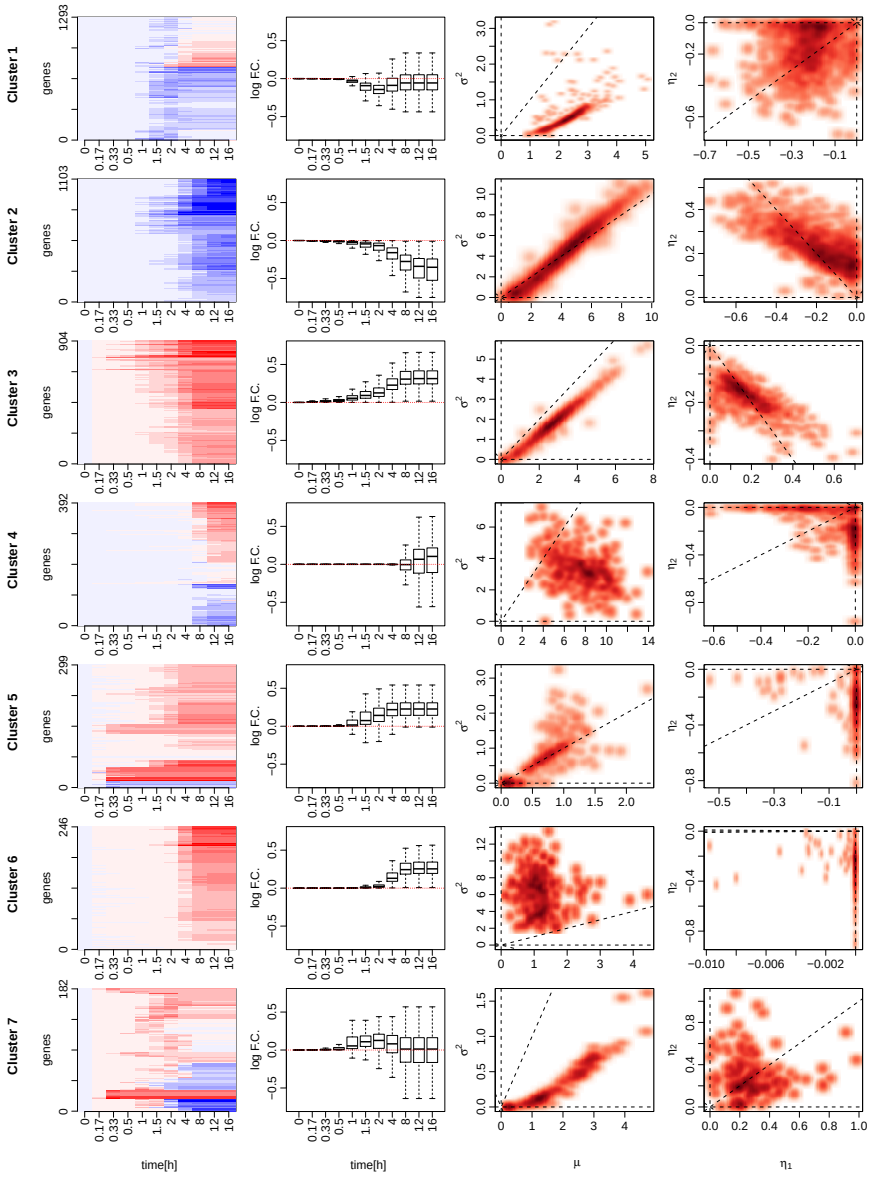

Fig. 4. Synthesis rate modulations in response to MYC activation for 7 clusters that are composed of at least 150 genes. (First column) Heatmap showing the log-fold-changes of the rate $\eta(t, 0)$, compared to the first time-point, for each gene in the cluster. (Second column) Boxplots showing the distribution of the synthesis rate log-fold-changes, compared to the first time-point. (Third column) μ versus σ^2 smooth-scatter plot. (Fourth column) η_1 versus η_2 smooth-scatter plot. The black dashed lines represent the horizontal and vertical axes and the bisector of the first and third quadrants.

of these analyses are mainly shown in the supplemental material. However, they are summarized in the main text, while Figure 5, which is reported as an example of a full output, refers to a specific case which has been selected both because of its interest and its

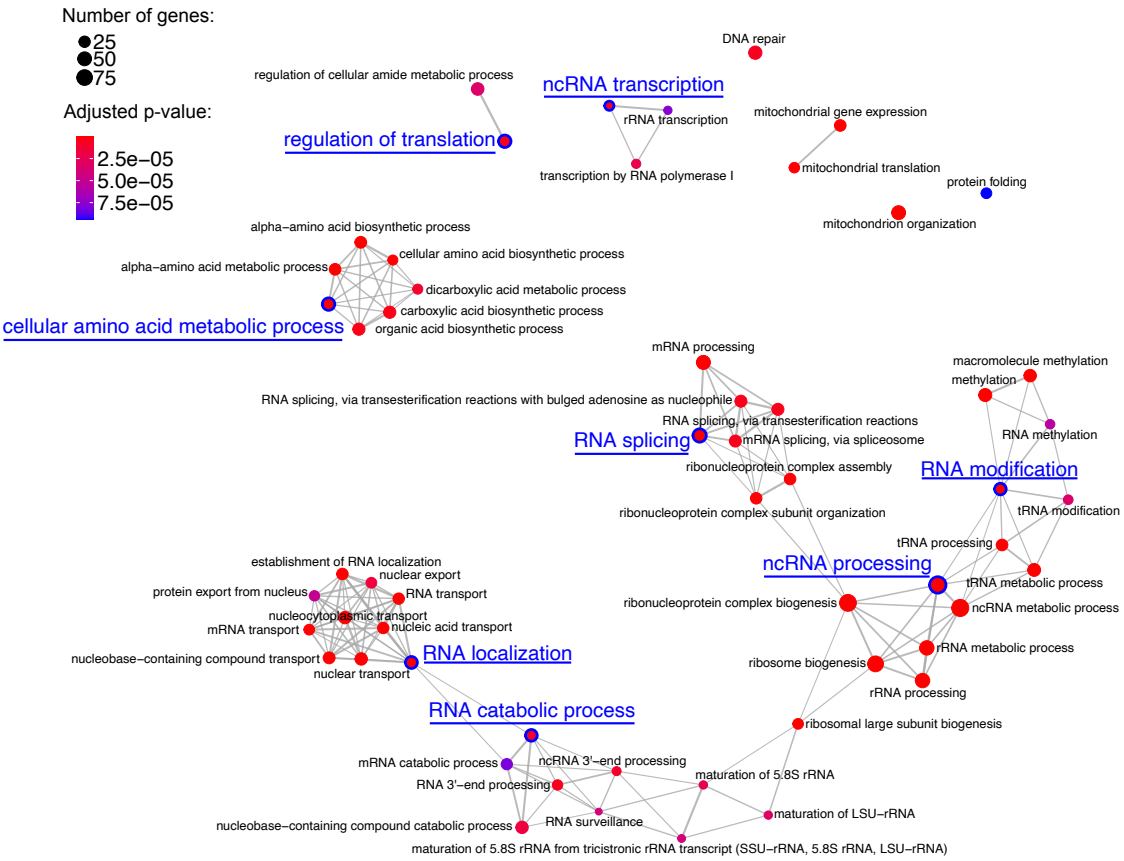

Fig. 5. Synthesis rate Cluster 3 Gene Ontology analysis. Biological processes (nodes of the network) associated with the genes belonging to the third cluster identified from the analysis of the synthesis rate modulations. The network structure is indicative of the semantic similarity of the terms; i.e., 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 colour 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 in blue.

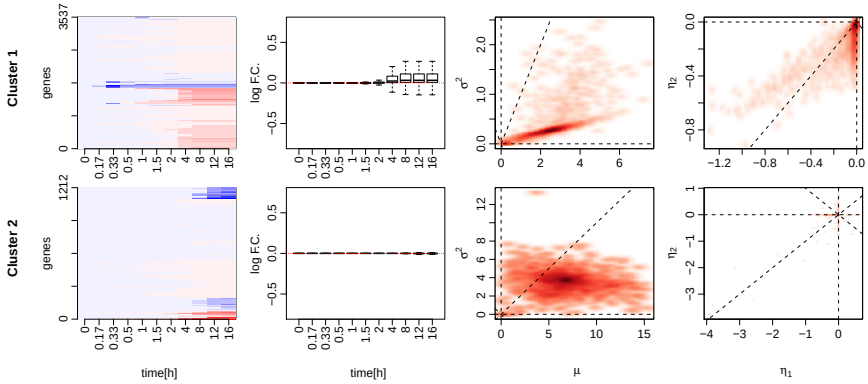

Fig. 6. Processing rate modulations in response to MYC activation for 2 clusters that are composed of at least 150 genes. (First column) Heatmap showing the log-fold-changes of the rate $\eta(t, 0)$, compared to the first time-point, for each gene in the cluster. (Second column) Boxplots showing the distribution of the processing rate log-fold-changes, compared to the first time-point. (Third column) μ versus σ^2 smooth-scatter plot. (Fourth column) η_1 versus η_2 smooth-scatter plot. The black dashed lines represent the horizontal and vertical axes and the bisector of the first and third quadrants.

graphical clarity. The GO enrichment analysis exploits an ontology which can be defined as a set of terms which, in our case, are pertinent to biological processes (e.g. "regulation of mRNA stability", "cellular response to stress" etc.), associated by means of relations (grey edges in Figure 5 - e.g. "is a", "regulates" etc.). 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 in the group of interest, compared to a background (a number of associated genes, that is proportional to the node size in Figure 5), that is a larger set which 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 (node colour in Figure 5), or q-values, selecting the most significant results. These terms (labels in Figure 5) 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).

Synthesis rate The graphical description of the clusters associated with the synthesis rate is shown in Figure 4.

Cluster 1 accounts for 1293 genes characterised by peak-type functions (η_1 and $\eta_2 < 0$) with the minimum between 1 and 3 hours and a small variance, which indicates quick responses. This behaviour can be explained as a side effect of MYC activation, which causes the polarisation of the resources required for the proficient transcription of the induced target genes (discussed in the following clusters). This response is compensated in the recovery of the transcriptional activity of these genes which then results, for a subset of such genes, in a light induction ($\eta_1 > \eta_2$). This indicates that these transcriptional units

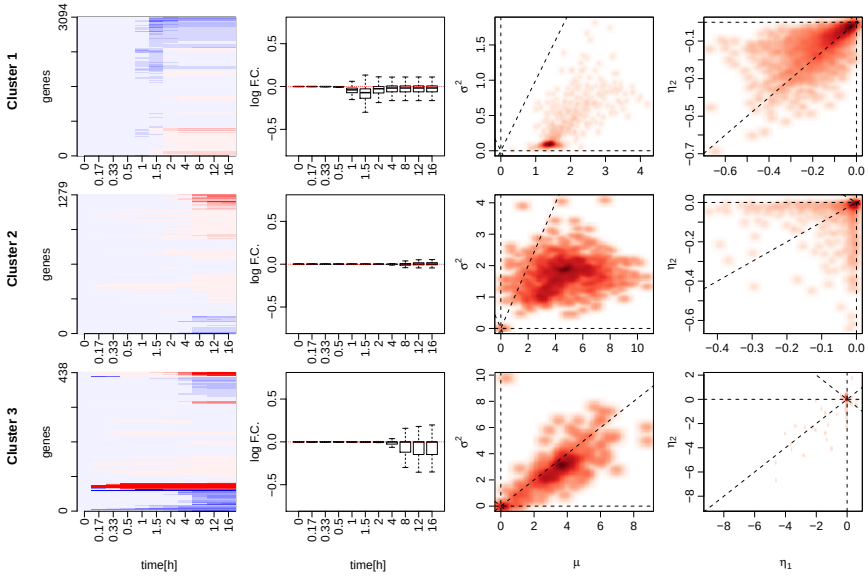

Fig. 7. Degradation rate modulations in response to MYC activation for 3 clusters that are composed of at least 150 genes. (First column) Heatmap showing the log-fold-changes of the rate $\eta(t, 0)$, compared to the first time-point, for each gene in the cluster. (Second column) Boxplots showing the distribution of the degradation rate log-fold-changes, compared to the first time-point. (Third column) μ versus σ^2 smooth-scatter plot. (Fourth column) η_1 versus η_2 smooth-scatter plot. The black dashed lines represent the horizontal and vertical axes and the bisector of the first and third quadrants.

are required by the cell for its homoeostasis.

The enrichment analysis points to such as basic processes as DNA metabolism and cell-cycle regulation, which are in line with the expectations, because they are known to be perturbed in an MYC-dependent manner (Dang, 2012; Chen et al., 2018); see Figure 1-SM.

Cluster 2 accounts for 1103 genes, characterised by a monotonic decrease of the synthesis rate ($\eta_1 < 0$ and $\eta_2 > 0$). The temporal response is more heterogeneous with μ and σ^2 spanning large domains. This cluster contains the genes that are involved in cell growth and development, cell adhesion, migration, and apoptotic signalling regulation (Figure 2-SM). All these terms 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 ($\eta_1 > 0$ and $\eta_2 < 0$) concentrated in the first 5 hours of the time-course. This behaviour indicates a potential direct MYC regulation. These genes are involved in coding and non-coding RNA metabolisms (Figure 5). They affect *RNA localisation*, 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*) which 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 regulation layers of the transcript 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 - ($\eta_1 \approx 0$ and $\eta_2 < 0$ or $\eta_1 < 0$ and $\eta_2 \approx 0$, respectively) late responding genes (μ 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 and 6 are composed of 299 and 246 monotonically induced genes ($\eta_1 \approx 0$) in response to MYC activation. These two clusters were split, due to their temporal responses, which are faster and more homogeneous in *Cluster 5* than in *Cluster 6*. The latter is partially related to non-coding RNA processing (see Figure 4-SM).

Finally, *Cluster 7* is composed of 182 genes characterised by a weak peak-like induction (η_1 and $\eta_2 > 0$), that is earlier in the time-course and quicker than those belonging to Cluster 1, as can be seen from the values of μ and σ^2 respectively.

Our model also provides a clustering of genes according to their steady-state values of the synthesis rate (β_1). The model identifies 7 clusters, enumerated with a progression of letters, with more than 150 genes. The Gene Ontology enrichment analysis returned more confused and less significant terms than those discussed above. *Cluster B* is partially related to cell-cycle regulation, while *Cluster E* and *Cluster F* are related to RNA metabolism (Figures 15-SM, 18-SM and 19-SM). Interestingly, a remarkable percentage ($\approx 30\%$) of the genes in *Cluster E* and *Cluster F* is also part of *Cluster 3* (Figure 26-SM). Noticeably, these are also the clusters that are characterised by the fastest 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 27-SM).

Processing rate A graphical description of the clusters associated with the processing rate is given in Figure 6. *Cluster 1* is composed of 3537 elements, which respond to a great extend 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 behaviour 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 characterised by small and late, both positive and negative, monotonic responses. These are probably secondary responses 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. It would be interesting to further investigate this population of transcripts through an experiment, with an ad-hoc temporal design. However, this is

beyond the scope of this manuscript.

The Gene Ontology enrichment analysis of *Cluster 2* does not provide any significant 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 (Figures 5-SM and 6-SM) and they disappear when the 4897 differentially expressed genes are used as the background instead of all the annotated ones (Figures 17-SM and 18-SM).

The same is true for all of the 5 clusters with more than 150 genes which our method returns for the processing steady-state rate.

Degradation rate The modulations of the degradation rate are divided into three clusters with more than 150 genes each, see Figure 7. *Cluster 1* is composed of 3094 elements characterised by very quick peak-type responses and with the minimum between 1 and 2 hours. As we have seen for k_2, this behaviour may be due to the coupling with the synthesis rate. *Cluster 2* and *Cluster 3* account for 1279 and 438 genes, respectively, characterised by late, both positive and negative, monotonic responses, which are probably secondary. The situation is similar to the one previously described for late processing rate modulations especially for *Cluster 2*, and could point to indirect regulations of the stability of the transcripts under-sampled due to the temporal design of the experiment. The Gene Ontology enrichment analysis results are analogues to those we discussed for the processing rate (Figures 7-SM and 8-SM), and also to those of the 5 steady-state clusters.

7. Final remarks

Motivated by a real data application, we here propose a Bayesian approach to the analysis of RNA expression levels. In our proposal, the experimental data are hypothesised 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, which are time- and gene-dependent. The KRs are the main object of inference, since they characterise the RNA life cycle and provide important insights into the analysis of gene expression levels. The temporal evolution of KRs is encoded in a single family of functions, defined by only 5 parameters that can be easily 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 KRs 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 of 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 a valuable piece of information.

Some identification problems have arisen, owing to the complex structure of our model. However, we have managed to solve them using latent variables and identification constraints in such a way that the MCMC algorithm can still only use Gibbs updates for the mixture parameters.

The results obtained with the new method are biologically relevant. The enrichment

analysis of the clusters results in meaningful sets of terms in the context of MYC biology, 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 specific biological system. However, our method manages to identify a remarkable fraction of genes as post-transcriptionally regulated, thus pointing to weak responses that are compatible with the adjustments of processing and degradation rates in response to the transcriptional perturbations and indirect secondary responses concentrated in the last portion of the time-course. Our method tends to describe the regulation of RNA metabolism more as a choral action of the three kinetic rates than the frequentist approach. The inclusion of Dirichlet-based clustering in the inference improves the goodness of fit.

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 in a much larger space, which would be difficult to handle. We preferred to acquire a full picture of each single response for our analysis, but we anticipate that our framework could also be adapted in this way.

We conclude by stressing that this proposal represents an inference framework for chemical reaction network coefficients which could be used to improve the methods currently available in computational biology to study dynamic phenomena in large omics datasets.

8. Implementation

The source 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.

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) for his critical comments on the manuscript.

References

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.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). "Gene Ontology: tool for the unification of biology." Nature Genetics, 25(1): 25–29.

Casella, G. and George, E. I. (1992). "Explaining the Gibbs Sampler." The American Statistician, 46(3): 167–174.

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.

de Souto, M. C., Costa, I. G., de Araujo, D. S., Ludermir, T. B., and Schliep, A. (2008). "Clustering cancer gene expression data: a comparative study." BMC Bioinformatics, 9(1): 497.

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

Eser, P., Demel, C., Maier, K. C., Schwalb, B., Pirkl, N., Martin, D. E., Cramer, P., and Tresch, A. (2014). "Periodic mRNA synthesis and degradation co-operate during cell cycle gene expression." Molecular Systems Biology, 10(1): 717.

Feinberg, M. (2019). Foundations of chemical reaction network theory, volume 202 of Applied Mathematical Sciences. Springer International Publishing.

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." The Annals of Statistics, 1(2): 209–230.

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.

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.

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.

Miller, C., Schwalb, B., Maier, K., Schulz, D., Dümcke, S., Zacher, B., Mayer, A., Sydow, J., Marcinowski, L., Dölken, L., Martin, D. E., Tresch, A., and Cramer, P. (2011). "Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast." Molecular Systems Biology, 7(1): 458.

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."

Oyelade, J., Isewon, I., Oladipupo, F., Aromolaran, O., Uwoghiren, E., Ameh, F., Achas, M., and Adebiyi, E. (2016). "Clustering Algorithms: Their Application to Gene Expression Data." Bioinform Biol Insights, 10: 237–253.

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.

Pavelka, N., Pelizzola, M., Vizzardelli, C., Capozzoli, M., Splendiani, A., Granucci, F., and Ricciardi-Castagnoli, P. (2004). "A power law global error model for the identification of differentially expressed genes in microarray data." BMC Bioinformatics, 5(1): 203.

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.

Saelens, W., Cannoodt, R., and Saeys, Y. (2018). "A comprehensive evaluation of module detection methods for gene expression data." Nature Communications, 9(1): 1090.

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.

Sun, M., Schwalb, B., Schulz, D., Pirkl, N., Etzold, S., Larivière, L., Maier, K. C., Seizl, M., Tresch, A., and Cramer, P. (2012). "Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation." Genome Research.

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.

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.