A nonlinear mixed-effects approach for the mechanistic interpretation of time-series transcriptomics data

Thibault Etienne, Charlotte Roux, Eugenio Cinquemani, Laurence Girbal, Muriel Cocaign-Bousquet, Delphine Ropers

A nonlinear mixed-effects approach for the mechanistic interpretation of time-series transcriptomics data

Thibault A. Etienne[1,2], Charlotte Roux[2], Eugenio Cinquemani[2], Laurence Girbal[2], Muriel Cocaign-Bousquet[2], and Delphine Ropers[*1]

[1]Univ. Grenoble Alpes, Inria, 38000 Grenoble, France
[2]TBI, Université de Toulouse, CNRS, INRAE, INSA, Toulouse, France

April 26, 2022

Motivation

Mechanistic models are essential to unravel the molecular mechanisms driving cellular responses. However, the integration of high-throughput data with mechanistic knowledge is limited by the availability of scalable computational approaches able to disentangle biological and technical sources of variation.
Results: We present an approach based on nonlinear mixed-effects modelling for the parameter estimation of large-scale mechanistic models from time-series transcriptomics data. It allows to factor out technical variability, to compensate for the limited number of conditions and time points by a population approach, and it incorporates mechanistic details to gain insight on the molecular causes of biological variability. We applied our approach for the biological interpretation of microarray and RNA-Seq gene expression profiles, with different levels of technical noise, but it is generalisable to numerous types of data. When integrated in a model describing the degradation kinetics of all cellular mRNAs, the data allowed to identify the targets of post-transcriptional regulatory mechanisms. Our approach paves the way for the interpretation of high-throughput biological data with more comprehensive mechanistic models.
Availability: The Monolix script for estimation and output files are freely available at `https://gitlab.inria.fr/tetienne/eccb_script`, together with the microarray data. The RNA-Seq dataset is being prepared for publication (Roux *et al.*, in preparation) and will be made available on demand upon acceptance of the article.
Contact: Delphine.Ropers@inria.fr

1 Introduction

Temporal omics data provide unprecedented views of cellular inner workings. They account for the molecular responses of living organisms to perturbations, resulting from regulatory mechanisms that ensure adaptation and survival. Statistical modelling tools are commonly used to analyse and interpret these data, from the clustering of gene expression profiles according to their similarity or characteristics to network inference (for reviews, Ding and Bar-Joseph 2020; Grigorov 2011; Lu *et al.* 2011; Noor *et al.* 2019; Straube *et al.* 2015). Achieving mechanistic insight requires moving from data-driven to mechanistic modelling approaches, by including prior knowledge of the relations between molecular responses and cell components (Androulakis *et al.*, 2007; Noor *et al.*, 2019). Mechanistic model parameters have a biological interpretation that provides insight on molecular mechanisms at stake. Mechanistic modelling is, however, a daunting task and examples of these models are poorly abundant in the literature (Noor *et al.*, 2019). Main issues arise at the time of estimating the parameters due to model and data complexity. Combinations of technological and practical limitations result in small numbers of biological and technical replicates that can be used and of time points that can be measured. Appropriate data normalisation is required to disentangle sources of noise, resulting from measurement errors and biological process heterogeneity, within and across studies (Liu *et al.*, 2019). The consequence of these shortcomings has been a coarse-graining of the models to help with estimation, at the price of an incomplete interpretation of the data: much biological insight remains buried.

The aim of this article is to develop a method for the estimation of large-scale mechanistic model parameters from dynamical transcriptomics data, addressing the above issues. We extend the nonlinear mixed-effects

*Correspondence: `delphine.ropers@inria.fr`

(NLME) modelling framework that allows to study variability between individuals in a population. Frequently used in pharmokinetics-pharmocodynamics studies of the variability of patient responses to drug absorption and elimination, NLME models have been successfully applied in other scientific fields as well (epidemiology, agriculture...) and, more recently, in cell biology to study cell-to-cell variability (Almquist *et al.*, 2015; Davidian and Giltinan, 2003; Lavielle, 2014; Lindstrom and Bates, 1990; Llamosi *et al.*, 2016; Loos *et al.*, 2018; Oddi *et al.*, 2019). In NLME modelling, the same individual response model is used to describe a physical or biological phenomenon observed for each individual and a statistical model describes why the model parameters vary between individuals. The information shared between individuals results in better estimates for both the individual and population parameters (Almquist *et al.*, 2015; González-Vargas *et al.*, 2016).

Unlike traditional NLME approaches, we model "mRNA-to-mRNA" variability, in an effort to identify biochemical principles behind the biological variability of mRNA concentrations. The idea is that the same cellular machineries determine mRNA intracellular levels, whose mechanism of action can be described by the same individual response model for each mRNA. For instance, in bacteria, a unique RNA polymerase transcribes all mRNAs and a single multiprotein complex, degradosme, is responsible for bulk mRNA degradation. Messenger RNAs can thus be regarded as forming a population of individuals subjected to the same cellular machineries. The variability between individual mRNAs reflects the intervention of additional regulatory mechanisms adjusting transcription or degradation to environmental conditions. In this study, we will focus on mRNA concentrations averaged over populations of cells, because many transcriptomics data are obtained in these conditions. We will show that NLME modelling is a suitable framework to analyse large, dynamical microarray and RNA-Seq datasets: 1) it separates the technical and biological variability of high-throughput data; 2) it leverages the limited number of replicates and time points by imposing a parameter distribution over the population; (3) it allows a knowledge-based interpretation of the datasets by incorporating mechanistic details; (4) it empowers *a posteriori* analyses of the biological variations between mRNAs to identify regulatory targets. Our method is a versatile approach, generalisable to other case studies.

We illustrate our approach by applying it to the study of bacterial mRNA decay, a key cellular process to adjust gene expression to environment (Dressaire *et al.* 2013; Esquerré *et al.* 2014; for review, Carpousis 2007). Degradation of bacterial mRNAs by the degradosome leads to short-lived messengers, typically between 3 and 8 minutes in *Escherichia coli*, up to 50 to 100 minutes in some extreme cases (Bernstein *et al.*, 2002; Chen *et al.*, 2015; Dressaire *et al.*, 2018; Esquerré *et al.*, 2014). The half-lives are generally determined following transcriptional arrest, by means of transcriptomics experiments monitoring at the genome-wide level the residual mRNA concentrations along time. The degradation profile of each mRNA reflects the activity of degradosome and regulatory mechanisms operational at the time of drug addition. It is classically fitted by simple exponential models, assuming a first-order reaction rate and independent degradation reactions, from which is then determined mRNA half life (Bernstein *et al.*, 2002; Chen *et al.*, 2015; Laguerre *et al.*, 2018; Moffitt *et al.*, 2016). In general, the degradation profiles are relatively similar, stemming from the fact that mRNAs are degraded in bulk by the same machinery, but differ at the level of mRNA abundance prior to drug addition and in their curvature. To mechanistically characterise this variability, we adapted a genome-wide model of mRNA degradation described in (Etienne *et al.*, 2020). It explicitly relates the degradation profiles with the degradation machinery and additional regulatory mechanisms involving, for instance, small RNAs and ribonucleoproteins (Etienne *et al.*, 2020). Using our NLME framework, we estimated mechanistic model parameters from a microarray and a RNA-Seq dataset on *Escherichia coli* bacteria growing in exponential phase (Esquerré *et al.* 2014; Roux *et al.*, in preparation). The method uncovers mRNAs known to be post-transcriptionally regulated, notably by small RNAs, such as mRNAs coding for enzymes in the central carbon metabolism and for ribosomal proteins.

2 Modelling approach

Fig. 1 gives an overview of our method, which takes as input preprocessed microarray or RNA-Seq time-series data. A mechanistic model relates their time-dependent evolution with cellular components and regulatory mechanisms, but model parameters are unknown. We model "mRNA-to-mRNA" variability using the NLME framework to estimate population and individual mRNA parameters. Biological variation between individual parameter values can be interpreted in relation to regulatory mechanisms.

2.1 Transcriptomics data

The two transcriptomics datasets used in this study were obtained in the same conditions (Esquerré *et al.* 2014; Roux *et al.*, in preparation). They quantify the residual mRNA concentration over time of *E. coli* bacteria grown exponentially at 0.63 h^{-1} in minimal medium supplemented with glucose until antibiotic treatment by the transcriptional inhibitor rifampicin. Three samples were taken right before the addition of rifampicin, and nine after. Measurements for each of the 4254 mRNAs at each time point consist of eight microarray probe

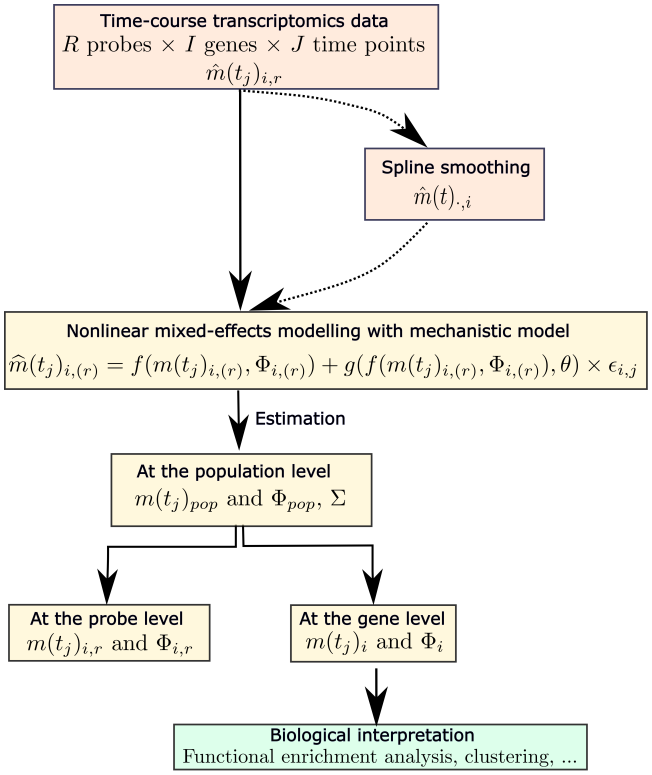

Figure 1: Nonlinear-mixed effects approach for the mechanistic interpretation of time-series transcriptomics data. The framework uses preprocessed microarray or RNA-Seq data, obtained from samples collected at time point t_j with $j = \{0, ..., J\}$ for cellular genes $i = \{1, ..., I\}$. They quantify the abundance $\hat{m}_i(t_j)$ of each mRNA. Probe intensities $r = \{1, ..., R\}$ are available for microarray data. An optional smoothing step enables to remove their variability (Section 3.2). NLME modelling using a mechanistic model gives estimates of population- and individual-mRNA parameters. The variability of the later has biological interpretation.

intensities corrected for the background or raw read counts for the RNA-Seq (Supplementary Section S1). Lowly expressed genes with signal intensities or read counts close to the background were filtered out. This leaves us with the same 2809 mRNAs in each dataset.

2.2 Mechanistic modelling

A variety of models can be used in NLME modelling. Here we use an ordinary-differential equation (ODE) model to describe mRNA concentrations averaged over a population of cells. We have shown in (Etienne *et al.*, 2020) that bacterial mRNA degradation can be assimilated to a macro-reaction catalysed by the degradosome, which we assume to follow Michaelis-Menten kinetics to reproduce the monotonous decrease of mRNA concentrations following transcription arrest. Owing to the fact that microarray and RNA-Seq experiments quantify mRNA abundances, we reformulate the model to describe the time-dependent evolution of mRNA quantities (see Supplementary Section S2 for details):

$$\frac{dm_i(t)}{dt} = -\frac{Vm \times m_i(t)}{Km_i + \frac{m_i(t)}{V}}, \text{with: } m_i(t_0) = m_i(\tau_i), \tag{1}$$

where $m_i(t)$ is the abundance of mRNA i (expressed in AU or count) at time t (min) and V is the cell volume (in milligram of cell dry weight, mgDW). Vm, the maximal velocity of the reaction (in UA or count/mgDW/min), is identical for all mRNAs. Km_i is the Michaelis constant of a given mRNA i (in UA or count/mgDW). Addition of rifampicin at time 0 inhibits the initiation of transcription. Degradation starts after a short delay τ_i (in min), during which RNA polymerase already engaged in the elongation of transcription completes mRNA synthesis. Km_i is an inverse measure of the affinity of degradosome for its target mRNA. For instance, mRNAs with small Km_i value are degraded faster than mRNAs with an average Km. This value depends on the sequence and structure characteristics of the mRNA, but also on regulatory factors such as small RNAs or RNA-binding proteins, which can either facilitate or block the binding of degradosome to mRNAs.

2.3 Formulation of the NLME problem

The fact that the same machinery degrades the bulk of mRNAs allows to consider a population approach based on NLME modelling for the estimation problem. We consider that the bulk mRNAs form a joint population, wherein the degradation kinetics of each individual mRNA can be described by the same response model, described in Eq. 1. Model parameters for each individual mRNA are drawn from the parameter distribution of the population. Similarly to NLME approaches applied to single-cell data (Dharmarajan *et al.*, 2019; Llamosi *et al.*, 2016), the objective is to harness differences between mRNA levels in order to improve the estimation of population distribution parameters and of individual parameters.

Concretely, we describe the time-series data with the following measurement model:

$$\hat{m}_{i,t_j} = f(m_i(t_j), \Phi_i) + g(f(m_i(t_j), \Phi_i), \theta) \times \epsilon_{i,j}. \tag{2}$$

Here, $\hat{m}_{i,t_j}$ is the measured concentration of mRNA $i = 1, \ldots, N$ at time t_j, $\epsilon_{i,j} \sim \mathcal{N}(0,1)$ is the residual error, and the function g the residual error model with a vector of noise parameters θ. Function f is the solution to the ODE model in Eq. 1.

The vector of parameters $\Phi_i = \{m_i(t_0), \tau_i, Vm_i, Km_i\}$ is drawn from a probabilistic distribution h characterised bya vector Φ_{pop} corresponding to the typical value of model parameters, also called fixed effects, and a variance-covariance matrix of the random effects Ω describing the variability between individual mRNAs. Here we will consider that h is a lognormal distribution to constrain individual parameters to positive values:

$$\Phi_i \sim Log\mathcal{N}(\Phi_{pop}, \Omega) \Leftrightarrow log(\Phi_i) = log(\Phi_{pop}) + \eta_i\,, \tag{3}$$

with $\eta_i \sim \mathcal{N}(0, \Omega)$.

Additional levels of variability can be considered, for instance in Section 3.2, to explicit technical variability between microarray probes (Lavielle and Mbogning, 2014; Panhard and Samson, 2009). This requires to introduce a new vector of parameters, $\Phi_{i,r}$, for individual mRNA i and probe r:

$$\Phi_{i,r} \sim Log\mathcal{N}(\Phi_{pop,r}, \Omega) \Leftrightarrow log(\Phi_{i,r}) = log(\Phi_{pop,r}) + \eta_i + \eta_{i,r} \tag{4}$$

with $\eta_i \sim \mathcal{N}(0, \Sigma)$, $\eta_{i,r} \sim \mathcal{N}(0, \Gamma)$, and $\Omega = \Sigma + \Gamma$ (Panhard and Samson, 2009). Σ is the variance-covariance matrix for inter-individual variability (between mRNAs) and Γ the variance-covariance matrix for intra-individual variability (between probes of a given mRNA).

The residual error model g in Eq. 2 should be adapted to the transcriptomics data. Measurement variance is considered to be proportional to probe intensities or read counts (Meacham *et al.*, 2011; Weng *et al.*, 2006). Systematic errors in transcriptomics data include non-zero background intensity level and hybridisation bias

Table 1: Estimated mechanistic model parameters from smoothed microarray data, microarray probe intensities, and RNA-Seq data. The parameters of the distribution (mode and the logarithm of the standard deviation) are given for each NLME model parameter, together with their standard error. The catalytic efficiency is the ratio of Vm_{pop} and Km_{pop}. a and b are residual error model parameters. The condition number is the ratio of the eigenvalues of the variance-covariance matrix

Data type and units:	**Smoothed microarray data** [AU]	**Microarray probe intensities** [AU]	**Raw RNA-Seq data** [count]
Vm_{pop} [AU (count)·mgDW^{-1}·min^{-1}]	$1.77e^{+3} \pm 10.50$	$3.64e^{+3} \pm 25$	$1.15e^{+5} \pm 834$
Km_{pop} [AU (count)·mgDW^{-1}]	$4.28e^{+3} \pm 75.1$	$7.01e^{+3} \pm 108$	$1.7e^{+5} \pm 2.19e^{+3}$
Catalytic efficiency [min^{-1}]	0.41	0.52	0.68
$m(to)_{pop}$ [AU (count)]	10.9 ± 0.24	8.57 ± 0.20	426 ± 13.9
τ_{pop} [min]	$0.33 \pm 1.66e^{-3}$	$0.52 \pm 1.7e^{-3}$	$0.74 \pm 2.1e^{-3}$
ω_{Km} [AU (count)·mgDW^{-1}]	$0.80 \pm 1.37e^{-3}$	$\sigma_{Km} = 0.71 \pm 1.19e^{-3}$ $\gamma_{Km} = 0.16 \pm 1.88e^{-4}$	$0.50 \pm 9.63e^{-4}$
$\omega_{m(to)}$ [AU (count)]	$1.17 \pm 1.58e^{-3}$	$\sigma_{m(to)} = 1.25 \pm 1.66e^{-3}$ $\gamma_{m(to)} = 0.55 \pm 2.87e^{-4}$	$1.72 \pm 2.33e^{-3}$
ω_{τ} [min]	$1.65 \pm 4.90e^{-3}$	$\sigma_{\tau} = 1.50 \pm 3.1e^{-3}$ $\gamma_{\tau} = 0.14 \pm 6.3e^{-4}$	$1.22 \pm 2.46e^{-3}$
a [AU (count)]	$0.84 \pm 5e^{-4}$	$0.24 \pm 8.75e^{-5}$	$3.82 \pm 7.26e^{-3}$
b [-]	$0.06 \pm 5.61e^{-5}$	$0.15 \pm 3.79e^{-5}$	$0.24 \pm 1.55e^{-4}$
Condition number	15.0	10.6	5.9

related to the binding efficiency of microarray probes for instance, or bias related to the library size, transcript length, and GC content in the case of RNA-Seq data. Data preprocessing allows to remove part of the systematic error, through background correction of probe intensities and normalisation to reduce the variability between measurements and time points in the case of dynamical data. We use a combined error model to take into account the technical errors of RNA-Seq and microarrays:

$$g(f(m_i(t_j), \Phi_i), \theta) = f(m_i(t_j), \Phi_i)) \times b + a) \tag{5}$$

where $\theta = \{a, b\}$. The multiplicative error parameter b is typically related to the mean of probe binding efficiencies and a is the basal measurement error (Meacham *et al.*, 2011; Weng *et al.*, 2006). These error parameters will be estimated together with the population parameters (Φ_{pop}, Σ, Γ, and Ω), from which, together with the data, are then derived the specific parameters for individual mRNAs.

3 Estimation of mechanistic model parameters from microarray data

3.1 Estimation from smoothed data

Can we estimate mechanistic model parameters from a single dynamical microarray dataset using NLME modelling? In this section, we address the question, by working on the most favourable scenario. We preprocess the background-corrected microarray probe intensities to take out part of the systematic measurement error. Among the possible approaches (Laguerre *et al.*, 2018; Oh and Li, 2021; Ritchie *et al.*, 2015), we have chosen spline smoothing to adjust for variability between the eight probes and the twelve time points for each mRNA i (Bates *et al.*, 2015; Ritchie *et al.*, 2015) (Supplementary Section S1.1). We obtain one degradation profile per mRNA and a total of 2809 smoothed degradation profiles to be used for estimation of the fixed and random effects of model parameters. The later are defined as follows: mRNAs being degraded at the same maximal rate, Vm is set to a fixed value, while Km, the delay before degradation τ, and initial concentrations at time t_0, $m(t_0)$, are allowed to vary between mRNAs. This gives:

$$\begin{cases} log(Vm_i) = & Vm_{pop} \\ log(\tau_i) = & log(\tau_{pop}) + \eta_i^{\tau} \\ log(m(t_0)_i) = & log(m(t_0)_{pop}) + \eta_i^{m(t_0)} \\ log(Km_i) = & log(Km_{pop}) + \eta_i^{Km} \end{cases}$$

with $\eta_i^{\phi} \sim \mathcal{N}(0, \sigma_{\phi}^2)$ for $\phi = \{\tau, m(t_0), Km\}$.

Estimation of NLME model parameters was run with the Monolix software in 4997 CPU seconds (see Supplementary Section S3 for details). As shown in Table 1 (second column), the population parameters are estimated with high certainty, with a small standard error for most of them and small error model parameters. The good estimation result reflects the size of the dataset (2809 degradation profiles) used for estimating parameter distributions. Our mechanistic model is sufficiently complex to explain the data, according to the normality of the error residuals with a mean of 0 (Supplementary Section S4.1). The Visual Predicted Check in

Supplementary Fig. S5 shows that model simulations with the estimated parameters are able to reproduce the central trend and variability of the degradation profiles. This is illustrated in Fig. 2(a) for mRNA *satP*. The predicted degradation profile using the estimated individual parameters correctly fits the data. This mRNA has a lower initial concentration at time t_0, and a slower degradation kinetics than the typical mRNA, representative of the population, whose degradation profile is shown as a black line.

Michaelis-Menten equations are known to pose practical identifiability problems, notably to obtain unique estimates of *Vm* and *Km*, while their ratio is identifiable (e.g., Choi *et al.* 2017; Holmberg 1982; Stroberg and Schnell 2016). In the NLME framework, the fact that individual parameters are tied by the population distribution can render otherwise intractable identification problems feasible (Lavielle and Aarons, 2016). Our population parameters seem to be practically identifiable, since they have low standard errors and the estimation algorithm converges to almost the same solution when started several times from random initial conditions (Supplementary Fig. S8). In addition, the condition number corresponding to the ratio of eigenvalues of the variance-covariance matrix of the estimates is satisfactory (Table 1). Correlations between parameters obtained from this matrix are shown in Fig. 2(d). All of them are statistically significant (p-value $< 2.2\cdot \times 10^{-6}$ with a t-test). Most likely, our parametrisation choice contributed to the small correlation of 0.35 between population parameters Vm_{pop} and Km_{pop} (with individuals set to a fixed *Vm* value, while Km_i varies), which can be seen in the multistart estimation results in Supplementary Fig. S8. The two parameters, not just their ratio, are informative. We observe correlations between other estimates, though. For instance, between $m(t_0)_{pop}$ and Km_{pop} (-0.66) and between Ω_τ and τ (-0.61). It may reflect residual identifiability issues (e.g., a bimodal distribution of τ_{pop} would have been more appropriate, see Supplementary Fig. S11 for η_i^τ), as well as a biological phenomenon: mRNA abundance is a known determinant of mRNA stability (Esquerré *et al.*, 2015; Nouaille *et al.*, 2017), which in our model is largely determined by the *Km* parameter. This may explain the correlation between $m(t_0)_{pop}$ and Km_{pop}. In the same line, mRNAs in *E. coli* are mainly regulated at the transcriptional level (Esquerré *et al.*, 2014). Abundant mRNAs are thus actively transcribed and on-going transcription at the time of rifampicin addition is likely to delay the onset of degradation, hence the small correlation of 0.33 between abundance and delay at the population level.

Are the estimated parameters biologically relevant? We cannot say for Vm_{pop} and Km_{pop} whose nonstandard units prevent their comparison to literature data. Their ratio in standard units, 0.41 min^{-1} in Table 1, represents the catalytic efficiency of degradosome and mRNA turnover. It indicates that the typical individual mRNA is degraded every 2.4 min, which agrees well with the median half life of 2.8 min determined for the same dataset in (Esquerré *et al.*, 2014). The τ_{pop} value in Table 1 corresponds to an average delay of about 20s before the onset of degradation, in agreement with the expected delay of rifampicin action <30s observed experimentally (e.g. Chen *et al.* 2015; Pato and von Meyenburg 1970). *E. coli* has many long mRNAs such as polycistronic mRNAs, which often delay the onset of degradation (Chen *et al.*, 2015). This most likely explains the relatively large value of the random effect for parameter τ ($\omega_\tau = 1.65$ min).

The largest population parameter for random effect is $\omega_{m(t_0)}$, which means that differences in mRNA concentrations explain a large part of the variability of the degradation profiles. *Km* values vary as well, but to a lesser extent. They reflect variations of the affinity of the degradosome for its target mRNAs and possible post-transcriptional regulations. This will be further studied in Section 5. We conclude from these results that NLME modelling is a suitable framework for mechanistic parameter estimation from dynamical transcriptomics data, at least when data is preprocessed. Despite the fact that only one dataset is used for estimation, the numerous individuals in the population share sufficient information to obtain good inference results.

3.2 Estimation from microarray probe intensities

In this section, we complexify the estimation problem. Since mRNA abundance is quantified by means of eight probes varying in their binding efficiency and mRNA complementarity, we have a population of 22,472 probes and as many degradation profiles, which could be useful information to improve parameter inference. However, probe intensities are background corrected but not further preprocessed. Can the NLME model handle richer, but noisier and larger data, at the price of an increased computational cost? Technical variability between probes can be seen for mRNA *satP* in Fig. 2(b) (black points). While the curvature of the eight degradation profiles is similar, probe intensities are variable at time t_0, all the more so due to the use of three measurements for each probe at this time point.

We introduce a new random effect in our NLME model, $\eta_{i,r}$, to describe technical variability between probes specific to a given mRNA i:

$$\begin{cases} log(Vm_i) = & Vm_{pop} \\ log(\tau i) = & log(\tau_{pop}) + \eta_i^\tau + \eta_{i,r}^\tau \\ log(m(t_0)_i) = & log(m(t_0)_{pop}) + \eta_i^{m(t_0)} + \eta_{i,r}^{m(t_0)} \\ log(Km_i) = & log(Km_{pop}) + \eta_i^{Km} + \eta_{i,r}^{Km} \end{cases}$$

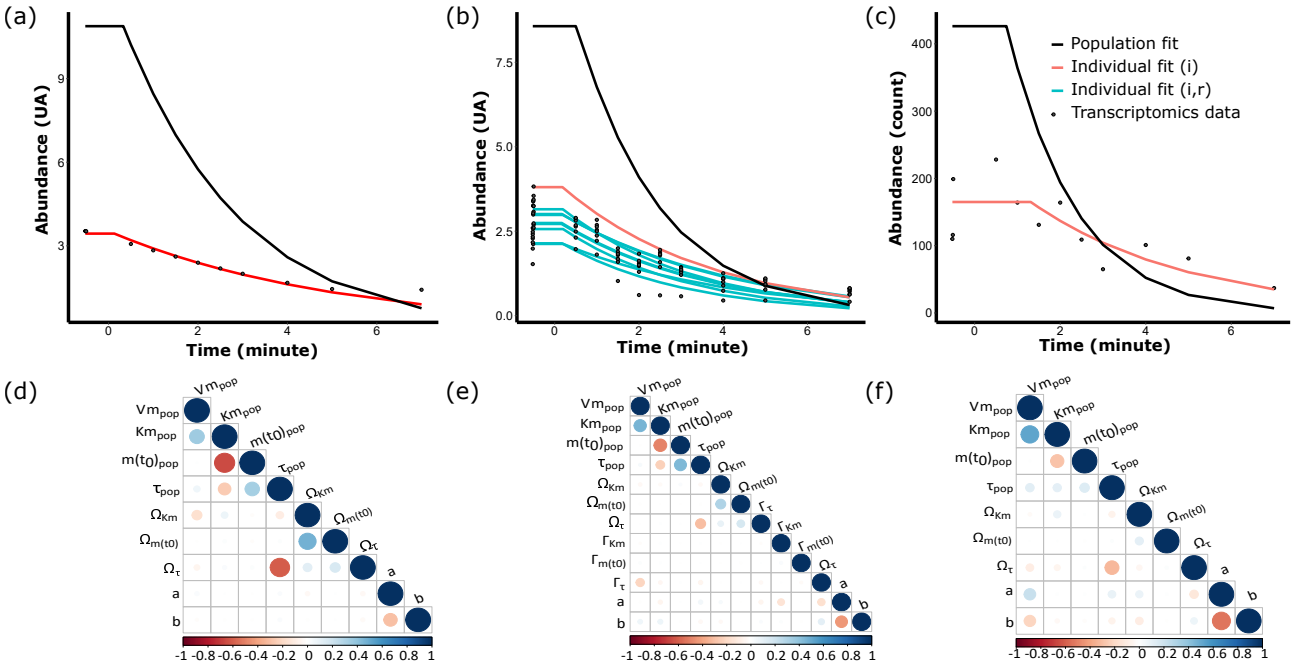

Figure 2: Estimation of mechanistic model parameters from time-series transcriptomics data. Top panels: examples of fit for mRNA *satP* obtained with (a) smoothed microarray data, (b) microarray probe intensities, and (c) normalised RNA-Seq data. Continuous lines represent the predicted degradation profiles generated with estimated population parameters (black) and with individual parameters for mRNAs (red) and probe intensities (cyan). Bottom panels: correlation plots of the estimated parameters for (d) smoothed microarray data, (e) microarray probe intensities, (f) RNA-Seq data. All correlations are statistically significant (p-value $< 2.2\cdot \times 10^{-6}$ with a t-test). They are given according to the colour code shown below each plot.

with $\eta_i^\phi \sim \mathcal{N}(0, \sigma_\phi^2)$ and $\eta_{i,r}^\phi \sim \mathcal{N}(0, \sigma_\phi^2)$ for $\phi = \{\tau, m(t_0), Km\}$.

As expected, the estimation problem is computationally more expensive, with a CPU time of $2.4{\cdot}10^6$s. Accurate and robust estimation results were obtained with the microarray probe intensities, as shown in Supplementary Information S4 for the residual distributions, VPC, and the convergence of the multistart algorithm to the same solution. The condition number of the variance-covariance matrix of the estimates is smaller when probe intensities are used (Table 1) and correlations between population parameters are markedly reduced too (Fig. 2(e).

The γ parameter values in Table 1 (third column), indicative of the technical variability between probe intensities, remains smaller than the variability between mRNAs, which is biological variability. The largest random effect between probes and mRNAs, observed for $\gamma_{m(t_0)}$, reflects the important noise of probe intensities at time $t0$. Some of them have clearly a weaker binding efficiency, leading to an estimated initial concentration $m(t_0)$ below that obtained with smoothed data. Interestingly, the standard deviation of random effects between mRNAs, σ_{Km}, $\sigma_{m(t_0)}$, and σ_τ, which corresponds to biological variability, is similar in both cases. They lead to consistent random effects between mRNAs for both datasets (Supplementary Fig. S11). For instance, mRNAs with weak affinities have high Km random effects in both datasets, and conversely. In the case of the random effects for initial concentrations, strong consistency between estimates from smoothed and nonsmoothed data is achieved when considering the two levels of variability (Supplementary Fig. S12), which indicates that information is gained when explicitly accounting for fluctuations between probes in the model.

The only changes, concerning error model parameters, are indicative of data preprocessing being performed within the NLME framework. The absence of data smoothing is reflected in the doubling of parameter b value, associated with binding efficiency bias. To the contrary, explicit consideration of the technical variability has decreased the basal measurement error represented by parameter a. The parameter value might represent residual systematic bias related to acquisition and quantification of array images for instance, or a systematic hybridisation bias. We note a small divergence between the Vm_{pop} and Km_{pop} values obtained from smoothed and nonsmoothed data. We cannot exclude that estimates from smoothed data were biased by the preprocessing step performed outside of the NLME framework in Section 3.1.

Overall, these results suggest that mechanistic parameter estimation from noisy and large microarray data is possible with the NLME framework. It allows accurate statistical inference from (almost) raw data through a good separation of the technical and biological variability.

4 Estimation of mechanistic model parameters from RNA-Seq data

RNA sequencing has become a popular approach for the quantification of gene expression. It provides a higher coverage and greater resolution than microarrays and former DNA sequencing approaches, and alleviate hybridisation bias (Marguerat and Bähler, 2010; Hrdlickova *et al.*, 2017; Todd *et al.*, 2016). Thus, with the wealth of RNA-Seq data being generated, it is interesting to inquire whether they are also prone to mechanistic interpretation. The difficulty is that we have only one noisy time series per mRNA for parameter estimation, as illustrated in Fig. 2(c) for mRNA *satP*.

NLME parametrisation and error modelling are the same as described in Section 3.1. There are no technical replicates, thus technical random effect has been omitted, and we consider only one level of variability, between mRNAs. Population parameters are estimated accurately, with a low standard error (Table 1, fourth column). The mechanistic model is able to explain count data and describe their variability, and the estimation results are robust, as shown in Supplementary Section S4. RNA-Seq and microarray units are different, which limits the comparison of estimation results. The Vm_{pop} and Km_{pop} values lead to a somewhat higher, but similar mRNA turnover, in the range of those measured experimentally (Table 1) (Bernstein *et al.*, 2002; Esquerré *et al.*, 2014). The value of the average delay, τ_{pop}, remains close to the experimental observations (Chen *et al.*, 2015; Pato and von Meyenburg, 1970).

In the absence of technical random effect, part of the technical variability is accounted for by the error model parameters, which are higher (Table 1). While we have opted for the same combined error model for microarray and RNA-Seq data for the sake of comparison, the correlation between a and b values indicates that a simplified model would be more appropriate (-0.53, Fig. 2(f)). Apart from Vm_{pop} and Km_{pop} that are more correlated in this estimation (0.52), other correlations are weak and the condition number of the variance-covariance matrix is the smallest of the three estimations (Table 1). Hence, even in the absence of technical replicate, the estimated biological random effects are consistent with that obtained from another type of data (Supplementary Fig. S11).

Altogether, inference of mechanistic model parameters from RNA-Seq data is also possible with the NLME framework, despite data noise and the absence of technical replicate. Mechanistic models provide physiologically-interpretable parameters, which will be analysed below. In the context of NLME models, in particular, we would like to mechanistically interpret the random effects, because they represent the effect of underlying regulatory mechanisms.

5 Identification of regulatory mechanisms of mRNA degradation

Classically, transcriptomics data are used in differential expression analyses, conducted on the fold change of expression levels characterising gene expression variability between conditions. Our NLME framework gives us access to random effects that represent individual variations from the mean, mode, or median. These can be interpreted as fold changes and can be used directly in Gene Set Enrichment Analysis (GSEA) (Yu *et al.*, 2012).

Here we will perform GSEA on one specific model parameter, the Michaelis constant. The random effect associated with this parameter, η_i^{Km}, subsums fluctuations of mRNA sequence and structure characteristics affecting the binding of the degradation machinery, as well as the effect of regulatory factors facilitating or inhibiting this binding (see Supplementary Eq. S2). Activatory effects decrease the value of individual parameter Km_i and thus, the corresponding random effect η_i^{Km} can take a negative value, while inhibitions increasing the Km lead to higher, positive random effects. Non-regulated mRNAs are expected to have small random effects, close to zero. They are already well described by population parameters.

Fig. 3 shows the results of a GSEA conducted on the parameters estimated from the smoothed microarray dataset. It reproduces the η_i^{Km} distributions of mRNAs in the GSEA enriched Gene Ontology (GO) categories. One category, *organic acid processes* related to carbon metabolism, is strongly post-transcriptionally regulated. The category includes genes involved in glycolysis and the Krebs cycle, whose half life is known to be controlled at high growth rate (Esquerré *et al.* 2014; see Roux *et al.* 2021 for review). It is interesting to note that mRNAs involved in the *organic acid catabolic process* category tend to be stabilised. This agrees with the stabilisation of mRNAs involved in the COG category *Carbohydrate transport and metabolism* in (Esquerré *et al.*, 2015). It may reflect the need to accumulate more of the corresponding enzymes to sustain high growth rate on glucose, which is the experimental condition studied. On the contrary, the negative random effect of Km for mRNAs in *organic acid biosynthetic processes* indicates that they are less needed in these growth conditions. These processes are involved in fast adaptation to fluctuating environments. Alleviating this post-transcriptional control can trigger a quicker cellular response through rapid accumulation of the corresponding enzymes, as mRNA instability and the relatively low cost of transcription compared to protein synthesis are determinant for fast changes of gene expression (Pérez-Ortín *et al.*, 2019). The five other categories shown in Fig. 3, encompassing amino-acid biosynthesis, cell motility, and translation, are known to include numerous post-transcriptionally regulated genes (Modi *et al.*, 2011; Wagner and Romby, 2015). According to our study, the degradation of these genes is stimulated, although it can be compensated for by high transcription rates. For instance, genes in the

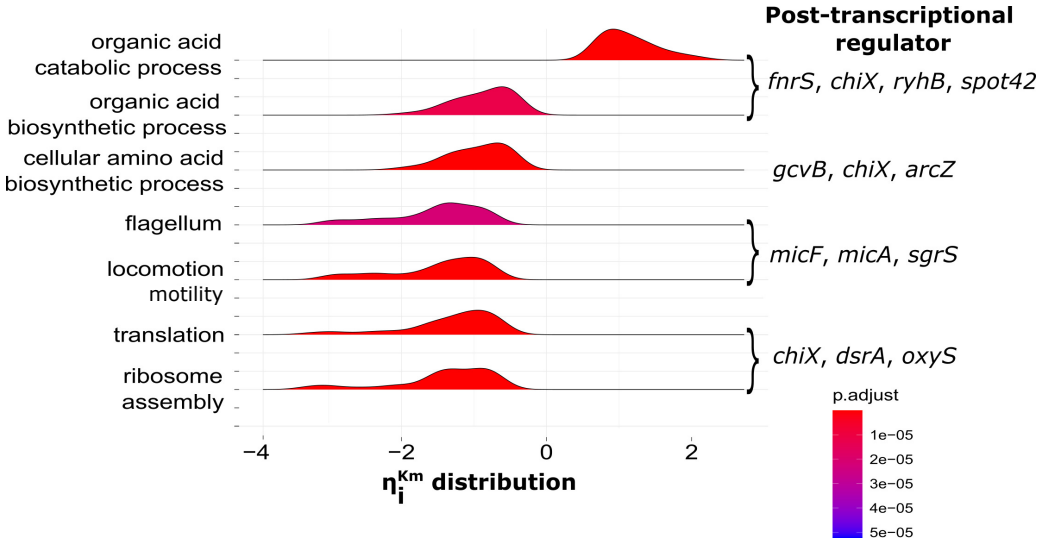

Figure 3: Gene set enrichment analysis conducted on random effects η_i^{Km} for the smoothed microarray data (Supplementary Section S6). Distributions are coloured according to their adjusted p-value

translation and *ribosome assembly* categories, notably genes coding for ribosomal proteins, are known to be actively transcribed at high growth rate, in particular in these experimental conditions (Esquerré *et al.*, 2014). In our study, this is reflected by the high random effect obtained for the initial concentrations of these mRNAs.

The above results show that our method is able to retrieve known post-transcriptionally regulated genes. We can now make hypotheses on the regulatory mechanisms involved. Half of the genes in each category shown in Fig. 3 are known and/or predicted targets of small RNAs (Modi *et al.*, 2011; Wright *et al.*, 2014). The main ones are cited on the right side of the figure. They facilitate or block degradosome binding to the target mRNA, acting alone or in complex with a RNA-binding protein such as Hfq or ProQ (for reviews, Holmqvist *et al.* 2020; Modi *et al.* 2011; Reyer *et al.* 2021; Santiago-Frangos and Woodson 2018). For instance, the small RNA *gcvB* is the master sRNA regulator of amino-acid metabolism and transport in a wide range of Gram-negative bacteria (Urbanowski *et al.* 2000; Miyakoshi *et al.* 2022 and references therein). It represses more than 50 genes in *E. coli* at the post-transcriptional level (Miyakoshi *et al.*, 2022), in agreement with the negative random effect of individual *Km* parameters observed for the category *cellular amino acid biosynthetic process* in Fig. 3.

Small RNAs have emerged as major post-transcriptional regulators that affect translation initiation and/or mRNA stability, both negatively and positively (Wagner and Romby, 2015). Yet their targets have not be all identified. For instance in Fig. 3, sole half of the genes in the different GO biological process categories have been identified as post-transcriptionally regulated. Our approach has thus allowed to identify new target genes.

6 Discussion and conclusion

The approach developed in this paper responds to the pressing need of mechanistically interpreting high-throughput biological data. To that aim, we have extended the application of NLME models to analyse "mRNA-to-mRNA" variability in a population of cells, by considering that mRNAs form a joint population submitted to the same cellular machinery, degradosome in our study. This allows to infer multidimensional parameter distributions describing the population from dynamical transcriptomics data, and then to derive specific parameters for individual mRNAs. We have shown that the framework allows consistent interpretation of heterogeneous datasets with different levels of technical noise. It factors out technical variability and incorporates mechanistic details, thus allowing to leverage biological variability for the identification of regulatory mechanisms and their targets.

Such mechanistic interpretation of dynamical omics data is still rare in the literature, because statistical inference of nonlinear models is challenging. One notable exception is the estimation of a model of cancer-related signalling pathways from exome- and transcriptome-sequencing data (Fröhlich *et al.*, 2018). The 4100 unknown parameters of this large ODE model have been accurately estimated from data comprising more than 6,900 experimental conditions. Imposing a parameter distribution over the population in our study leverages the limited number of experimental observations despite a larger number of unknown parameters. This is a remarkable result, as this implies that reliable estimation from less comprehensive high-throughput datasets is possible.

This study provides the first example of the same NLME framework applied to different types of data, and to such a large amount of longitudinal data. We have shown that the framework is flexible and can be

used to analyse data of different types, with different levels of technical noise, from raw RNA-Seq data to nonsmoothed and smoothed microarray probe intensities. The framework is flexible and can be extended to other case studies, with different data, models, conditions, organisms, in single cells or populations of cells... provided adaptation of the model for the source of variability and of the residual error model. In the context of mRNA degradation, the possibility to analyse a large number of data is a clear improvement over classical studies, wherein determination of reliable mRNA half lives is often restricted to about 20% of cellular mRNAs (e.g., Chen *et al.* 2015). Useful information is lost, which makes it difficult to draw general conclusions from a subset of data.

There is an advantage to integrate additional data, for instance to improve parameter identifiability. In our study, the addition of literature information such as the value of the maximal velocity could solve the small discrepancies observed for Vm_{pop} and Km_{pop} (Section 3). However, integration of heterogeneous datasets is easier if they are expressed in compatible units. This would require an absolute quantification of transcriptome to obtain data in standard international units.

Current limitation to the integration of larger datasets is essentially computational. Multiplying by eight the dimension of the data and adding one level of variability in Section 3.2 has increased by ~ 50 the CPU time. The problem arises when the dependency structure of the individual parameters is complex. Efforts to overcome this limitation have started, notably with the development of f-SAEM, a fast mixing sampler for the maximum likelihood estimation of NLME population parameters (Karimi *et al.*, 2020), or of the platform Pumas with parallelization capabilities for simulation and estimation (Rackauckas *et al.*, 2022). Another means of addressing this issue is data reduction, for instance through nonparametric smoothing of aggregated individuals with similar gene expression profiles (Lu *et al.*, 2011).

Acknowledgements

The authors thank Inria MICROCOSME team members for numerous exchanges on this work. Aline Marguet (MICROCOSME) and Adeline Leclercq Samson (Univ. Grenoble Alpes) are more specifically thanked for discussions on mixed-effect modelling.

Funding

This work was supported by the French National Research Agency under project RIB-ECO [ANR-18-CE43-0010] and MEMIP [ANR-16-CE33-0018], as well as by the Institut National de Recherche pour l'Agriculture, l'alimentation et l'environnement (INRAE) and Inria, from which T.A.E. received a doctoral fellowship.

References

Almquist, J., Bendrioua, L., Adiels, C. B., Goksör, M., Hohmann, S., and Jirstrand, M. (2015). A nonlinear mixed effects approach for modeling the cell-to-cell variability of Mig1 dynamics in yeast. *PloS one*, **10**(4), e0124050.

Androulakis, I., Yang, E., and Almon, R. (2007). Analysis of time-series gene expression data: Methods, challenges, and opportunities. *Annu. Rev. Biomed. Eng.*, **9**(1), 205–228.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.*, **67**(1), 1–48.

Bernstein, J. A., Khodursky, A. B., Lin, P.-H., Lin-Chao, S., and Cohen, S. N. (2002). Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. *Proc. Natl. Acad. Sci. U.S.A.*, **99**(15), 9697–9702.

Carpousis, A. J. (2007). The RNA degradosome of *Escherichia coli*: an mRNA-degrading machine assembled on RNase E. *Annu. Rev. Microbiol.*, **61**, 71–87.

Chen, H., Shiroguchi, K., Ge, H., and Xie, X. S. (2015). Genome-wide study of mRNA degradation and transcript elongation in *Escherichia coli*. *Mol. Syst. Biol.*, **11**(1), 781.

Choi, B., Rempala, G. A., and Kim, J. K. (2017). Beyond the Michaelis-Menten equation: Accurate and efficient estimation of enzyme kinetic parameters. *Sci. Rep.*, **7**(1), 1–11.

Davidian, M. and Giltinan, D. M. (2003). Nonlinear models for repeated measurement data: an overview and update. *J. Agric. Biol. Environ. Stat.*, **8**(4), 387.

Dharmarajan, L., Kaltenbach, H.-M., Rudolf, F., and Stelling, J. (2019). A simple and flexible computational framework for inferring sources of heterogeneity from single-cell dynamics. *Cell Syst.*, **8**(1), 15–26.

Ding, J. and Bar-Joseph, Z. (2020). Analysis of time-series regulatory networks. *Curr. Opin. Syst. Biol.*, **21**, 16–24.

Dressaire, C., Picard, F., Redon, E., Loubière, P., Queinnec, I., Girbal, L., and Cocaign-Bousquet, M. (2013). Role of mRNA stability during bacterial adaptation. *PloS one*, **8**(3), e59059.

Dressaire, C., Pobre, V., Laguerre, S., Girbal, L., Arraiano, C. M., and Cocaign-Bousquet, M. (2018). PNPase is involved in the coordination of mRNA degradation and expression in stationary phase cells of *Escherichia coli*. *BMC Genom.*, **19**(1), 848.

Esquerré, T., Laguerre, S., Turlan, C., Carpousis, A. J., Girbal, L., and Cocaign-Bousquet, M. (2014). Dual role of transcription and transcript stability in the regulation of gene expression in *Escherichia coli* cells cultured on glucose at different growth rates. *Nucleic Acids Res*, **42**(4), 2460–2472.

Esquerré, T., Moisan, A., Chiapello, H., Arike, L., Vilu, R., Gaspin, C., Cocaign-Bousquet, M., and Girbal, L. (2015). Genome-wide investigation of mRNA lifetime determinants in *Escherichia coli* cells cultured at different growth rates. *BMC Genomics*, **16**(1), 275.

Etienne, T. A., Cocaign-Bousquet, M., and Ropers, D. (2020). Competitive effects in bacterial mRNA decay. *J. Theor. Biol.*, page 110333.

Fröhlich, F., Kessler, T., Weindl, D., Shadrin, A., Schmiester, L., Hache, H., Muradyan, A., Schütte, M., Lim, J.-H., Heinig, M., *et al.* (2018). Efficient parameter estimation enables the prediction of drug response using a mechanistic pan-cancer pathway model. *Cell Syst.*, **7**(6), 567–579.

González-Vargas, A. M., Cinquemani, E., and Ferrari-Trecate, G. (2016). Validation methods for population models of gene expression dynamics. *IFAC-PapersOnLine*, **49**(26), 114–119.

Grigorov, M. G. (2011). Analysis of time course omics datasets. In *Bioinformatics for Omics Data*, pages 153–172. Springer.

Holmberg, A. (1982). On the practical identifiability of microbial growth models incorporating michaelis-menten type nonlinearities. *Math. Biosci.*, **62**(1), 23–43.

Holmqvist, E., Berggren, S., and Rizvanovic, A. (2020). RNA-binding activity and regulatory functions of the emerging sRNA-binding protein ProQ. *Biochim. Biophys. Acta. Gene Regul. Mech.*, **1863**(9), 194596.

Hrdlickova, R., Toloue, M., and Tian, B. (2017). RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip. Rev. RNA*, **8**(1), e1364.

Karimi, B., Lavielle, M., and Moulines, E. (2020). f-SAEM: A fast stochastic approximation of the EM algorithm for nonlinear mixed effects models. *Comput. Stat. Data Anal.*, **141**, 123–138.

Karlsson, M. O. and Holford, N. (2008). A tutorial on visual predictive checks. In *abstr*, volume 1434, page 17.

Laguerre, S., González, I., Nouaille, S., Moisan, A., Villa-Vialaneix, N., Gaspin, C., Bouvier, M., Carpousis, A. J., Cocaign-Bousquet, M., and Girbal, L. (2018). Large-scale measurement of mRNA degradation in *Escherichia coli*: To delay or not to delay. *Methods Enzymol.*, **612**, 47–66.

Lavielle, M. (2014). *Mixed effects models for the population approach: models, tasks, methods and tools.* CRC press.

Lavielle, M. and Aarons, L. (2016). What do we mean by identifiability in mixed effects models? *J. Pharmacokinet. Pharmacodyn.*, **43**(1), 111–122.

Lavielle, M. and Mbogning, C. (2014). An improved SAEM algorithm for maximum likelihood estimation in mixtures of non linear mixed effects models. *Stat. Comput.*, **24**(5), 693–707.

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, pages 673–687.

Liu, X., Li, N., Liu, S., Wang, J., Zhang, N., Zheng, X., Leung, K.-S., and Cheng, L. (2019). Normalization methods for the analysis of unbalanced transcriptome data: a review. *Front. Bioeng. Biotechnol.*, **7**.

Llamosi, A., Gonzalez-Vargas, A., Versari, C., Cinquemani, E., Ferrari-Trecate, G., Hersen, P., and Batt, G. (2016). What Population Reveals about Individual Cell Identity: Single-Cell Parameter Estimation of Models of Gene Expression in Yeast. *PLoS Comp. Biol.*, **12**(2), e1004706.

Loos, C., Moeller, K., Fröhlich, F., Hucho, T., and Hasenauer, J. (2018). A hierarchical, data-driven approach to modeling single-cell populations predicts latent causes of cell-to-cell variability. *Cell Syst.*, **6**(5), 593–603.

Lu, T., Liang, H., Li, H., and Wu, H. (2011). High-dimensional ODEs coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification. *J. Am. Stat. Assoc.*, **106**(496), 1242–1258.

Marguerat, S. and Bähler, J. (2010). RNA-seq: from technology to biology. *Cell. Mol. Life Sci.*, **67**(4), 569–579.

Meacham, F., Boffelli, D., Dhahbi, J., Martin, D. I., Singer, M., and Pachter, L. (2011). Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinform.*, **12**(1), 1–11.

Miyakoshi, M., Okayama, H., Lejars, M., Kanda, T., Tanaka, Y., Itaya, K., Okuno, M., Itoh, T., Iwai, N., and Wachi, M. (2022). Mining RNA-seq data reveals the massive regulon of GcvB small RNA and its physiological significance in maintaining amino acid homeostasis in *Escherichia coli*. *Mol. Microbiol.*, **117**(1), 160–178.

Modi, S. R., Camacho, D. M., Kohanski, M. A., Walker, G. C., and Collins, J. J. (2011). Functional characterization of bacterial sRNAs using a network biology approach. *Proc. Natl. Acad. Sci. U.S.A.*, **108**(37), 15522–15527.

Moffitt, J. R., Pandey, S., Boettiger, A. N., Wang, S., and Zhuang, X. (2016). Spatial organization shapes the turnover of a bacterial transcriptome. *Elife*, **5**, e13065.

Noor, E., Cherkaoui, S., and Sauer, U. (2019). Biological insights through omics data integration. *Curr. Opin. Syst. Biol.*, **15**, 39–47.

Nouaille, S., Mondeil, S., Finoux, A.-L., Moulis, C., Girbal, L., and Cocaign-Bousquet, M. (2017). The stability of an mRNA is influenced by its concentration: a potential physical mechanism to regulate gene expression. *Nucleic Acids Res*, **45**(20), 11711–11724.

Oddi, F. J., Miguez, F. E., Ghermandi, L., Bianchi, L. O., and Garibaldi, L. A. (2019). A nonlinear mixed-effects modeling approach for ecological data: Using temporal dynamics of vegetation moisture as an example. *Ecol. Evol.*, **9**(18), 10225–10240.

Oh, V.-K. S. and Li, R. W. (2021). Temporal dynamic methods for bulk RNA-Seq time series data. *Genes*, **12**(3), 352.

Panhard, X. and Samson, A. (2009). Extension of the SAEM algorithm for nonlinear mixed models with 2 levels of random effects. *Biostatistics*, **10**(1), 121–135.

Pato, M. L. and von Meyenburg, K. (1970). Residual rna synthesis in *Escherichia coli* after inhibition of initiation of transcription by rifampicin. In *Cold Spring Harb. Symp. Quant. Biol*, 35, page 479–504, USA. Cold Spring Harbor Laboratory Press.

Pérez-Ortín, J. E., Tordera, V., and Chávez, S. (2019). Homeostasis in the central dogma of molecular biology: the importance of mRNA instability. *RNA Biol.*, **16**(12), 1659–1666.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rackauckas, C., Ma, Y., Noack, A., Dixit, V., Mogensen, P. K., Elrod, C., Tarek, M., Byrne, S., Maddhashiya, S., Calderón, J. B. S., Hatherly, M., Nyberg, J., Gobburu, J. V., and Ivaturi, V. (2022). Accelerated predictive healthcare analytics with Pumas, a high performance pharmaceutical modeling and simulation platform. *bioRxiv*.

Reyer, M. A., Chennakesavalu, S., Heideman, E. M., Ma, X., Bujnowska, M., Hong, L., Dinner, A. R., Vanderpool, C. K., and Fei, J. (2021). Kinetic modeling reveals additional regulation at co-transcriptional level by post-transcriptional sRNA regulators. *Cell Rep.*, **36**(13), 109764.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**(7), e47–e47.

Roux, C., Etienne, T., Hajnsdorf, E., Ropers, D., Carpousis, A., Cocaign-Bousquet, M., and Girbal, L. (2021). The essential role of mRNA degradation in understanding and engineering *E. coli* metabolism. *Biotechnol. Adv.*, page 107805.

Santiago-Frangos, A. and Woodson, S. A. (2018). Hfq chaperone brings speed dating to bacterial sRNA. *RNA*, **9**(4), e1475.

Straube, J., Gorse, A.-D., of Excellence Team, P. C., Huang, B. E., and Lê Cao, K.-A. (2015). A linear mixed model spline framework for analysing time course 'omics' data. *PloS one*, **10**(8), e0134540.

Stroberg, W. and Schnell, S. (2016). On the estimation errors of KM and V from time-course experiments using the michaelis–menten equation. *Biophys. Chem.*, **219**, 17–27.

Todd, E. V., Black, M. A., and Gemmell, N. J. (2016). The power and promise of rna-seq in ecology and evolution. *Mol. Ecol.*, **25**(6), 1224–1241.

Urbanowski, M. L., Stauffer, L. T., and Stauffer, G. V. (2000). The *gcvB* gene encodes a small untranslated RNA involved in expression of the dipeptide and oligopeptide transport systems in *Escherichia coli*. *Mol. Microbiol.*, **37**(4), 856–868.

Wagner, E. G. H. and Romby, P. (2015). Small RNAs in bacteria and archaea: who they are, what they do, and how they do it. *Adv. Genet.*, **90**, 133–208.

Wang, W. and Yan, J. (2021). Shape-restricted regression splines with R package splines2. *Journal of Data Science*, **19**(3), 498–517.

Weng, L., Dai, H., Zhan, Y., He, Y., Stepaniants, S. B., and Bassett, D. E. (2006). Rosetta error model for gene expression analysis. *Bioinformatics*, **22**(9), 1111–1121.

Wright, P. R., Georg, J., Mann, M., Sorescu, D. A., Richter, A. S., Lott, S., Kleinkauf, R., Hess, W. R., and Backofen, R. (2014). CopraRNA and IntaRNA: predicting small RNA targets, networks and interaction domains. *Nucleic Acids Res*, **42**(W1), W119–W123.

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**(5), 284–287.

A nonlinear mixed-effects approach for the mechanistic interpretation of time-series transcriptomics data: Supplementary material

Thibault A. Etienne[1,2], Charlotte Roux[2], Eugenio Cinquemani[1], Laurence Girbal[2], Muriel Cocaign-Bousquet[2] and Delphine Ropers[1,*]

[1]Univ. Grenoble Alpes, Inria, 38000 Grenoble, France
[2]TBI, Université de Toulouse, CNRS, INRAE, INSA, Toulouse, France.

Contents

6.1 Data preprocessing

Transcriptomics data used in this study come from microarray and RNA-Seq experiments. Their preprocessing was carried out using the R software (R Core Team, 2021).

6.1.1 Microarray data

The dataset is published in (Esquerré *et al.*, 2014). It corresponds to *E. coli* K12 MG1655 cells grown exponentially on glucose at 0.63h^{-1} prior to addition of the transcriptional inhibitor rifampicin. Cultures were sampled before and after drug addition and total RNA was extracted. Two µg of total RNA were converted to cDNA and subsequently labelled and hybridised to *E. coli* gene expression arrays. Two probesets of eight different probes were used per gene. Averaging the signal between probesets provided us with the intensity of eight different probes per mRNA. Probe intensities were then corrected for background using random probesets (Esquerré *et al.*, 2014). Rather than applying an intrareplicate quantile normalisation to the intensities of the eight probes used per mRNA, as done in (Esquerré *et al.*, 2014), we corrected the variation between probes and between time points through spline smoothing (Wang and Yan, 2021). This allowed to obtain a single time series per gene. An example of spline smoothing is provided for mRNA *thrC* in Fig. 4.

An additional filtration step was performed to retain only 2809 genes sufficiently expressed to allow for the observation of the degradation kinetics. Probe intensity for these genes is at least twice the background at time 0.

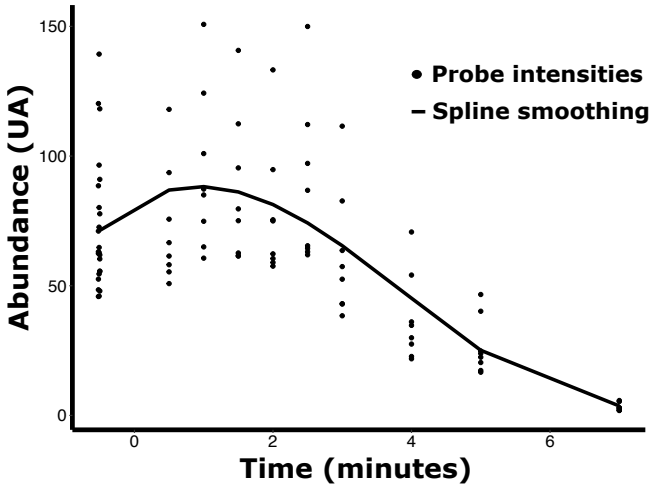

Figure 4: Monitoring of *thrC* mRNA degradation by microarrays. The black dots represent the background-corrected intensities of the eight probes at twelve time points, three time points before drug addition and nine after. The black line represents B-spline smoothing of the 8 probe intensities.

6.1.2 RNA-Seq data

The strain and the experimental conditions are the same as in Section 6.1.1. Total RNA (1 µg) was ribodepleted using the riboPOOLs kit and used to construct a sequencing library using the Ion Total RNA-Seq Kit v2 and the Ion Xpress RNA-Seq Barcode Kit. Reads were mapped with the `tmap` package onto the *E. coli* genome (version NC_000913.3, GenBank) and counted with `HTSeq-count` ("intersection non empty" mode) (Roux *et al.*, in preparation). We use directly the raw data of the 2809 genes selected in the microarray preprocessing step.

6.2 Mechanistic model of mRNA degradation

6.2.1 Simplification of mechanistic model of mRNA degradation

In (Etienne *et al.*, 2020), the following mechanistic model describes the time-dependent evolution of the concentration $c_i(t)$ of each mRNA i, averaged over a population of cells subjected to treatment by rifampicin:

$$\frac{dc_i(t)}{dt} = -\frac{k_{cat} \times E_0 \times c_i(t)}{Km_i^{app} + E_0 + c_i(t)}, \tag{6}$$

where k_{cat} is the catalytic constant in min^{-1}, E_0 the concentration of degradosome in μM, Km_i^{app} the apparent Michaelis constant of mRNA i in μM, and $c_i(t)$ the concentration of mRNA i in the same units. Km_i^{app} is said apparent because

it reflects the effect of activators or inhibitors on the kinetics of the degradation reaction. For instance:

$$Km_i^{app} = Km_i^t \left(1 + \frac{x_{small\ RNA}}{K_{small\ RNA}} + \frac{K_{regulatory\ protein}}{x_{regulatory\ protein}} + ...\right) , \tag{7}$$

where the x's denote the concentration of regulatory factors and the K's, their dissociation constant. Km_i^t is the "true" Michaelis constant of mRNA i, which solely depends on the mRNA features. In this example, the small RNA has an inhibitory effect that decreases the affinity of the degradosome for its target mRNA, while the regulatory protein stimulates degradation.

For the purpose of this paper, we do not need to describe the enzyme concentration explicitly. We introduce new parameters: Vm denotes the maximal velocity of the reaction $Vm = k_{cat} \times E_0$ (unit: µM·min^{-1}), while $Km_i = Km_i^{app} + E_0$ (unit: µM). This formulation leads to the classical Michaelis-Menten expression usually obtained after application of the standard quasi-steady-state assumption:

$$\frac{dc_i(t)}{dt} = -\frac{Vm \times c_i(t)}{Km_i + c_i(t)} . \tag{8}$$

This is our starting model, which will adapt below so as to integrate the experimental data (Section S2).

6.2.2 Adaptation of mechanistic model to data

Mechanistic model with quantities While model variables and parameters are expressed in the International System (SI) of units, this is not the case of the microarray and RNA-Seq data. In addition, the latter two correspond to abundances of mRNAs.

The integration of data into the model can be performed in two different manners: either we transform the data to obtain concentrations, possibly in SI units if experimental design has allowed the absolute quantification of mRNA abundances, or we adapt the model to describe the time-dependent evolution of quantities in their units. We have chosen the latter approach in this study. This allows to estimate model parameters and initial concentrations easily, without the requirement of additional data to allow for data conversion. The inconvenient is that the estimates will not be expressed in SI units, which impedes their comparison with literature data. However, the objective of the study is to determine how the variability of individual parameters account for the variability of degradation profiles. We do not need SI units to reach such conclusions.

Let $m_i(t)$ denotes the quantity of mRNA i at time t. It is expressed in Arbitrary Units (AU) (microarrays), or in read counts (RNA-Seq). $m_i(t)$ is proportional to the cellular abundance $n_i(t)$ of mRNA i at time t averaged over the population of cells: $m_i(t) = \alpha\, n_i(t)$. $n_i(t)$ is expressed in µmol, while α is a conversion factor expressed in AU/µmol (microarrays) or in count/µmol (RNA-Seq). The concentration $c_i(t)$ is proportional to the quantity of mRNA i:

$$c_i(t) = \frac{m_i(t)}{V} , \tag{9}$$

$$= \frac{\alpha\, n_i(t)}{V} , \tag{10}$$

where V is the cell volume expressed in Liter (L). It is constant as rifampicin addition stops cell growth.

If we know α or if we measure the absolute abundance of mRNA i, $n_i(t)$, we can reformulate Eq. 8 as a function of $n_i(t)$. As this is not the case here, we replace $c_i(t)$ by its expression (9) in Eq. 8:

$$\frac{dm_i(t)}{dt} = -\frac{Vm \times m_i(t)}{Km_i + \frac{m_i(t)}{V}} . \tag{11}$$

Explicit consideration of cell volume Analysing the degradation kinetics in various conditions may involve a change of volume between conditions, which implies to take into account the volume explicitly, as done in Eq. 11. Such information is not readily available in most cases, but alternative data can be used in place. In (Esquerré *et al.*, 2014), for instance, the multiplication of the microarray probe intensities in AU, measured from a starting material of 2 µg of total RNAs, by the total RNA extraction yield (in µg total RNA per mg of dry cell weight) provides mRNA concentration values in UA/mgDW. The extraction yield E_y is inversely proportional to cell volume and thus can be used as proxy for the cell volume. It has been measured in the experimental conditions studied here: $E_y = 45.2$ µg/mgDW in *E. coli* cells growing exponentially at 0.63 h^{-1} (Esquerré *et al.*, 2014), which gives $V = 0.022$ mgDW. We will use this value in our study. This changes the units of parameters Vm and Km_i to AU (or counts)/mgDW/min and AU (or counts)/mgDW, respectively.

6.3 Parameter estimation procedure

6.3.1 Initialisation of parameter values

Our data in AU or counts does not permit to use available literature values for initialising the unknown parameters. We used instead the mean of the degradation constants in (Esquerré *et al.*, 2014), which is similar to the mRNA turnover or catalytic efficiency of degradation given by the ratio $Vm/Km \approx 0.5$, from which initial values were given for the two parameters: $Vm = 102$ AU or count/gDW/min and $Km = 180$ AU or count/gDW based on several preliminary estimation tests. Standard deviations were initialised to 1. Parameters of the residual error model, a and b, were initialised to 1 and 0.3, respectively.

6.3.2 Parameter estimation with NLME modelling

Kinetic parameters were estimated using the SAEM-MCMC algorithm implemented in Monolix (Monolix version 2020R1. Antony, France: Lixoft SAS, 2020) using default settings indicated in the software documentation for auto-criteria of convergence at https://monolix.lixoft.com/ (maximum/minimun of the iterations, autocriteria, *etc*).

Estimation of population parameters The estimation of the population parameters is the key task in nonlinear mixed effect (NLME) modelling. It was performed using the Stochastic Approximation Expectation-Maximization (SAEM) algorithm of Monolix. It is achieved in three steps:

1. The burn-in phase: it is the SAEM initialisation, wherein individual parameters are sampled from their conditional distribution by MCMC sampling using initial values of the population parameters. The number of iterations of this phase was set to five iterations.
2. The Exploratory phase, during which the parameter value at iteration k is built using information collected at that iteration only (it does not take into account the value of the parameter at the previous iteration). We set 500 iterations for the RNA-Seq and smoothed microarray data, and 1000 iterations for the nonsmoothed microarray data.
3. The smoothing phase. During this step the parameter value at iteration k is built using information collected at the previous iteration. We fixed 200 iterations for both datasets.

Estimation of individual parameters The estimation algorithm for population parameters gives a rough estimate of the individual parameters already. However, it can be estimated by two more precise estimators: the conditional mode and the conditional distribution. We used them both.

Estimation of standard errors The standard errors represent the uncertainty of the estimated population parameters. They were determined in Monolix through the calculation of the Fisher Information Matrix using stochastic approximation.

6.4 Assessing the quality of estimation

6.4.1 Distribution of residuals

Different types of residuals can be determined to detect potential misspecifications in the mechanistic model or the residual error model: the PWRES (Population Weighted RESiduals), the IWRES (Individual Weighted RESiduals), and the NPDEs (Normalised Prediction Distribution Errors). The distributions of the residuals are shown in Fig. 5 for the smoothed microarray data, Fig. 6 for the nonsmoothed microarray data, and Fig. 7 for the RNA-Seq data.

The following two tests indicate that the model correctly describe the experimental data, as residuals are symmetrically scattered around 0 and normally distributed (Shapiro-Wilk test).

Symmetry test:

	Smoothed microarrays	Microarray probe intensities	RNA-Seq
IWRES	$< 2.2e^{-16}$	$< 2.2e^{-16}$	$1.79e^{-3}$
PWRES	$< 2.2e^{-16}$	$< 2.2e^{-16}$	$< 2.2e^{-16}$
NPDE	$< 2.2e^{-16}$	$< 2.2e^{-16}$	$8.54e^{-10}$

Shapiro-Wilk test:

	Smoothed microarrays	Microarray probe intensities	RNA-Seq
IWRES	$< 2.2e^{-16}$	$< 2.2e^{-16}$	$< 2.2e^{-16}$
PWRES	$< 2.2e^{-16}$	$< 2.2e^{-16}$	$< 2.2e^{-16}$
NPDE	$< 2.2e^{-16}$	$< 2.2e^{-16}$	$1.05e^{-15}$

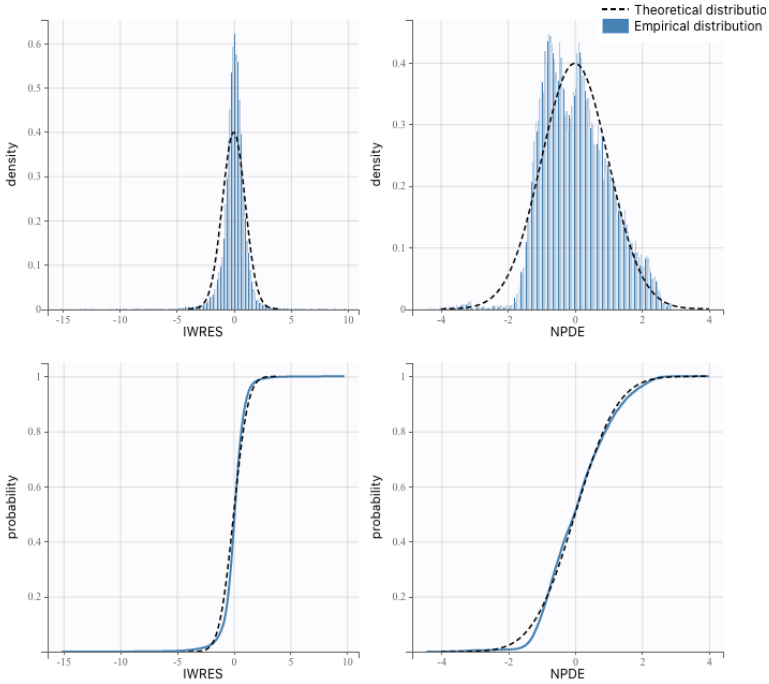

Figure 5: Residual distributions for smoothed microarray data

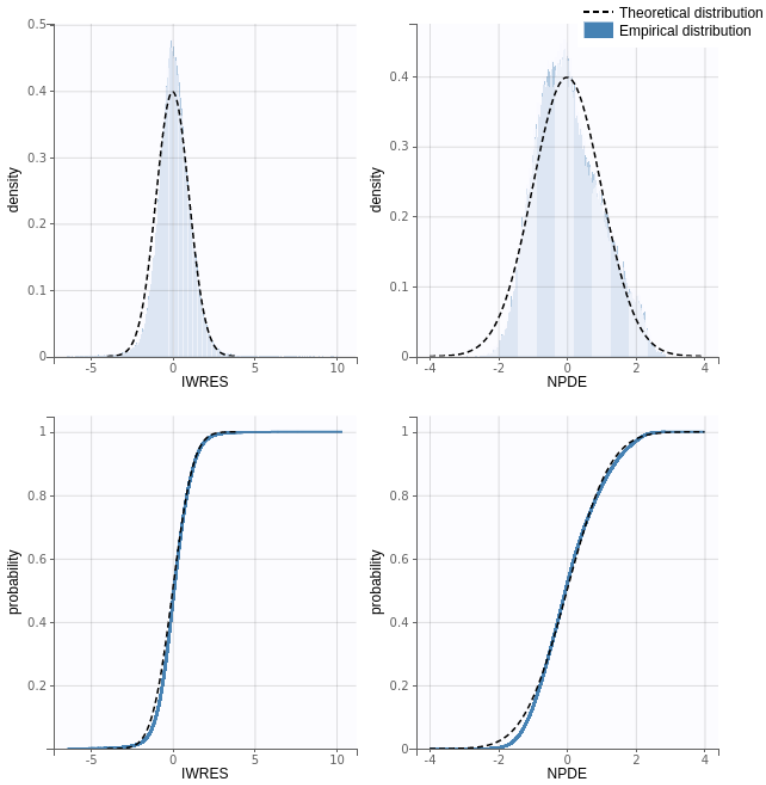

Figure 6: Residual distributions for microarray probe intensities

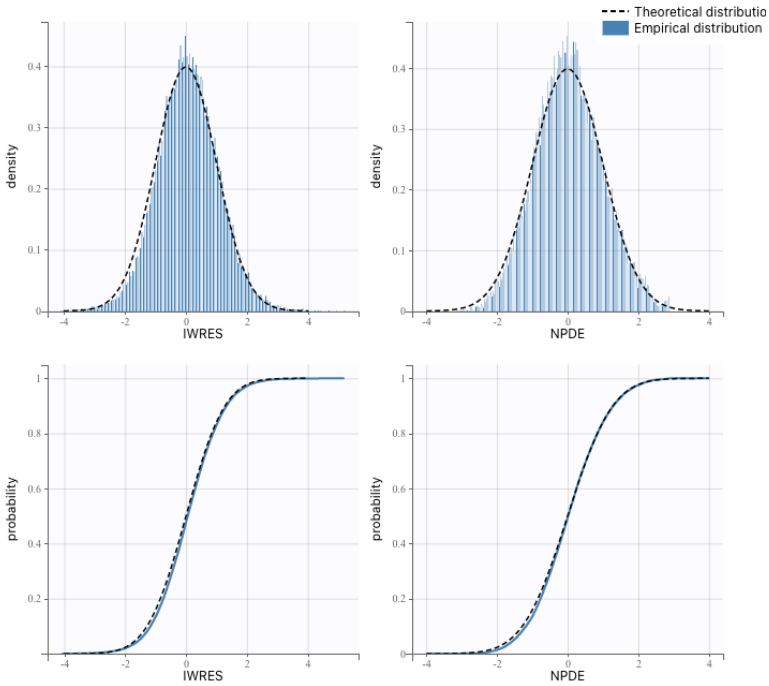

Figure 7: Residual distributions for RNA-Seq data

6.4.2 Visual Predictive Check

We use the classical Visual Predictive Check (VPC) tool to assess graphically whether simulations from the mechanistic model are able to reproduce both the central trend and the variability in the experimental data along time. It summarises, in the same graphic, the structural and statistical models by computing several quantiles of the empirical distribution of the 2809 degradation profiles, after having regrouped them into bins over successive intervals (Karlsson and Holford, 2008). The VPC are shown in Figs. 8, 9, and 10 for the estimations from the (smoothed and nonsmoothed) microarray data and the RNA-Seq data. Outliers (in red) are observed, due to the strong variability of degradation profiles. In this specific case, stratification of the VPCs could reduce the number of outliers, by splitting the degradation profiles in two or more groups (Karlsson and Holford, 2008).

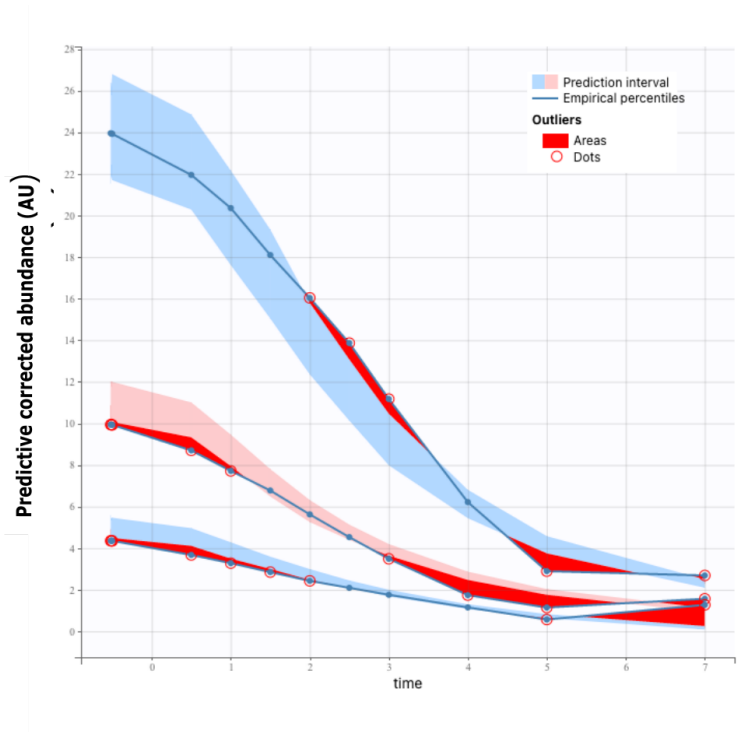

Figure 8: Visual predictive check for the smoothed microarray data

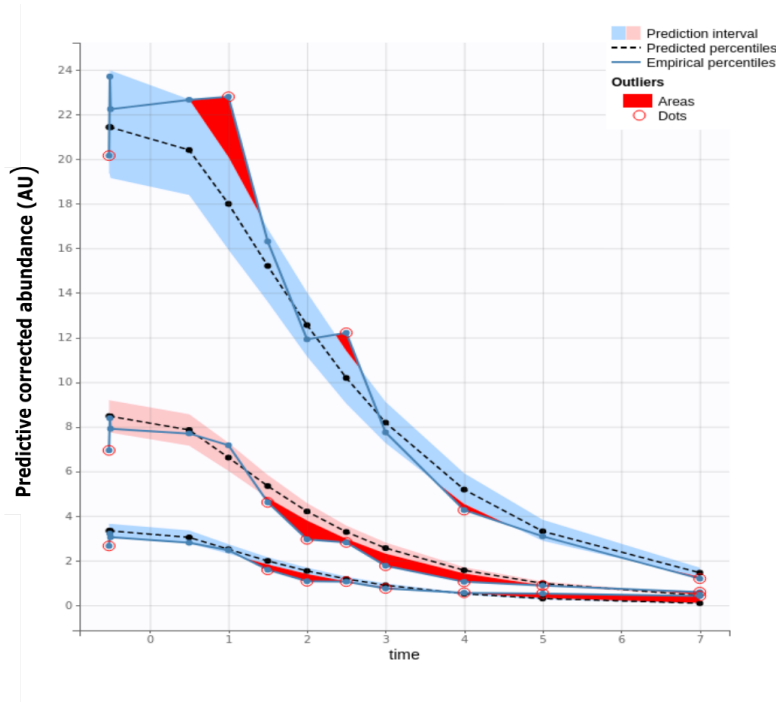

Figure 9: Visual predictive check for the microarray probe intensities

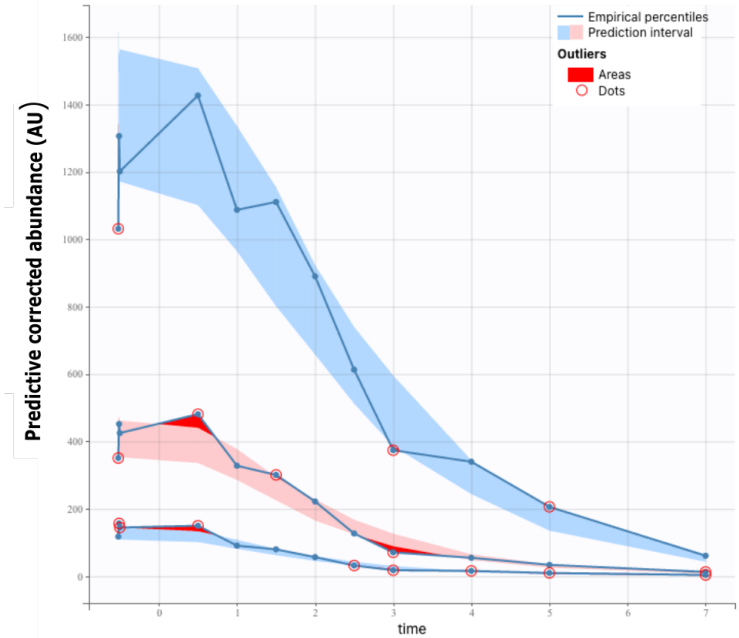

Figure 10: Visual predictive check for the RNA-Seq data

6.4.3 Correlation matrix of the estimates

The inverse of the Fisher Approximation Matrix can be used to detect possible practical identifiability issues. It provides an approximation of the variance-covariance matrix of the estimated population parameters. A large condition number (the ratio of the largest and smallest eigenvalues of the matrix) is indicative of strong correlations between estimates. The (small) value of the condition number in our case is shown in Table 1 of the main text for the three estimations, while correlations are graphically represented in Fig. 2 of the main text.

6.4.4 Convergence assessment

Convergence of the estimation algorithm is another approach that can be used to detect identifiability issues (Lavielle and Aarons, 2016). We started the algorithm from ten different initial conditions. Results are shown below in Figs. 11-13. The algorithm converges to (almost) the same solution for each of the three estimations.

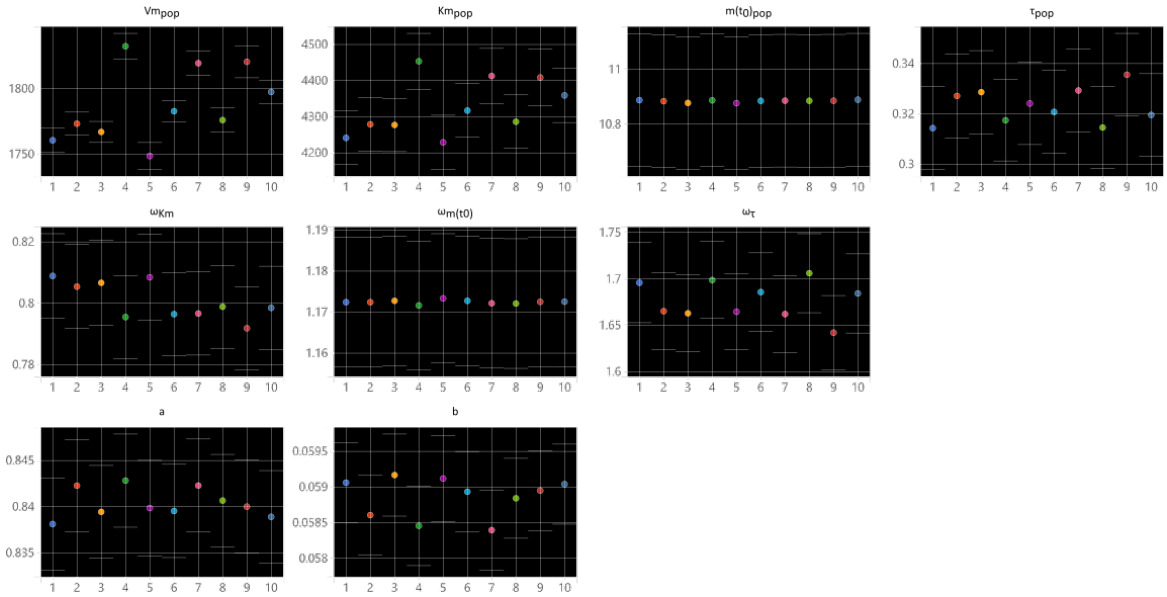

Figure 11: Estimated parameter values from 10 different initial conditions using the smoothed microarray data

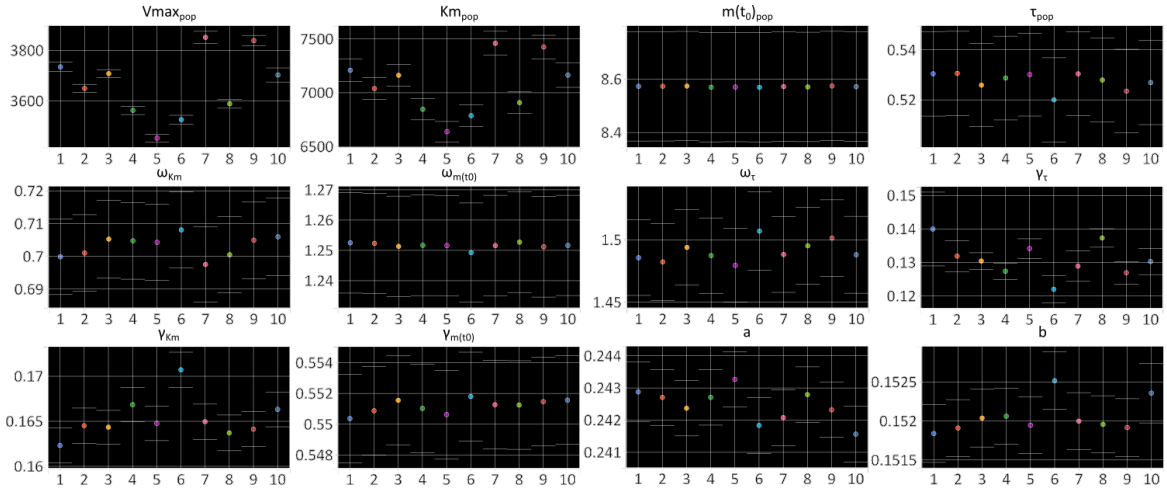

Figure 12: Estimated parameter values from 10 different initial conditions using the microarray probe intensities

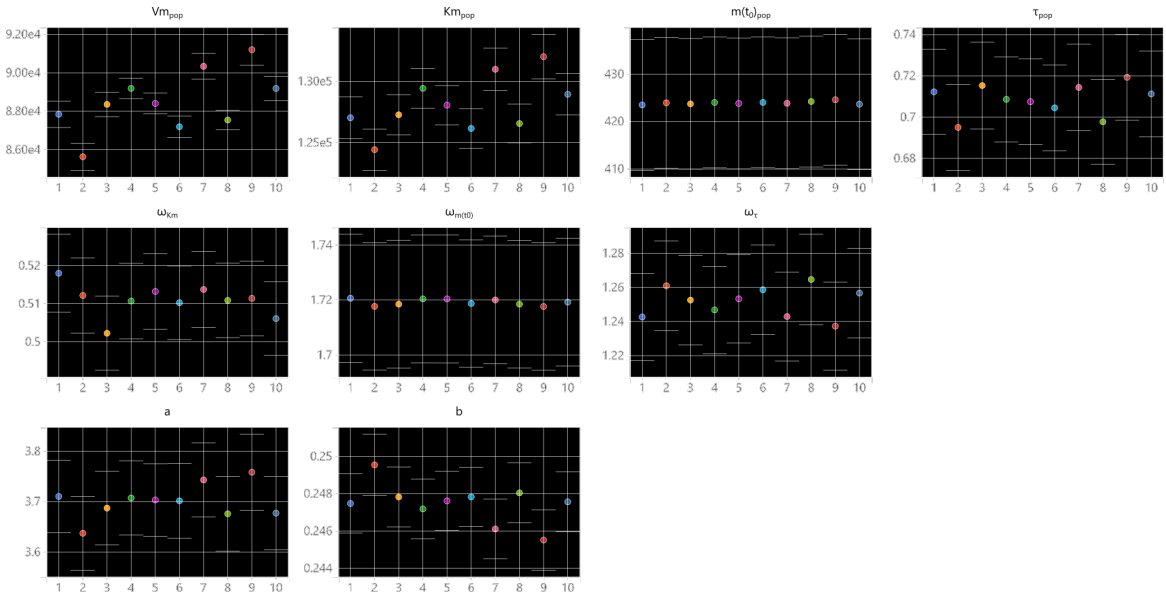

Figure 13: Estimated parameter values from 10 different initial conditions using the RNA-Seq data

6.4.5 Comparison of random effects estimated from the different data sets

Fig. 14 shows the consistency between random effects obtained from the different estimations. They are compared two by two. Lowest and highest random effects for the initial concentration are less comparable, in particular for the comparison of estimates from smoothed and nonsmoothed microarray data (middle left panel). This is solved by integrating the second level of variability describing fluctuations between probes (Fig. 14).

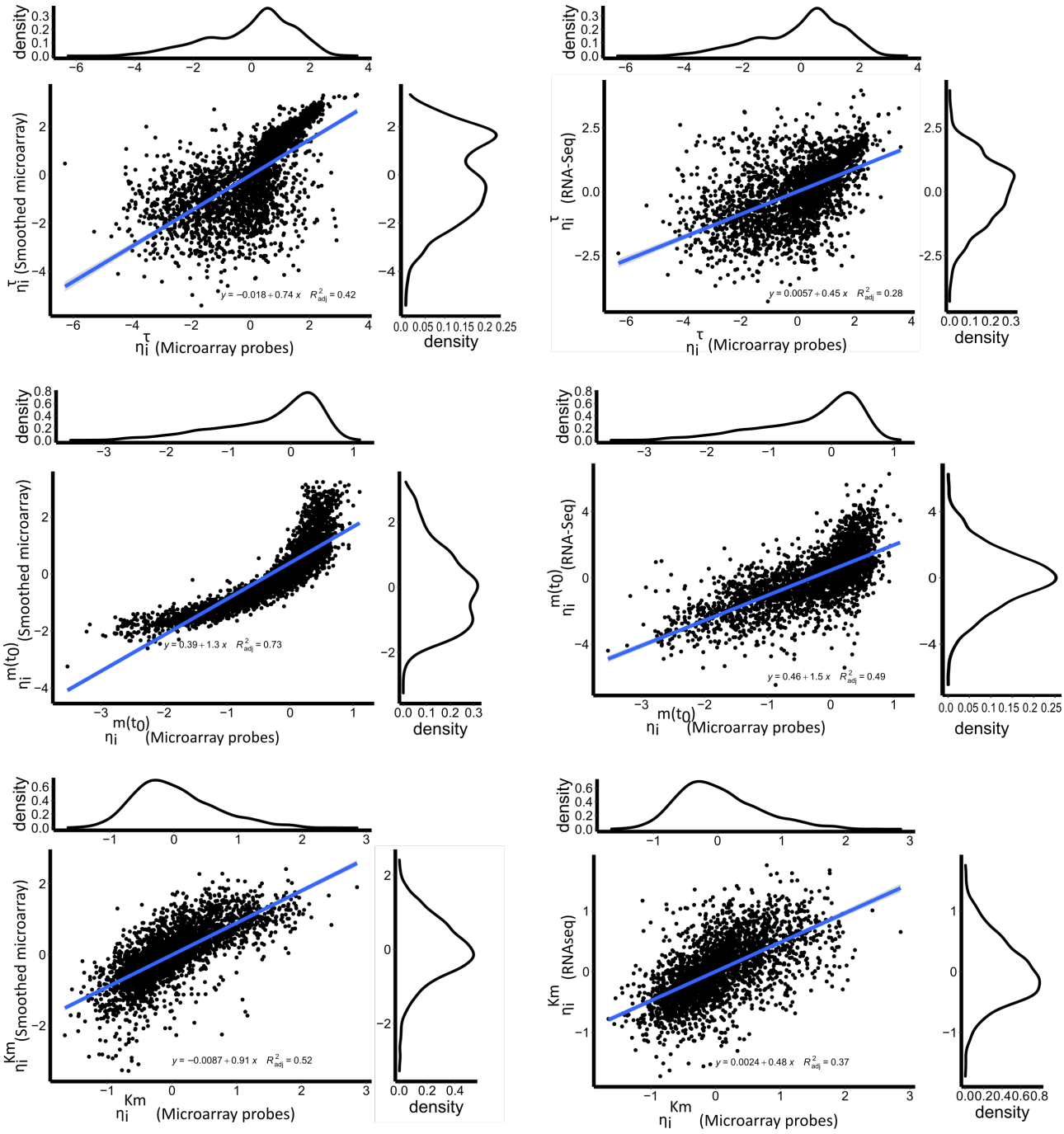

Figure 14: Comparison of the random effects between mRNAs for τ (top panels), $m(t_0)$ (middle), and Km (bottom), between smoothed and nonsmoothed microarray data (left column) and between nonsmoothed microarray and RNA-Seq data

6.5 Postprocessing of the estimation results

Functional enrichment analysis was performed with the R package `clusterProfiler` (Yu *et al.*, 2012), by using the function `gseGO`. We chose the "ALL" ontology (biological process + molecular function + cellular component) and a pvalueCutoff equal to 0.05. We corrected the p-values with the Benjamini-Hochberg procedure.

Functional groups possibly include mRNAs regulated at the post-transcriptional level by small RNAs. They can be identified by the prediction tool CopraRNA (Wright *et al.*, 2014).

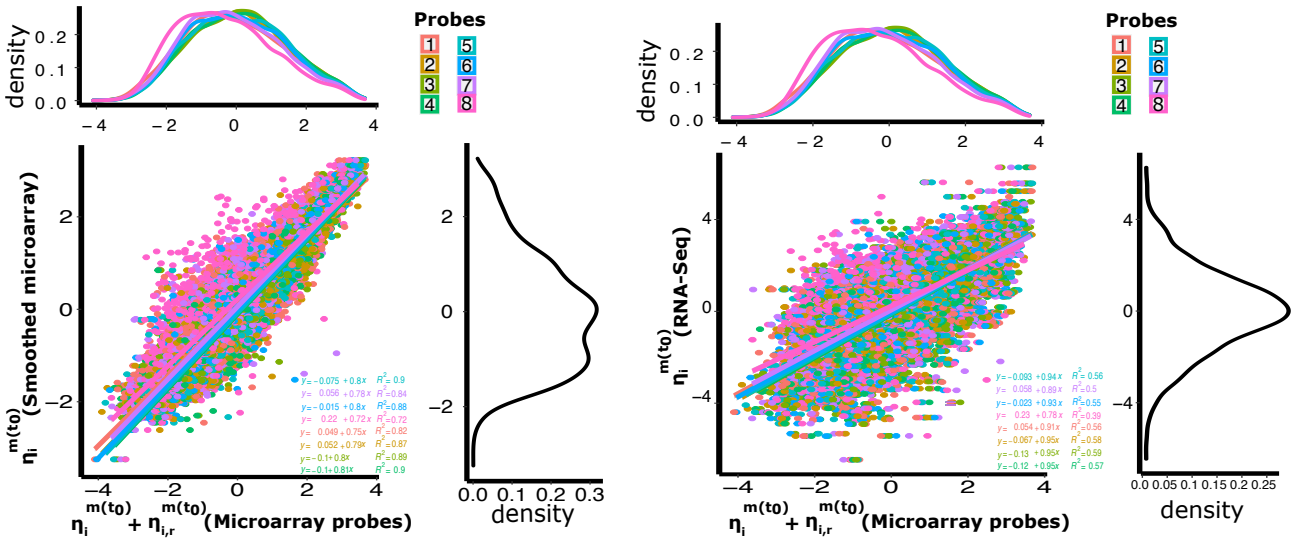

Figure 15: Comparison of the random effects for $m(t_0)$, between mRNA and between probes. Left: comparison of the estimates from smoothed and nonsmoothed microarray data. Right: comparison of RNA-Seq and nonsmoothed microarray data. The colour code corresponds to the eight different probes

6.6 Data availability

The Monolix script for estimation and output files are freely available at https://gitlab.inria.fr/tetienne/eccb_script, together with the microarray dataset. The RNA-Seq dataset is being prepared for publication (Roux *et al.*, in preparation) and will be made available on demand upon acceptance of the article.

Output files in *.txt* format include estimated population and individual parameters for each of the three estimations, together with the statistical results on model assessment (see the accompanying README file for details).