Estimating RNA dynamics using one time point for one sample in a single-pulse metabolic experiment

Micha Hersch [1,2], Adriano Biasini [1], Ana Claudia Marques [1] and Sven Bergmann [1,2]

[1] Department of Computational Biology, University of Lausanne,CH-1015 Lausanne, Switzerland and
[2] Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland.

Abstract

Over the past decade, experimental procedures such as metabolic labeling for determining RNA turnover rates at the transcriptome-wide scale have been widely adopted. Several computational methods to estimate RNA processing and degradation rates from such experiments have been suggested, but they all require several RNA sequencing samples. Here we present a method that can estimate RNA synthesis, processing and degradation rates from a single sample. To this end, we use the Zeisel model and take advantage of its analytical solution, reducing the problem to solving a univariate non-linear equation on a bounded domain. This makes our method computationally efficient, while enabling inference of rates that correlate well with previously published data sets. Using our approach on a single sample, we were able to reproduce and extend the observation that dynamic biological processes such as transcription or chromatin modifications tend to involve genes with higher metabolic rates, while stable processes such as basic metabolism involve genes with lower rates.
In addition to saving experimental work and computational time, having a sample-based rate estimation has several advantages. It does not require an error-prone normalization across samples and enables the use of replicates to estimate uncertainty and perform quality control. Finally the method and theoretical results described here are general enough to be useful in other settings such as nucleotide conversion methods.

1 Introduction

Since the advent of molecular biology, a consensus has emerged that the regulation of gene expression underlies most biological processes including development, disease and adaptation [19, 15, 18]. While gene expression regulation has mostly been associated with activating the production of RNA (e.g. through transcription factors), it has become apparent that the regulation of RNA splicing and RNA stability also plays an important role in determining the expression level of a gene [8, 1]. Taking advantage of high throughput RNA quantification protocols, methods designed to distinguish the effects of RNA synthesis, processing and degradation at the transcriptome-wide level have been developed. Among them, RNA metabolic labeling techniques relying on chemically modified ribonucleotides such as 4-thiouridine (4sU) and 5'-Bromouridine (BrU) have been widely adopted, as their impact on cellular function is minimal [9]. Briefly, incubating cells with modified ribonucleotides for a limited period of time (referred to as the pulse), and their concomitant incorporation in newly synthesized transcripts, allows distinguishing newly transcribed from preexisting RNA, which can be biochemically separated and quantified. This quantification, initially performed using microarray [7] and later using RNA-seq [20, 23], can then be used to estimate RNA decay rates. More recently, methods that rely on nucleotide conversion have been used to the same effect, with the advantage of circumventing the cumbersome biochemical enrichment and separation step.

In the last decade, several methods to estimates RNA dynamics from metabolic labeling experiment data have been developed [26, 21, 2]. Typically, labeled transcript abundance are fitted to an exponential approach to equilibrium, from which the RNA half-life can be estimated [24, 16]. This requires time-course experiments in order to have enough points for fitting, as well as a way to normalize RNA concentrations across samples, either using spike-ins [22], or using internal controls such as intron concentrations [17]. The INSPEcT method [6] goes beyond first order dynamics and takes into account the RNA processing rates, which are estimated along with the degradation and synthesis rates. This increases the number of parameters in the model and thus the number of samples needed for the estimation.

In this work, we build on this approach. However, by considering the intron to exon ratio for each transcript in both the labeled and unlabeled RNA pools, we bypass the need for normalization across samples. Moreover by using the analytical solution to our RNA model, we can infer processing and degradation rates from a single sample and time point. This has several advantages, such as reducing the experimental load and costs, as well as enabling comparisons across samples and time points. Applying our method to our own experimental data and using a single sample and time point, we obtain mRNA degradation rates that correlate well with previously published rates obtained with three replicates and seven time points [12].

2 Method

2.1 Overview

This paragraph summarizes the general strategy of the method, with references to relevant equations indicated in parentheses. We use the Zeisel model of RNA dynamics [27] to model both the unlabeled and

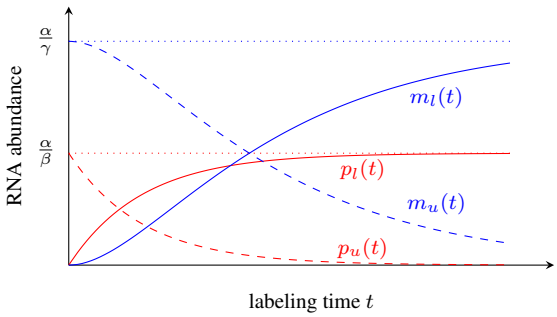

Fig. 1. Evolution of unlabeled and labeled, premature and mature RNA during labeling according to the Zeisel model. Dotted horizontal lines correspond to steady-state levels, dashed lines correspond the unlabeled RNA and solid lines to labeled RNA. Processing and degradation rates can be estimated from the ratios of the two dashed lines and of the two solid lines at a single time point.

the labeled RNA (1). Using the standard procedure for solving systems of linear differential equations, we find its general solution and its free parameters by setting the initial conditions for both the unlabeled (or pre-existing) and the labeled RNA (3,5), as illustrated in Fig 1. We can then express, for a given gene, the ratios for both unlabeled and labeled RNA of intron to exon expression level as functions of the processing and degradation rate of that gene (8,9). Those two ratios are independent from the RNA synthesis rate. Using the intron to exon ratios as observables, we are left with two non-linear equations and two unknowns, namely the processing and degradation rates. Those equations are then reparametrized with dimensionless parameters and reduced to a single non-linear equation with one unknown (22). This resulting equation is only defined on a bounded domain (24). Our rates can thus be inferred by numerically solving that equation on a bounded domain, which is very fast. In addition, we prove in Appendix 2 that under certain conditions, that equation has a single solution (but in general it can also have two or no solution).

2.2 Model

Like previous work [6], we use the Zeisel model of RNA synthesis, processing and degradation [27].

$$\dot{p} = \alpha - \beta p \tag{1}$$

$$\dot{m} = \beta p - \gamma m\,, \tag{2}$$

where p is the premature RNA, m the mature RNA, and α, β, γ are RNA the synthesis, processing and degradation rates. This model can be solved analytically (see appendix 1). In particular, enforcing the boundary conditions corresponding to the unlabeled RNA, namely that it is a steady-state when the pulse starts (t = 0) and then pre-mature RNA is not produced anymore, results in

$$p_u(t) = \frac{\alpha}{\beta}\exp(-\beta t) \tag{3}$$

$$m_u(t) = \frac{\alpha}{\gamma-\beta}\exp(-\beta t) - \frac{\beta\alpha}{\gamma(\gamma-\beta)}\exp(-\gamma t)\,, \tag{4}$$

where the u subscript indicates that this corresponds to the unlabeled RNA pool.
Enforcing boundary conditions corresponding to the labeled RNA, namely that it is not expressed at $t = 0$ leads to

$$p_l(t) = \frac{\alpha}{\beta}\big(1-\exp(-\beta t)\big) \tag{5}$$

$$m_l(t) = \frac{\alpha}{\gamma}\big(1+\frac{\beta}{(\gamma-\beta)}\exp(-\gamma t)\big) - \frac{\alpha}{\gamma-\beta}\exp(-\beta t) \tag{6}$$

where the l subscript indicates that this corresponds to the labeled RNA pool.

2.3 Inferring processing and degradation rates

We consider that the exonic RNA abundance χ corresponds to the processed and mature RNA, while the intronic RNA abundance ι correspond to the processed RNA only. Furthermore, we assume that χ and ι are suitably normalised for exonic and intronic length so that they are proportional to the number of transcripts. We can then compute:

$$\frac{\iota}{\chi} = \frac{p(T)}{p(T)+m(T)}\,, \tag{7}$$

where T is the duration of the labeling.

In the case of unlabeled fraction, we have

$$\begin{aligned}\frac{\iota_u}{\chi_u} &= \frac{p_u(T)}{p_u(T)+m_u(T)}\\ &= \frac{E_\beta}{\beta\big((\frac{1}{\beta}+\frac{1}{\gamma-\beta})E_\beta - \frac{\beta}{\gamma(\gamma-\beta)}E_\gamma\big)}\\ &= \frac{E_\beta}{\frac{\gamma}{\gamma-\beta}E_\beta - \frac{\beta^2}{\gamma(\gamma-\beta)}E_\gamma}\\ &= \frac{(\gamma-\beta)E_\beta}{\gamma E_\beta - \frac{\beta^2}{\gamma}E_\gamma}\\ &= \frac{\gamma(\gamma-\beta)E_\beta}{\gamma^2 E_\beta - \beta^2 E_\gamma}\end{aligned} \tag{8}$$

where we define $E_\beta = \exp(-\beta T)$ and $E_\gamma = \exp(-\gamma T)$ as abbreviations.

For the labeled fraction, we have

$$\begin{aligned}\frac{\iota_l}{\chi_l} &= \frac{p_l(T)}{p_l(T)+m_l(T)}\\ &= \frac{\big(1-E_\beta\big)}{\big(1-E_\beta\big) - \frac{\beta}{\gamma-\beta}E_\beta + \frac{\beta}{\gamma}\big(1+\frac{\beta}{\gamma-\beta}E_\gamma\big)}\\ &= \frac{\big(1-E_\beta\big)}{\frac{\gamma+\beta}{\gamma} - \frac{\gamma}{\gamma-\beta}E_\beta + \frac{\beta^2}{\gamma(\gamma-\beta)}E_\gamma}\\ &= \frac{\gamma(\gamma-\beta)\big(1-E_\beta\big)}{\gamma^2-\beta^2+\beta^2E_\gamma-\gamma^2E_\beta}\\ &= \frac{\gamma(\gamma-\beta)\big(1-E_\beta\big)}{\gamma^2\big(1-E_\beta\big)-\beta^2\big(1-E_\gamma\big)}.\end{aligned} \tag{9}$$

We notice that this last expression is of the same form as the one for the unlabeled fraction (8), but replacing exponentials by their complement to one. Importantly those two fractions do not depend on α, which (unlike [10]) allows our method to estimate processing and degradation rates independently from the synthesis rate.

Denoting $a = \frac{\iota_u}{\chi_u}$ and $b = \frac{\iota_l}{\chi_l}$ as the observable unlabeled and labeled fractions of intron abundance, we are left with a system of two

equations and two unkowns β and γ, which we now set out to solve. First, we reparametrize our system with $\beta = k\gamma$ and define $E_{k\gamma} = E_\beta = \exp(-k\gamma T)$ leading to

$$a = \frac{(1-k)E_{k\gamma}}{E_{k\gamma} - k^2 E_\gamma} \tag{10}$$

$$b = \frac{(1-k)\big(1-E_{k\gamma}\big)}{\big(1-E_{k\gamma}\big) - k^2\big(1-E_\gamma\big)}\,. \tag{11}$$

We thus have

$$a\Big(E_{k\gamma} - k^2 E_\gamma\Big) = (1-k)E_{k\gamma} \tag{12}$$

$$b\Big(\big(1-E_{k\gamma}\big) - k^2(1-E_\gamma)\Big) = (1-k)\big(1-E_{k\gamma}\big)\,. \tag{13}$$

Summing (12) and (13) yields

$$E_{k\gamma}(a-b) + k^2 E_\gamma(b-a) + b(1-k^2) = 1-k \tag{14}$$

$$\begin{aligned} \Leftrightarrow \quad E_{k\gamma} - k^2 E_\gamma &= \frac{(1-k) - b(1-k^2)}{a-b} \\ &= \frac{(1-k)\big(1-b(1+k)\big)}{a-b}\,. \end{aligned} \tag{15}$$

Dividing (12) by (13) and inserting (15) results in

$$\begin{aligned} \frac{E_{k\gamma}}{1-E_{k\gamma}} &= \frac{a}{b}\frac{E_{k\gamma} - k^2 E_\gamma}{\big(1-E_{k\gamma}\big) - k^2(1-E_\gamma)} \\ &= \frac{a}{b}\frac{E_{k\gamma} - k^2 E_\gamma}{(1-k^2) - \big(E_{k\gamma} - k^2 E_\gamma\big)} \\ &= \frac{a}{b}\frac{(1-k)(1-b(1+k)}{(1-k^2)(a-b) - (1-k)\big(1-b(1+k)\big)} \\ &= \frac{a}{b}\frac{1-b(1+k)}{(1+k)(a-b) - 1 + b(1+k)} \\ &= \frac{a}{b}\frac{1-b(1+k)}{(1+k)a-1} = -\frac{a-ab(1+k)}{b-ab(1+k)} \end{aligned} \tag{16}$$

It follows that

$$E_{k\gamma}\big(b-ab(1+k)\big) = \big(E_{k\gamma}-1\big)\big(a-ab(1+k)\big) \tag{17}$$

$$\Leftrightarrow \quad (b-a)E_{k\gamma} = ab(1+k) - a\,, \tag{18}$$

an thus

$$\exp(-k\gamma T) = E_{k\gamma} = \frac{kab+ab-a}{b-a} = \frac{a(bk+b-1)}{b-a} \tag{19}$$

$$\Leftrightarrow \quad k\gamma T = \log\big(\frac{b-a}{a(bk+b-1)}\big) \tag{20}$$

Moreover, from (10), we have that

$$a = \frac{1-k}{1-k^2\exp\big((k-1)\gamma T\big)} \Leftrightarrow \exp\big((k-1)\gamma T\big) = \frac{\big(1-\frac{1-k}{a}\big)}{k^2}$$

$$\Leftrightarrow \quad (k-1)\gamma T = \log\big(\frac{k+a-1}{k^2 a}\big) \tag{21}$$

Multiplying (20) by $\frac{k-1}{k}$ and subtracting (21) results in

$$0 = \frac{k}{k-1}\log\big(\frac{k+a-1}{k^2 a}\big) - \log\Big(\frac{(b-a)}{a(bk+b-1)}\Big) \tag{22}$$

Our system of two equations can thus be reduce to a single equation which does not explicitly depend on T and can be solved numerically. In practice a and b are approximated by r_u and r_l, defined as the length-normalized intronic to exonic read count ratio (or TPM ratio) for the unlabeled and for the labeled sampled respectively. This equation also provides upper and lower bounds for k as both $\frac{k+a-1}{a}$ and $bk+b-1$ must be strictly positive for their logarithm to be defined and

$$0 < \exp(-\beta T) = E_{k\gamma} = \frac{kab+ab-a}{b-a} < 1 \quad \forall \beta T > 0 \tag{23}$$

for (19) to hold. Developing those three conditions results in the following domain of definition $\mathcal{D}$ for k:

$$\max(\frac{1}{b}-1, 1-a) < k < \frac{1}{a} - 1\,, \tag{24}$$

where $0 < a < b < 1$. Note that the right-hand side of (22) is in general continuous in $k = 1$, but not in $k = 1-a$. Furthermore, it can be shown (see Appendix 2) that for $b > \frac{1}{2-a}$, (22) has a single solution in the domain given by (24), which can be found very efficiently. This enables the estimation of the processing and degradation rates for a single sample. Moreover, since the reduced equation is independent from T, uncertainty on its true value does not affect the relative values of the resulting rates. Hence replicates can be used to assess the reliability of the estimates and time courses allow to test whether the rates are constant as assumed by the model.

If (22) does not have a solution, estimates can be obtained by minimizing (in log space) the squared Euclidian distance between the observed (i.e., r_u, r_l) and derived values of a and b:

$$f(k,\gamma T) = \Big(\log(r_u) - \log\big(\frac{(1-k)}{\exp(-k\gamma T) - k^2\exp((k-1)-\gamma T)}\big)\Big)^2 + \Big(\log(r_l) - \log\big(\frac{(1-k)(1-\exp(-k\gamma T))}{(1-\exp(-k\gamma T)) - k^2(1-E_\gamma)}\big)\Big)^2,$$

The ratios r_u, r_l must be smaller than one to make sense within our model and genes where this is not the case should be discarded. The log function is used to give exon and intron counts equal standing.
The above bivariate function can be reduced to a univariate function f^* using (20):

$$f^*(k) = f\Big(k, \frac{1}{k}\log\big(\frac{r_l - r_u}{r_u(r_l k + r_l - 1)}\big)\Big) \tag{25}$$

Once processing and degradation rates are obtained, the (relative) synthesis rates α can be easily obtained from (4), where m_u is approximated by χ_u (which is likely the most reliably measured specie).

3 Results

3.1 Simulated data

In order to confirm that our method can be applied in principle, we evaluated our method on simulated data, where the data was generated using the exact model used to develop the method (see equations 3 and following). We then generated 50000 random value for α, β, and γ ranging between $\exp(-5)$ and $\exp(5)$ and computed the corresponding values for ι and χ. We then computed r_u and r_l by taking the ratio. Estimates $\hat{\beta}$ and $\hat{\gamma}$ where then inferred by using r_u and r_l as an input to the method and compare the original β and γ.
Numerically solving equation (22), yielded either one or two solutions. The results for the unambiguous cases are shown in Fig 2. We see that in virtually all cases, the method yields accurate estimates of the processing and degradation rates. For a few points, the method is less accurate at the upper boundary of the parameter space, probably due limited floating

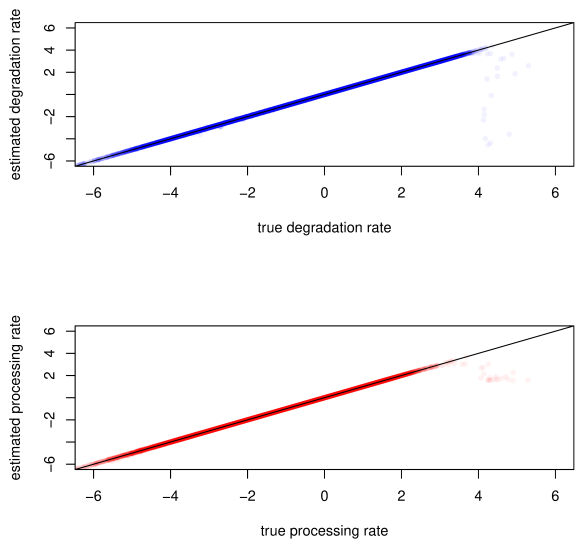

Fig. 2. Simulated data. The method correctly estimates processing and degradation rates. Points with ambiguous solutions are not shown. Some points corresponding to high rates cannot be estimated correctly as the system as already reached steady-state during the simulated "pulse".

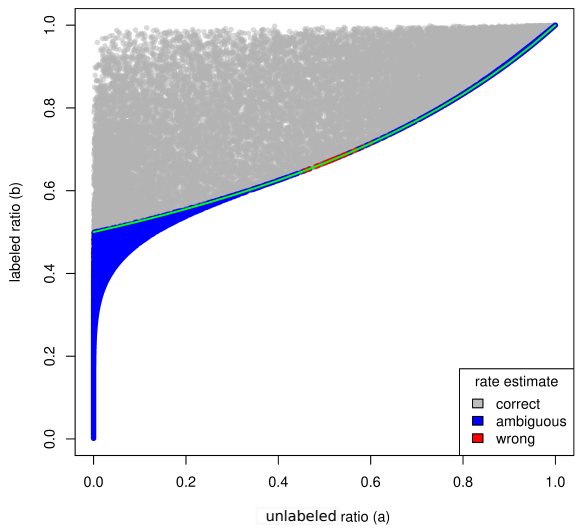

Fig. 3. Simulated data. The measurement space can be partitioned into ambiguous and unambiguous regions. The green line corresponds to $b = \frac{1}{2-a}$. Above that line, rates are correctly and unambiguously estimated. Boundary cases are sometimes wrongly estimated, probably due to numerical errors (red dots).

point precision. Indeed if the labeling time is too long with respect to the metabolic rates, virtually all unlabeled RNA are degraded and the rates cannot be reliably estimated.

As we are considering single-sample estimates, it is possible to chart the observable space given by a and b and see when the method provides unambiguous results. Fig. 3 confirms that for $b > \frac{1}{2-a}$ the method

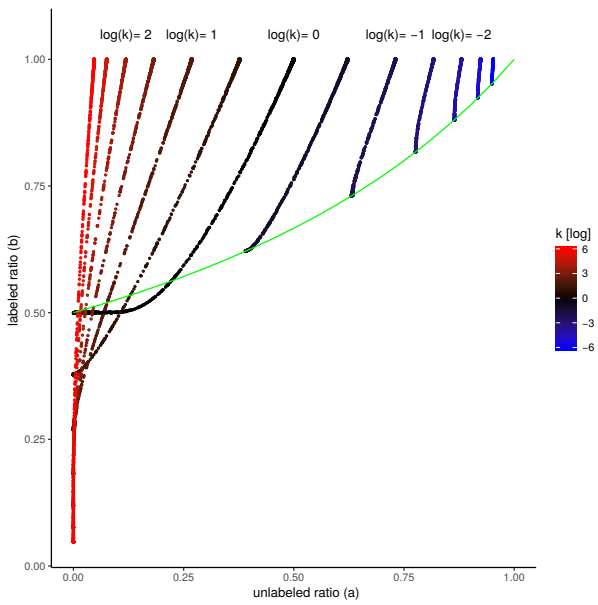

Fig. 4. Observable space of the dynamical system. Trajectories in the phase space are solely determined by the k parameter. They start at time $T = 0$ at the top ($b = 1$) and go down. For $k < 1$ the trajectories (in blue) remain above the green line defined by $b = (2 - a)^{-1}$ and do not cross. For $k > 1$ (in red), they cross each other below the system follows the trajectory depends on the actual values of β and γ.

provides a unique (and correct) solution as proven in appendix 2. Below this line (displayed in green), the methods provides ambiguous results as two distinct set of values β and γ can account for the same value of a and b (in blue).

It is also possible to visualize the trajectories of the observables a and b for various values of k, as depicted in Fig. 4. When $T = 0$, trajectories start from the top of the space at $(\frac{1}{1+k}, 1)$. When $k < 1$, as time passes the system moves down to $(a, b) \to (1 - k, \frac{1}{1+k})$. For $k \geq 1$, trajectories move to $(0, \frac{1}{1+k})$. Note that this is the expected case, as the splicing of mRNA occurs in general faster than its degradation. Note that, in this case, trajectories cross below the green line, explaining why two solutions can be found for a single value of (a, b). The speed at which the system follows those trajectories depends on γ.

3.2 Real data

In order to assess the performance of the method on real data, we applied our method on the 4sU labeling experiment described in [3]. Briefly, mouse embryonic stem cells were plated at a density of 40000 cells/cm^2 on gelatin-coated 10cm tissue culture plates and grown for approximately 14 hours. After addition of 4sU to the growth medium, cells were incubated at 37C for 10 minutes (10 minutes labeling pulse). RNA was then extracted and processed according to the protocol described in [4]. Reads that did not map to mouse ribosomal RNA sequences were aligned to intronic and exonic sequences using STAR V2.5 and quantified using RSEM V1.1.17, yielding intron and exon expression levels for unlabeled and labeled RNA.

For a single sample, the observable space represented in Figs 3 and 4 is represented (in log coordinates) in Fig 5. We see that, while the points are centered on the expected region of the observable space, many transcripts lie above $r_l = 1$ (in blue) or below the diagonal (in red), which

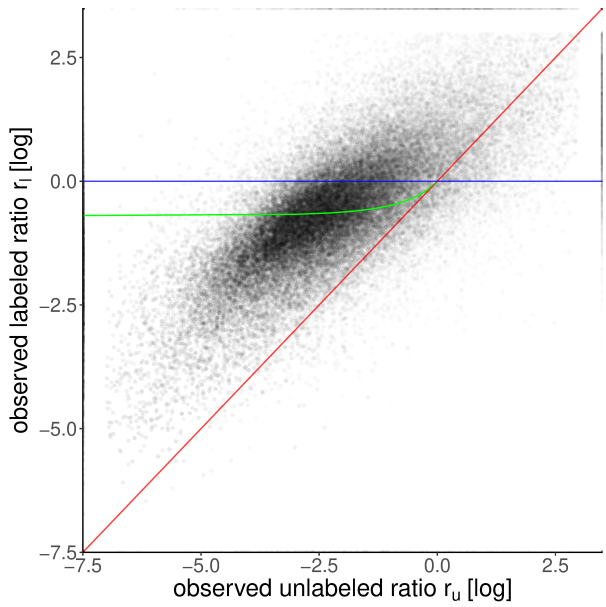

Fig. 5. Real data. Each point corresponds to a transcript with its transparency reflecting log expression value. Like in the previous figure, the green line is defined $y = (2 - x)^{-1}$. For transcripts lying between the abscissa (in blue) and the green line, estimates of processing and degradation rates can be obtained by solving (22). For transcripts lying between the diagonal (in red) and the green line, estimates can be obtained by minimizing (25). The observed ratios for the remaining transcripts are not coherent with the model and are discarded.

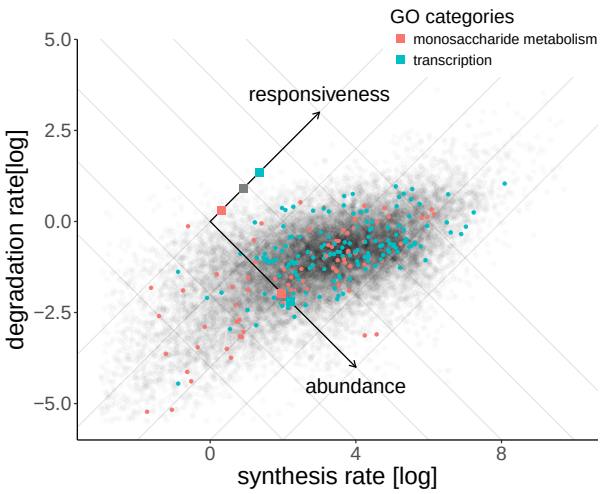

Fig. 6. Estimated mRNA synthesis and degradation rates obtained from a single sample. Those rates can also be considered in a different and maybe functionally more relevant frame of reference defined by the steady state abundance (first axis) and gene responsiveness (second and perpendicular axis), as illustrated by the background grid. Genes involved in fast adapting biological processes (such as transcription) tend to be more responsive than genes involved in stable functions (such as monosaccharide metabolism). The squares on the axes represent the projections of the mean rates for the respective categories (gray representing genes that belong to neither of the two categories) and indicate that mean transcript responsiveness (but not abundance) is strongly affected by the category. Those two GO categories were selected for illustration because they were previously reported to be mostly enriched in high and low turn-over genes respectively [23].

is not compatible with our model. Those transcripts, amounting to 25% of protein-coding genes with an exon TPM higher than 1, are discarded from further analyses.
The processing and degradation rates were computed either by solving (22) when $r_l > (2 - r_u)^{-1}$ or by optimizing (25) otherwise. It took a few seconds to estimate several tens of thousands of rates on a desktop computer. For those cases that had two solutions (6% of the transcripts), we selected the one corresponding to rates most consistent with the other transcripts.

The resulting synthesis and degradation rates are depicted in Fig. 6. They have a correlation of 65%, which is very similar to the 66% reported by [12] for the same system. This is also consistent with the emerging concept of a coupling between RNA transcription and decay [11]. Our data indicate that genes span a large range of dynamics, irrespective of their expression level. Indeed, genes with high synthesis and degradation rates can have the same steady-state expression level as genes with low synthesis and degradation rates. However, the former will reach this steady state faster than the latter. It thus make sense to consider our RNA metabolic rates in the functional frame of reference indicated in Fig. 6. One axis corresponds to the steady state concentration, given by the log-ratio of synthesis over degradation rates (or equivalently by the difference of log of the rates). The second axis correspond to the responsiveness of the gene, i.e. how fast it reaches steady state (computed by the sum of the log of the synthesis and degradation rates). It has been observed before that genes involved in more reactive and dynamic biological processes such as chromatin remodeling or transcription regulation tend to have a higher turnover than genes involved in more stable processes such as basic metabolism [23]. We checked that our data confirm this observation by looking at the Gene Ontology (GO,[5])) categories most associated by [23] with high and low turnover, namely "transcription" and "monosaccharide metabolism". Despite having similar steady-state abundances, transcripts of genes involved in transcription indeed have significantly faster dynamics and the ones involved in monosaccharide metabolism have significantly slower dynamics than the rest of the genes, as illustrated by the squares in Fig. 6. Other categories where our data confirms faster genes include chromatin modifications, cell cycle and transcription regulation.

We assessed the precision of our method by comparing the resulting degradation rates to those published for the same cell type by [12]. Those were obtained by using three replicates and seven time points and applying the SLAM-seq nucleotide-conversion method that, unlike metabolic labeling, does not require biochemical separation between the labeled and unlabeled RNA and is thus not affected by noise generated by the imperfect separation process (although that method has its own source of noise). From our data, we obtained gene degradation rates by taking, for each gene, the weighted average degradation rates of the corresponding transcripts The weights were given by the mean exonic expression levels (unlabeled and labeled). We expect a lower precision for transcripts close to the $r_l = 1$ line, for which the labeling time was likely somewhat too short, so to assess the correlation, we weighted the transcripts by $1 - r_l$. Fig. 7 compares degradation rates obtained in our experiments with those reported by [12], keeping only genes with an average expression value higher than 100 TPM. We expect

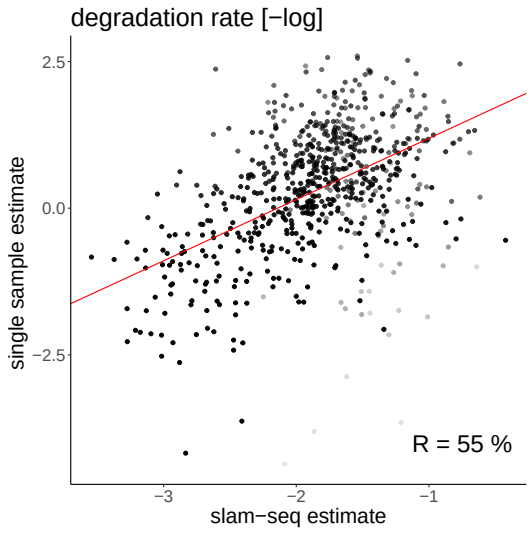

Fig. 7. Degradation rates estimated from a single sample plotted against degradation rates published in [12] (obtained using slam-seq). The red line is obtained through weighted linear regression. The weights are set as $1 - r_l$ as indicated by the transparency of the dots. The (weighted) correlation of 55% indicates that the estimated rates are meaningful. Only genes with a mean exon TPM above 100 are taken into account.

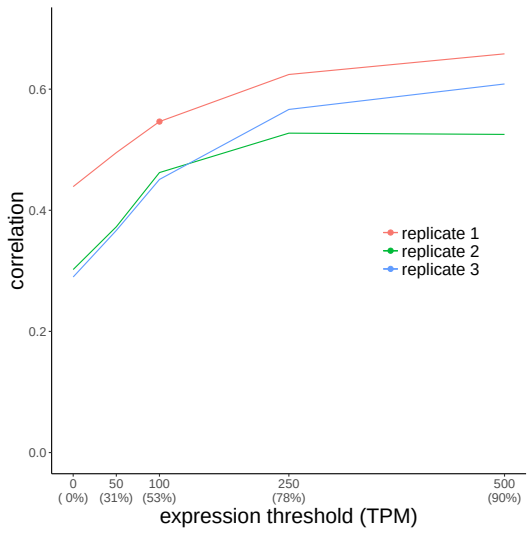

Fig. 8. Correlation between degradation rates obtained by [12] and the ones obtained our single-sample method as a function of expression level. Each line represents a biological replicate. The dot corresponds to the data shown in Fig 7. As expected, the correlation is higher for highly expressed genes, as the intro to exon ratios can be more reliably estimated. In this experiment, replicate 1 correlates better than the two others, indicating that it is probably of better quality.

a higher precision for highly expressed genes, as this allows for a more precises estimates of the intron to exon ratios. This is indeed the case, and depending on the expression threshold and the sample, the correlation between our data and the previously published rates, we obtain a correlation ranging between 30% and 67% for a single sample estimate (see Fig. 8). As those are experiments performed in different labs using different methods, those numbers show that our rates obtained on a single sample and time point are meaningful. For comparison, [13] reports correlations around 70% by using the *same* data, but changing only the method of analysis. Using three replicates, [4] reports a 26% correlation using the INSPEcT package.

4 Discussion

In this paper, we presented a method to estimates splicing and degradation rates of RNA transcripts from a single 4sU labeled sample. We validated our method first *in silico* and then on real data obtained from mouse embryonic stem cells. Using our method we first replicated, on a different cell type, previous findings about the enrichment in high or low turnover gene of specific cellular processes. Second, we showed that the rates obtained with our method correlate well (between 30% and 67%) with published rates obtained by applying SLAM-seq to the same cell types. Methods for such estimation have been published before, but they usually require a sufficient number of samples (around a dozen). In contrast to these methods, our method explicitly uses the analytical solution to the standard RNA dynamics model given by (26). Moreover, our method is self-normalizing as it only uses the ratio of intron to exon expression levels. It is thus not affected by differences in sequencing depth of the various samples. This approach makes our method also faster than other methods as it boils down to numerically solving on a bounded domain either a univariate equation or a one-dimensional optimization for each transcript. However, a caveat of our method is that a sizable fraction of mostly lowly expressed transcripts (about 25 % in our case) are inconsistent with the model and their dynamics cannot be estimated, while the method provides two solutions for another fraction (6%) of the transcripts. However, our theoretical considerations indicate that this issue is constitutive of the model, and is likely to also impact other methods (for example through mutliple local extremas in the likelihood function), although it may be more difficult to detect it. Using multiple samples or time-points is likely to help solve this problem.

While using a single sample allows to reduce costs, this is not the only merit of this approach. In practice most experiments will have biological replicates, in which case our methods enables obtaining point estimates of α, β and γ for each of them. This in turn allows for estimating their variance, as well as assessing sample quality (e.g. if one of them systematically gives very different estimates for all genes). Moreover because cell growth is likely to be limited during (short) labeling time, it is less likely to interfere in the estimation process than when using time course data, where it can have an effet [17]. In addition, when used in a time-course experiment, our method allows to investigate the evolution of those rates over time and assess whether those rates are stationnary. Finally, the theoretical results obtained in this paper, could be used to improve other methods. For example, the method could be used to analyze SLAM-seq data which would reduce the number of samples but also provide estimate for the processing rate. Another possible application is single cell RNA velocity [14], where the Zeisel model of RNA dynamics is also used, but splicing rates γ are set to be equal for all transcripts. While it has been documented (and is consistent with our data) that splicing rates are more homogeneous than degradation rates [21], this is potentially an approximation that could be improved with our framework to increase the accuracy of the method.

The method presented in this paper can be adapted for the case when unlabeled RNA is mixed with labeled RNA in a "total" rather than a "unlabeled " RNA pool. In that case, the intron to exon ratio in the total RNA pool is constant during labeling time and is given by $\frac{1}{1+k}$, and rates can be easily obtained from (11). This method is however likely to be

less precise than separating unlabeled from labeling RNA, as additional information can be gained from the decreasing unlabeled RNA pool.

Our method could be further improved in several ways. For example, unlike in [10], we did not consider the effect of leakage of unlabeled RNA in the labeled RNA pool because of unspecific capture. This leakage has the effect of dragging r_l down towards the diagonal, and could potentially be estimated from the data as it is shared across all transcripts.
Another improvement would be to embed this method in a probabilistic framework in order to quantify the estimate uncertainty (as in [13] for a simpler model) or to determine the optimal labeling time (as in [25]).

Acknowledgements

This work was funded by the Swiss National Science Foundation through grant no. FN 310030_152724/1 to S.B and PP00P3_150667 and the NCCR in RNA & Disease to A.C.M.

Author contribution and conflict of interest

.H. developed and implemented the method, analyzed the expression quantification data, interpreted the results, figured out the proof, generated the figures and wrote the manuscript, A.B. performed the experiments and interpreted the results, A.C.M. initiated the project, designed the study and generated the expression quantification data, S.B. reviewed the math, interpreted the results and supervised the process. All authors contributed to the manuscript.
The authors declare that they have no conflicts of interest.

Data and code availabiity

An R package implementing our method is available on github, together with the code used to generate the figures as well as the gene expression data used: https://github.com/BergmannLab/SingleSampleRNAdynamics
The raw data files data are available on the Gene Expression Omnibus accession number GEO:GSE150286 (main replicate) and GEO: GSE143277 (second and third replicates of Fig. 8)

References

[1]Tara Alpert, Lydia Herzel, and Karla M Neugebauer. Perfect timing: splicing and transcription rates in living cells. *WIREs: RNA*, 8(2):e1401, 2017.

[2]J David Barrass, Jane EA Reid, Yuanhua Huang, Ralph D Hector, Guido Sanguinetti, Jean D Beggs, and Sander Granneman. Transcriptome-wide RNA processing kinetics revealed using extremely short 4tu labeling. *Genome biology*, 16(1):282, 2015.

[3]Adriano Biasini, Stefano De Pretis, Jennifer Yihong Tan, Baroj Abdulkarim, Harry Wischnewski, Rene Dreos, Mattia Pelizzola, Constance Ciaudo, and Ana Claudia Marques. Translation is required for miRNA-dependent decay of endogenous transcripts. *BioRxiv*, 2020.

[4]Adriano Biasini and Ana Claudia Marques. A protocol for transcriptome-wide inference of RNA metabolic rates in mouse embryonic stem cells. *Frontiers in Cell and Developmental Biology*, 8:97, 2020.

[5]Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic acids research*, 47(D1):D330–D338, 2019.

[6]Stefano De Pretis, Theresia Kress, Marco J Morelli, Giorgio EM Mellon i, Laura Riva, Bruno Amati, and Mattia Pelizzola. INSPEcT: a computational tool to infer mrna synthesis, processing and d egradation dynamics from rna-and 4su-seq time course experiments. *Bioinformatics*, 31(17):2829–2835, 2015.

[7]Lars Dölken, Zsolt Ruzsics, Bernd Rädle, Caroline C Friedel, Ralf Zimmer, Jörg Mages, Reinhard Hoffmann, Paul Dickinson, Thorsten Forster, Peter Ghazal, et al. High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *Rna*, 14(9):1959–1972, 2008.

[8]Ran Elkon, Eitan Zlotorynski, Karen I Zeller, and Reuven Agami. Major role for mrna stability in shaping the kinetics of gene induction. *BMC genomics*, 11(1):259, 2010.

[9]Caroline C Friedel and Lars Dölken. Metabolic tagging and purification of nascent rna: implications for transcriptomics. *Molecular BioSystems*, 5(11):1271–1278, 2009.

[10]Mattia Furlan, Stefano de Pretis, Eugenia Galeota, Michele Caselle, and Mattia Pelizzola. Dynamics of transcriptional regulation from total RNA-seq experiments. *bioRxiv*, p. 520155, 2019.

[11]Ella Hartenian and Britt A Glaunsinger. Feedback to the central dogma: cytoplasmic mrna decay and transcription are interdependent processes. *Critical reviews in biochemistry and molecular biology*, 54(4):385–398, 2019.

[12]Veronika A Herzog, Brian Reichholf, Tobias Neumann, Philipp Rescheneder, Pooja Bhat, Thomas R Burkard, Wiebke Wlotzka, Arndt von Haeseler, Johannes Zuber, and Stefan L Ameres. Thiol-linked alkylation of RNA to assess expression dynamics. *Nature methods*, 14(12):1198, 2017.

[13]Christopher Jürges, Lars Dölken, and Florian Erhard. Dissecting newly transcribed and old RNA using GRAND-SLAM. *Bioinformatics*, 34(13):i218–i226, 2018.

[14]Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E Kastriti, Peter Lönnerberg, Alessandro Furlan, et al. RNA velocity of single cells. *Nature*, 560(7719):494, 2018.

[15]Tong Ihn Lee and Richard A Young. Transcriptional regulation and its misregulation in disease. *Cell*, 152(6):1237–1251, 2013.

[16]Andrew Lugowski, Beth Nicholson, and Olivia S Rissland. Determining mRNA half-lives on a transcriptome-wide scale. *Methods*, 137:90–98, 2018.

[17]Andrew Lugowski, Beth Nicholson, and Olivia S Rissland. DRUID: A pipeline for transcriptome-wide measurements of mRNA stability. *RNA*, 24(5):623–632, 2018.

[18]Katya L Mack, Mallory A Ballinger, Megan Phifer-Rixey, and Michael W Nachman. Gene regulation underlies environmental adaptation in house mice. *Genome research*, 28(11):1636–1645, 2018.

[19]Florence Petit, Karen E Sears, and Nadav Ahituv. Limb development: a paradigm of gene regulation. *Nature Reviews Genetics*, 18(4):245, 2017.

[20]Michal Rabani, Joshua Z Levin, Lin Fan, Xian Adiconis, Raktima Raychowdhury, Manuel Garber, Andreas Gnirke, Chad Nusbaum, Nir Hacohen, Nir Friedman, et al. Metabolic labeling of rna uncovers principles of rna production and degradation dynamics in mammalian cells. *Nature biotechnology*, 29(5):436, 2011.

[21]Michal Rabani, Raktima Raychowdhury, Marko Jovanovic, Michael Rooney, Deborah J Stumpo, Andrea Pauli, Nir Hacohen, Alexander F Schier, Perry J Blackshear, Nir Friedman, et al. High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. *Cell*, 159(7):1698–1710, 2014.

[22]Joseph Russo, Adam M Heck, Jeffrey Wilusz, and Carol J Wilusz. Metabolic labeling and recovery of nascent RNA to accurately

quantify mRNA stability. *Methods*, 120:39–48, 2017.

[23]Björn Schwanhäusser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337, 2011.

[24]Alexey Uvarovskii and Christoph Dieterich. pulseR: Versatile computational analysis of rna turnover from metabolic labeling experiments. *Bioinformatics*, 33(20):3305–3307, 2017.

[25]Alexey Uvarovskii, Isabel S Naarmann-de Vries, and Christoph Dieterich. On the optimal design of metabolic RNA labeling experiments. *PLoS computational biology*, 15(8):e1007252, 2019.

[26]Lukas Windhager, Thomas Bonfert, Kaspar Burger, Zsolt Ruzsics, Stefan Krebs, Stefanie Kaufmann, Georg Malterer, Anne L'Hernault, Markus Schilhabel, Stefan Schreiber, et al. Ultrashort and progressive 4su-tagging reveals key characteristics of RNA processing at nucleotide resolution. *Genome research*, 22(10):2031–2042, 2012.

[27]Amit Zeisel, Wolfgang J Köstler, Natali Molotski, Jonathan M Tsai, Rita Krauthgamer, Jasmine Jacob-Hirsch, Gideon Rechavi, Yoav Soen, Steffen Jung, Yosef Yarden, et al. Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. *Molecular systems biology*, 7(1), 2011.

Appendix

1 Derivation of the model solution

This is a first order linear ordinary differential equation in $p(t)$ and $m(t)$ that can be expressed in matrix form as

$$\begin{pmatrix} \dot{p} \\ \dot{m} \end{pmatrix} = \begin{pmatrix} -\beta & 0 \\ \beta & -\gamma \end{pmatrix} \begin{pmatrix} p \\ m \end{pmatrix} + \begin{pmatrix} \alpha \\ 0 \end{pmatrix} \tag{26}$$

The solution to this equation is given by

$$\begin{pmatrix} p \\ m \end{pmatrix} = k_1 \mathbf{v} \exp(\lambda_1 t) + k_2 \mathbf{w} \exp(\lambda_2 t) + \begin{pmatrix} \frac{\alpha}{\beta} \\ \frac{\alpha}{\gamma} \end{pmatrix}, \tag{27}$$

where k_1 and k_2 are scalar constants determined by the boundary conditions, λ_1, λ_2 are eigenvalues of the matrix in (26) and $\mathbf{v}$, $\mathbf{w}$ are the corresponding eigenvectors.
The eigenvalues are given by $\lambda_1 = -\beta$ and $\lambda_2 = -\gamma$. The first eigenvector $\mathbf{v}$ is obtained by solving

$$\begin{cases} -\beta \mathbf{v}_1 = -\beta \mathbf{v}_1 \\ \beta \mathbf{v}_1 - \gamma \mathbf{v}_2 = -\beta \mathbf{v}_2 \end{cases} \Rightarrow \mathbf{v}_1 = \frac{\gamma - \beta}{\beta} \mathbf{v}_2 \Rightarrow \mathbf{v} \propto \begin{pmatrix} \gamma - \beta \\ \beta \end{pmatrix} \tag{28}$$

Similarly the second eigenvector is obtained by solving

$$\begin{cases} -\beta \mathbf{w}_1 = -\gamma \mathbf{w}_1 \\ \beta \mathbf{w}_1 - \gamma \mathbf{w}_2 = -\gamma \mathbf{w}_2 \end{cases} \Rightarrow \mathbf{w}_1 = 0 \Rightarrow \mathbf{w} \propto \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

The solution to (26) is thus given by

$$\begin{pmatrix} p \\ m \end{pmatrix} = k_1 \begin{pmatrix} \gamma - \beta \\ \beta \end{pmatrix} \exp(-\beta t) + k_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \exp(-\gamma t) + \begin{pmatrix} \frac{\alpha}{\beta} \\ \frac{\alpha}{\gamma} \end{pmatrix}.$$

Expressed by its component this is equivalent to

$$p(t) = k_1(\gamma - \beta) \exp(-\beta t) + \frac{\alpha}{\beta} \tag{29}$$

$$m(t) = k_1 \beta \exp(-\beta t) + k_2 \exp(-\gamma t) + \frac{\alpha}{\gamma} \tag{30}$$

We now turn to the boundary conditions to determine k_1 and k_2. The boundary conditions are different for the unlabeled and the labeled RNA.

Unlabeled RNA
Like in [6], we assume the system to be in steady-state prior to labeling. The steady-state is given by solving (26) with $\dot{p} = \dot{m} = 0$.

$$\begin{cases} 0 = -\beta p + \alpha \\ 0 = \beta p - \gamma m \end{cases} \Rightarrow \begin{cases} p = \frac{\alpha}{\beta} \\ 0 = \beta \frac{\alpha}{\beta} - \gamma m \end{cases} \Rightarrow \begin{cases} p = \frac{\alpha}{\beta} \\ m = \frac{\alpha}{\gamma} \end{cases} \tag{31}$$

During labeling time, we assume that no unlabeled RNA is synthesized such that $\alpha = 0$. Assuming that we start labeling at time $t = 0$, we thus have

$$p_u(0) = \frac{\alpha}{\beta} \Rightarrow k_1(\gamma - \beta) = \frac{\alpha}{\beta} \Rightarrow k_1 = \frac{\alpha}{\beta(\gamma - \beta)} \tag{32}$$

Moreover we have

$$m_u(0) = \frac{\alpha}{\gamma} \Rightarrow \frac{\alpha}{\gamma - \beta} + k_2 = \frac{\alpha}{\gamma} \Rightarrow k_2 = \frac{\alpha}{\gamma} - \frac{\alpha}{\gamma - \beta} = \frac{-\beta\alpha}{\gamma(\gamma - \beta)}$$

This leads us to the solution for the unlabeled RNA

$$p_u(t) = \frac{\alpha}{\beta} \exp(-\beta t) \tag{33}$$

$$m_u(t) = \frac{\alpha}{\gamma - \beta} \exp(-\beta t) - \frac{\beta\alpha}{\gamma(\gamma - \beta)} \exp(-\gamma t), \tag{34}$$

where the u label indicates that this corresponds to the unlabeled RNA pool.

Labeled RNA
The solution for the labeled RNA could be obtained the same way as for the unlabeled RNA, but setting $\alpha \neq 0$ and $p_l(0) = m_l(0) = 0$. However, it is simpler to notice that the total RNA (labeled and non-labeled) stay at steady-state during the labeling such that we have the following solution for labeled RNA.

$$p_l(t) = \frac{\alpha}{\beta} - p_u(t) = \frac{\alpha}{\beta}\big(1 - \exp(-\beta t)\big)$$

$$m_l(t) = \frac{\alpha}{\gamma} - m_u(t) = \frac{\alpha}{\gamma}\big(1 + \frac{\beta}{(\gamma - \beta)} \exp(-\gamma t)\big) - \frac{\alpha}{\gamma - \beta} \exp(-\beta t)$$

where the l label indicates that this corresponds to the labeled RNA pool.

2 Proof of unicity of solution

In this appendix, we prove that (22) has a single solution for $b > (2 - a)^{-1}$. We first note that $b > (2 - a)^{-1} \Leftrightarrow \frac{1-b}{b} < 1 - a$, so the lower bound for k is $k_- = 1 - a$. We then define the right-hand side of (22) as

$$\begin{aligned} g(x) =& \frac{x}{x-1} \log\big(\frac{1}{x^2}(1 - \frac{1-x}{a})\big) - \log\Big(\frac{(b-a)}{a\big(b(x+1)-1\big)}\Big) \\ =& \frac{x}{x-1}\Big(-2\log(x) + \log(a + x - 1) - \log(a)\Big) \\ & - \log(b-a) + \log(a) + \log\big(b(x+1) - 1\big) \end{aligned}$$

We then observe that on that lower bound $\lim_{x \overset{>}{\to} 1-a} g(x) = +\infty$ because $a + x - 1$ tends to zero and $x - 1$ is negative.

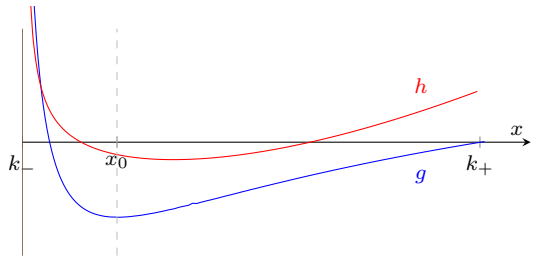

Fig. 9. Sketch of the proof that $g(x)$ has a single zero in $\mathcal{D} = (k_-, k_+)$. We first show that $\lim_{x\to k_-} g(x) = \infty$, that $\lim_{x\to k_+} g(x) = 0$ and $\lim_{x\to k_+} g'(x) > 0$, so that g must cross the x-axis on $\mathcal{D}$. To show that it only does it once, we consider a function $h(x)$ that has the same sign as $g(x)$ when $g'(x) = 0$. We show that h is convex on $\mathcal{D}$ and thus g cannot have a negative extrema, followed by a positive extrema, followed by a negative extrema. Hence it cannot have more than one zero on $\mathcal{D}$.

On the other hand, for the upper bound $k_+ = \frac{1-a}{a}$, we have

$$a + x - 1 = \frac{a^2 + 1 - a - a}{a} = \frac{(1-a)^2}{a}$$

and

$$b(x+1) - 1 = \frac{b}{a} - 1 = \frac{b-a}{a}$$

we can deduce that the upper bound of x is a zero of g:

$$\begin{aligned} g(\frac{1-a}{a}) =& \frac{1-a}{1-2a}\Big(-2\log(1-a) + 2\log(a) \\ &+ 2\log(1-a) - 2\log(a)\Big) - \log(b-a) \\ &+ \log(a) + \log(b-a) - \log(a) = 0. \end{aligned}$$

Moreover, the derivative of g is given by:

$$\begin{aligned} g'(x) =& -\frac{1}{(x-1)^2}\big(-2\log(x) + \log(a+x-1) - \log(a)\big) \\ &+ \frac{x}{x-1}\big(-\frac{2}{x} + \frac{1}{a+x-1}\big) + \frac{b}{b(x+1)-1}. \end{aligned} \tag{35}$$

Then

$$\begin{aligned} g'(\frac{1-a}{a}) =& \frac{-a^2}{(1-2a)^2}\big(-2\log(1-a)) + 2\log(a) \\ &+ 2\log(1-a) - 2\log(a)\big) \\ &+ \frac{1-a}{1-2a}\big(\frac{-2a}{1-a} + \frac{a}{(1-a)^2}\big) + \frac{ba}{b-a} \\ =& \frac{1}{1-2a}\frac{2a-1}{1-a} + \frac{ba}{b-a} \\ =& -\frac{a}{1-a} + \frac{ba}{b-a} \\ =& \frac{a^2(1-b)}{(b-a)(1-a)} > 0 \quad \text{if } 0 < a < b < 1. \end{aligned} \tag{36}$$

Since $g(k_+)$ reaches zero from below, while $g(k_-) > 0$, we can infer that $g(x)$ has a zero between k_- and k_+ as illustrated on Fig 9.

To show that this zero is unique, we look at the sign of $g'(x)$.

We can rewrite $g'(x)$ as

$$g'(x) = \frac{1}{(x-1)^2}\big(A(x) - B(x)\big) \tag{37}$$

where

$$A(x) = \frac{(1-x)\big(x(ab+b-1) - (3b-2)(1-a)\big)}{(x+a-1)(bx+b-1)}\Big) \tag{38}$$

$$B(x) = \log\big(\frac{a+x-1}{ax^2}\big) \tag{39}$$

Let x_0 be a zero of g', i.e., the position of a local extrema of g. We have

$$\begin{aligned} g(x_0) =& \frac{x_0}{x_0-1}\log\big(\frac{x_0+a-1}{ax_0^2}\big) - \log\Big(\frac{(b-a)}{a\big(b(x_0+1)-1\big)}\Big) \\ =& \frac{-x_0\big(x_0(ab+b-1) - (3b-2)(1-a)\big)}{(x_0+a-1)(bx_0+b-1)} \\ &- \log\Big(\frac{(b-a)}{a\big(b(x_0+1)-1\big)}\Big) \end{aligned} \tag{40}$$

The second equality holds because $g'(x_0) = 0$ by definition of x_0. By multiplying (40) by (bx_0+b-1), which is positive, we can then define a new function $h(x)$ whose sign is the same as the sign of $g(x)$ for $x = x_0$ (see Fig 9 for an illustration).

$$h(x) = C(x) - D(x)\,, \tag{41}$$

where

$$C(x) = \frac{-x\big(x(ab+b-1) - (3b-2)(1-a)\big)}{(x+a-1)} \tag{42}$$

$$D(x) = (bx+b-1)\log\Big(\frac{(b-a)}{a\big(bx+b-1\big)}\Big) \tag{43}$$

We can now compute the second derivatives of $C(x)$ and $D(x)$.

$$\begin{aligned} C'(x) =& -\frac{\big(2x(ab+b-1) - (3b-2)(1-a)\big)(x+a-1)}{(x+a-1)^2} \\ &+ \frac{x^2(ab+b-1) - x(3b-2)(1-a)}{(x+a-1)^2} \\ =& -\frac{x^2(ab+b-1) - 2x(1-a)(ab+b-1) + (3b-2)(1-a)^2}{(x+a-1)^2} \\ C''(x) =& -\frac{\big(2x(ab+b-1) - 2(1-a)(ab+b-1)\big)(x+a-1)}{(x+a-1)^3} + \\ &\frac{2\big(x^2(ab+b-1) - 2x(1-a)(ab+b-1) + (3b-2)(1-a)^2\big)}{(x+a-1)^3} \\ =& -\frac{2(1-a)^2(ab-2b+1)}{(x+a-1)^3} > 0 \quad \forall b > \frac{1}{2-a} \\ D'(x) =& b\log\big(\frac{b-a}{a}\big) - b\big(\log(bx+b-1) + 1\big) \\ D''(x) =& -\frac{b^2}{bx+b-1} < 0 \quad \forall x > k_- \end{aligned}$$

Hence

$$h''(x) = C''(x) - D''(x) < 0 \quad \forall x \in \mathcal{D}, \forall b > \frac{1}{2-a} \tag{44}$$

This means that h is convex, so there cannot be three points $x_1 < x_2 < x_3$ such that $0 > h(x_1) < h(x_2) > 0 > h(x_3)$. Hence the same can be said of three zeros of g', so $g(x)$ cannot have more that one zero.□