On the optimal design of metabolic RNA labeling experiments

Alexey Uvarovskii[1,2], Isabel S. Naarmann-de Vries[3,†] and Christoph Dieterich[1,2,†#]

1: Section of Bioinformatics and Systems Cardiology, Klaus Tschira Institute for Integrative Computational Cardiology and Department of Internal Medicine III, University Hospital Heidelberg
2: German Center for Cardiovascular Research (DZHK)
3: Department of Intensive Care Medicine, University Hospital Aachen, RWTH Aachen University
Correspondence to alexey.mipt@gmail.com and christoph.dieterich@uni-heidelberg.de
†: equal contribution #: Lead contact

September 26, 2018

Summary

Massively parallel RNA sequencing (RNA-seq) in combination with metabolic labeling has become the *de facto* standard approach to study alterations in RNA transcription, processing or decay. Regardless of advances in the experimental protocols and techniques, every experimentalist needs to specify the key aspects of experimental design: For example, which protocol should be used (biochemical separation vs. nucleotide conversion) and what is the optimal labeling time? In this work, we provide approximate answers to these questions using asymptotic theory of optimal design. Specifically, we derive the optimal labeling time for any given degradation rate and show that sub-optimal time points yield better rate estimates if they precede the optimal time point. Subsequently, we show that an increase in sample numbers should be preferred over an increase in sequencing depth. Lastly, we provide some guidance on use cases when laborious biochemical separation outcompetes recent nucleotide conversion based methods (such as SLAMseq).

Keywords

Metabolic labeling, Kinetic model, RNA expression, synthesis, decay, SLAMseq, TUC-seq, Timelapse-seq

Introduction

Changes in gene expression are frequently observed in pathological conditions. In the simplest model [Schwanhäusser et al., 2011], steady state RNA levels are governed by synthesis (transcription) and degradation rates (RNA stability). A paradigm is the generation of the hypoxic response in pathological conditions such as heart insufficiency [Ziaeian and Fonarow, 2016] and fast growing tumors [Wilson and Hay, 2011]. Hypoxia ($< 2\%$ O_2) results in a global decrease of total transcription [Johnson et al., 2008]. However, transcription of specific target genes is induced under hypoxic conditions by hypoxia inducible factor 1 (HIF1) [Semenza, 2003], which is composed of a stable β-subunit and an oxygen labile α-subunit [Huang et al., 1998]. Furthermore, different RNA binding proteins such as HuR and TTP as well as miRNAs regulate the stability of their cognate target mRNAs dependent on oxygen availability [Gorospe et al., 2011] and contribute to changes in gene expression profiles.

Metabolic labeling experiments are a versatile tool to discern dynamic aspects of biological processes. These experiments drive our understanding of key processes in molecular systems, such as synthesis and decay of metabolites, DNA, RNA and proteins. Pulse-chase experiments help to determine the kinetic parameters of synthesis and decay in various contexts. In the pulse phase of an experiment, the label is introduced to newly synthesized compounds and unlabeled or pre-existing molecules are only subjected to degradation or some other form of processing. In contrast, during the chase phase, the label in the system is gradually replaced by unlabeled compounds. A typical metabolic labeling experiment may include a pulse, a chase or both phases.

The first transcriptome-wide studies by Cleary et al. [2005] and Dölken et al. [2008] used 4-thiouridine (4sU) labeling in cell culture experiments to infer kinetic parameters. This approach has become quite popular in RNA biology, which is shown by a vastly increasing number of studies (see Wachutka and Gagneur [2017] for review).

Massively parallel RNA sequencing (RNA-seq) in combination with metabolic labeling has become the *de facto* standard approach to study alterations in RNA transcription, processing or decay at the transcriptome-wide level. At the time of writing, the most widely used approach involves metabolic labeling with thiol-labeled nucleoside analogs such as 4sU (4sU-tagging) [Baptista and Dölken, 2018]. Briefly, total cellular RNA is isolated and thiol groups are biotinylated. Subsequently, total cellular RNA can be efficiently separated into newly transcribed (labeled) and pre-existing (unlabeled) RNA.

A very recent achievement is the arrival of new methods involving chemical conversion of 4sU residues into cytosine analogs, which is observed as point mutations in RNA-seq data (T-to-C transitions), (see Herzog et al. [2017], Schofield et al. [2018] and Riml et al. [2017]). The absence of any biochemical separation method makes metabolic labeling more accessible due to lower input amounts and less laborious protocols.

Regardless of advances in the experimental protocols and techniques, a few important questions remain to be answered by any experimentalist, namely the specifics of experimental design: What should be measured (i.e. sequenced) and when? For example, which approach should I take (e.g. biochemical separation vs. nucleotide conversion), when should I collect my samples (e.g. time points in a pulse experiment) and how could this affect my estimates on kinetic parameters.

Within this manuscript, we use kinetic and statistical models to infer degradation rates from a pulse experiment (see Figure 1 and Equations 1-2), and derive several aspects on the optimal design of metabolic RNA labeling experiments.

Results

Model of experiment

We describe RNA-seq read counts with the negative binomial distribution, which is widely used in this setting and accounts for overdispersion [Anders and Huber, 2012]. For a given gene, read count follows $X \sim NB(m(\mu, d, t), k)$, where m is the mean read count, which depends on the time of labeling t, the degradation rate d and the expression level in the steady-state μ, and k is the overdispersion parameter of the negative binomial distribution NB. In this case, variance $\mathrm{var}(X) = m(m + k)/k$, where low k values correspond to high overdispersion in the data.

We describe RNA amount in the labeling experiments using simple first order kinetics:

$$\frac{\mathrm{d}m}{\mathrm{d}t} = s - dm, \tag{1}$$

where s is the synthesis rate and d is the degradation rate. In a steady-state, the expression level of a gene is $\mu = s/d$. The expression level μ can be derived from the total fraction, which ensures identifiability of at least this parameter. For that reason, we use μ and d to parametrize the model. In this manuscript, we only discuss the case of pulse labeling experiments throughout. However, our considerations extend to chase labeling experiments, where equations are the same, except that the labeled fraction behaves as the unlabeled one in the pulse experiment and *vice versa*. For simplicity, we assume that fraction cross-contamination is negligible, in which case, RNA amounts for a given gene are proportional to the means m_L, m_U and m_T derived from the kinetics for labeled, unlabeled and total fractions scaled by sample-specific factors x_i:

$$\begin{aligned} m_\mathrm{T}(t) &= 1 \cdot \mu \\ m_\mathrm{L}(t) &= x_\mathrm{L}\mu(1 - e^{-dt}) \\ m_\mathrm{U}(t) &= x_\mathrm{U}\mu e^{-dt} \end{aligned} \tag{2}$$

Here we treat mean read count in the total sample as a reference (coefficient is 1), to make the system identifiable. In the case of labeled and unlabeled fractions, expected read numbers must be scaled by additional coefficients, x_{U} and x_{L}, because RNA material can be normalized by different degrees during library preparation from chemically separated fractions.

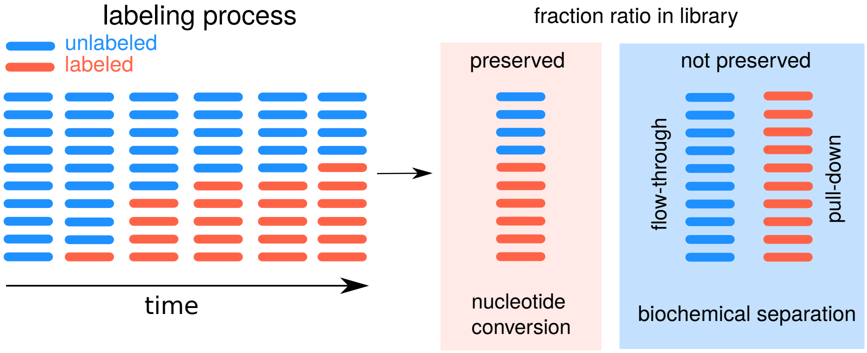

Figure 1: **Pulse labeling experiment types to measure degradation rates.** The conventional approach as in Duffy et al. [2015] utilizes biochemical separation, which does not preserve the fraction ratio (labeled vs. unlabeled) in the read counts. Alternative novel approaches (e.g. Herzog et al. [2017]) induce reverse transcription signature events (nucleotide conversions, typically T-to-C). Individual reads can be classified by the presence or absence of this characteristic nucleotide conversions. For this particular case, the fraction ratio is well reflected by the read counts.

Certain experimental approaches preserve the ratio of labeled to unlabeled fractions (see Figure 1; e.g. SLAMseq), $x_{\mathrm{U}} = x_{\mathrm{L}}$, because they do not involve any biochemical purification step. If the sequencing depth is approximately the same for all samples, we may assume for simplicity $x_{\mathrm{U}} = x_{\mathrm{L}} = 1$, and in this case, $m_{\mathrm{T}} = m_{\mathrm{L}} + m_{\mathrm{U}} = \mu$.

In the conventional approach, where labeled and unlabeled molecules are separated, $x_{\mathrm{U}} \neq x_{\mathrm{L}}$, and the fraction ratio must be inferred from the data itself or by using an external normalization by spiking in labeled and unlabeled known molecules. In the presence of cross-contamination, the estimations for the rates are biased depending on the relation of the labeling time and the degradation rate: if $dt \ll 1$ (slow rate), the bias is towards faster rate values, and, if $dt \gg 1$ (fast rate), it is towards slower rate values, see Extended methods for more details.

Best time to measure

To infer the values of model parameters $\boldsymbol{\theta}$ (where elements of the vector θ_i correspond to μ, d etc.), we maximize the likelihood function $\mathcal{L}(\boldsymbol{\theta}, X)$ given read counts X.

In this paper, we derive our results on the basis of the asymptotic theory, when the number of experiment repetitions $n \to \infty$, in which case the system can be treated analytically. However, it provides only an approximation, which depends on how close the log-likelihood is to its quadratic approximation. In this case, for $n \to \infty$ and under a number of regularity conditions, the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}$ is consistent and its covariance matrix can be approximated by the expected Fisher information matrix (FIM)

$$\mathcal{I}_{\mathrm{ij}} = -\mathbb{E}\frac{\partial^2 \log \mathcal{L}(\boldsymbol{\theta}, X)}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j}, \tag{3}$$

then distribution of $\hat{\boldsymbol{\theta}}$ is asymptotically normal:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \sim \mathcal{N}\left(0, \frac{1}{\mathcal{I}_1}\right), \tag{4}$$

where $\mathcal{I}_1$ corresponds to the FIM for a single experiment repetition [Chernoff, 1953, Pawitan, 2001]. We assume that the overdisperion parameter k is shared between all genes and neglect the uncertainty in d propagating from k, i.e. only two parameters, d and μ, are used to construct the FIM:

$$\mathcal{I} = \begin{pmatrix} \mathcal{I}_{\mathrm{dd}} & \mathcal{I}_{\mathrm{d}\mu} \\ \mathcal{I}_{\mathrm{d}\mu} & \mathcal{I}_{\mu\mu}. \end{pmatrix} \tag{5}$$

The FIM is additive, i.e. if $\mathcal{I}_\mathrm{U}$ and $\mathcal{I}_\mathrm{L}$ correspond to the labeled and unlabeled fractions, the total FIM for the experiment is $\mathcal{I} = \mathcal{I}_\mathrm{U} + \mathcal{I}_\mathrm{L}$

The diagonal terms of the inverse FIM estimate the variance of $\hat{\theta}_i$

$$\mathrm{var}(\hat{\theta}_i) = (\mathcal{I}^{-1})_{\mathrm{ii}}. \tag{6}$$

Since the FIM $\mathcal{I}$ depends on the experiment parameters, such as labeling time t and sequencing depth, it is our main interest is to reduce variance of the MLE by selecting the optimal conditions accordingly.

Noting that $\mathcal{I} = n\mathcal{I}_1$, where $\mathcal{I}_1$ is the FIM for a single measurement, we will omit n and will optimize the FIM for $n = 1$ instead.

In the case of multiple parameters, it may be not possible to achieve minimal variance for all parameters at the same time. Different criteria can be constructed as a combination of the elements of the inverse FIM [Chernoff, 1953, Van den Bos, 2007]. We are interested to optimize estimation of d only and do not consider variance of the expression level estimator $\hat{\mu}$ in the design criteria.

Let us consider first a simpler experimental setup, which preserves the fraction ratio (e.g. SLAMseq). The derivations for a general case are left to the Extended Methods section, and here we first discuss the case of the Poisson model, which corresponds to the case of no overdispersion ($k \to \infty$). Let X_L and X_U be the read counts corresponding to the labeled and unlabeled molecules for a given gene in a SLAM-seq sample, and let t be the time of labeling. In this case, the inverse FIM is diagonal:

$$\mathcal{I}^{-1}_{\mathrm{slam}} = (\mathcal{I}_\mathrm{L} + \mathcal{I}_\mathrm{U})^{-1} = \begin{pmatrix} \dfrac{e^{dt}-1}{\mu t^2} & 0 \\ 0 & \mu \end{pmatrix} \tag{7}$$

The parameters d and μ are information orthogonal, because $\mathcal{I}_{\mathrm{d}\mu} = 0$ and inference about d can be done as μ were known exactly.

Indeed, for $X_\mathrm{L} \sim \mathrm{Pois}(m_\mathrm{L}), X_\mathrm{U} \sim \mathrm{Pois}(m_\mathrm{U})$, the conditional distributions $P(X_\mathrm{L}|X_\mathrm{U} + X_\mathrm{L})$ and $P(X_\mathrm{U}|X_\mathrm{U} + X_\mathrm{L})$ are binomial with the rates $m_\mathrm{U}/(m_\mathrm{U} + m_\mathrm{L}) = e^{-dt}$ and $m_\mathrm{L}/(m_\mathrm{U} + m_\mathrm{L}) = 1 - e^{-dt}$ and do not depend on μ. This model was recently discussed in a Bayesian framework for SLAMseq experiments by [Jürges et al., 2018].

For a diagonal $\mathcal{I}$, the inverse term $\left(\mathcal{I}^{-1}_{\mathrm{slam}}\right)_{\mathrm{dd}} = ((\mathcal{I}_{\mathrm{slam}})_{\mathrm{dd}})^{-1} = ((\mathcal{I}_\mathrm{U})_{\mathrm{dd}} + (\mathcal{I}_\mathrm{L})_{\mathrm{dd}})^{-1}$. The maximum of the term $(\mathcal{I}_{\mathrm{slam}})_{\mathrm{dd}}$ corresponds to minimal asymptotic variance of $\hat{d}$ due to Equation 4. By optimizing $(\mathcal{I}_{\mathrm{slam}})_{\mathrm{dd}}$ in respect to t, we get

$$t_{\mathrm{slam}} = 1.59\tau, \tag{8}$$

where $\tau = 1/d$ is the characteristic time of degradation. That means, that if one optimizes the SLAMseq experiment and is concerned with characteristic time of degradation of τ, the measurement at time point 1.59τ corresponds to the asymptotically optimal design.

In Figure 2A, we depicted dependency of $(\mathcal{I}_{\mathrm{slam}})_{\mathrm{dd}}$ and corresponding values of $(\mathcal{I}_\mathrm{U})_{\mathrm{dd}}$ and $(\mathcal{I}_\mathrm{L})_{\mathrm{dd}}$ as functions of normalized time t/τ for the degradation rate $d = 1$.

Interestingly, $(\mathcal{I}_\mathrm{U})_{\mathrm{dd}}$ and $(\mathcal{I}_\mathrm{L})_{\mathrm{dd}}$ achieve maximum at $t_U = 2\tau$ and $t_L = 0.64\tau$, and the main contribution to the sum $(\mathcal{I}_{\mathrm{slam}})_{\mathrm{dd}} = (\mathcal{I}_\mathrm{U})_{\mathrm{dd}} + (\mathcal{I}_\mathrm{L})_{\mathrm{dd}}$ comes from the term corresponding to labeled counts at shorter labeling times, and from the term for unlabeled counts at times longer than τ, see Figure 2A.

Too early is better than too late

Usually one is interested to measure a rate with a certain relative precision. To reflect this, we normalize the variance of the estimator by d^2:

$$\frac{\mathrm{var}(\hat{d})}{d^2} \approx \frac{1}{\mathcal{I}_{\mathrm{dd}} d^2}, \tag{9}$$

Technically, such modification corresponds to the transformation $d = e^\eta$, because $\mathcal{I}(\eta) = \mathcal{I}(d)\left|d(\eta)'_\eta\right|^2 = \mathcal{I}(d)d^2$, as if the degradation rate were considered at the logarithmic scale, and any absolute changes in η would correspond to relative changes in d. However, we avoid introducing additional parameters and stick to the usage of d only.

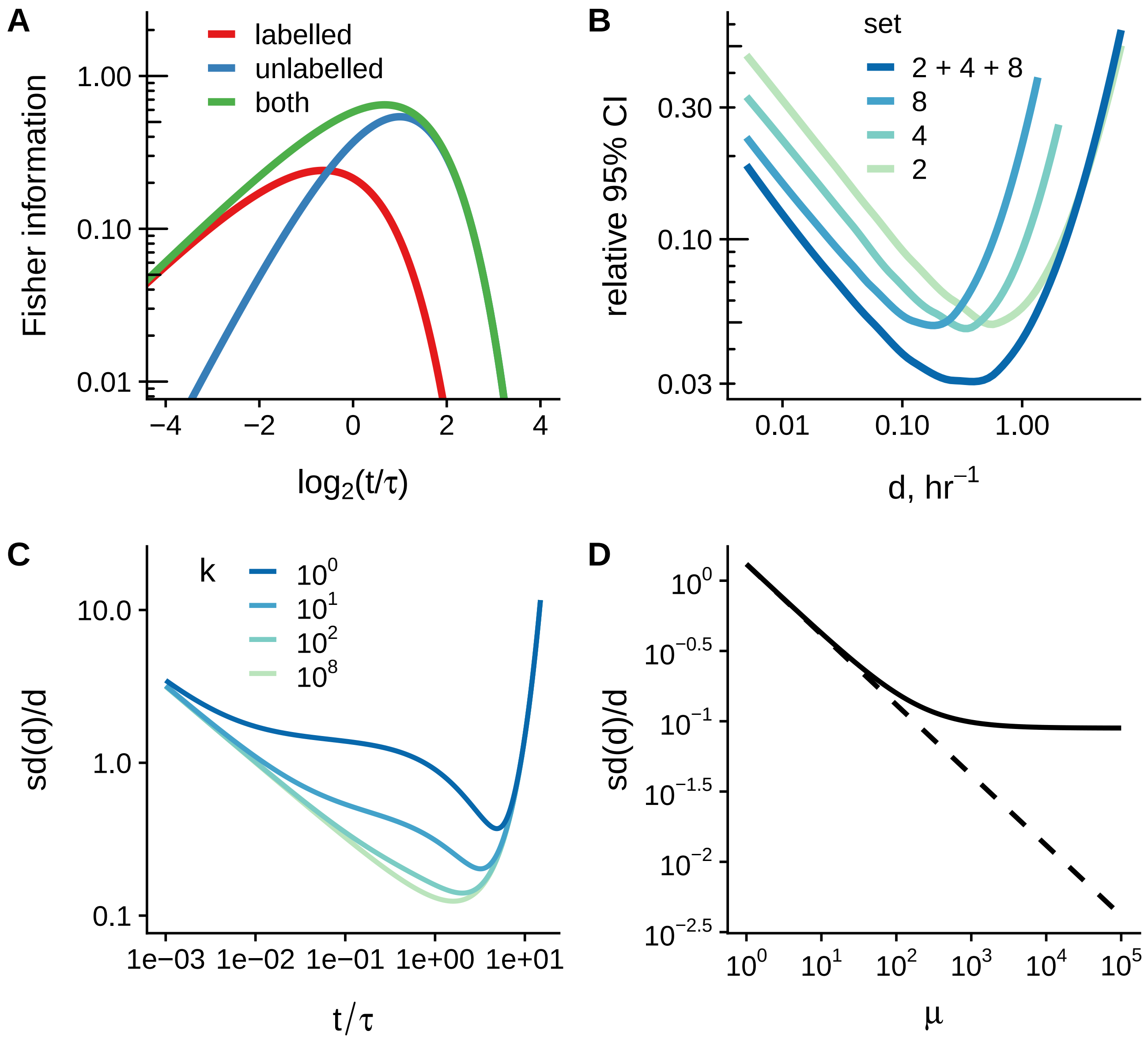

Figure 2: **Key characteristics of metabolic RNA labeling experiments A:** The diagonal term of the Fisher information matrix (FIM) $\mathcal{I}_{\mathrm{dd}}$, which corresponds to the degradation rate, as a function of the ratio of labeling time to the characteristic time of degradation $\tau = 1/d$ for the case of SLAMseq experiment. read counts follow the Poisson distribution. **B:** 95% confidence interval (CI) relative width of the degradation rates for different sets of time points included in the simulation of the SLAMseq experiment. We simulated counts for a range of rates d and assumed for simplicity that normalization factors are perfectly known but not the rates and expression levels. Smoothed data from 10 simulation runs is shown. **C:** Relative standard deviation ($sd(\hat{d})/d$) of the MLE estimator for d as a function of measurement time at different values of the overdispersion parameter k. With increasing overdispersion, the profile of the dependency flattens. However, near the optimal time point, variance of the estimation is more sensitive to time of labeling, which complicates the optimal design choice for different d ranges. Expression level is fixed to $\mu = 100$ reads in this example, the degradation rate is assumed to be $d = 1$. **D:** Relative standard deviation ($sd(\hat{d})/d$) for a model with overdispersion ($k = 100$, solid line) or with no overdispersion ($k \to \infty$, dashed line). The degradation rate is $d = 1$, the labeling time is $t = 1$.

For brevity we omit the index, since we consider only the *dd* diagonal term of $\mathcal{I}$ in this subsection. By introducing a non-dimensional time variable $\alpha = t/\tau$,

$$\begin{aligned} \mathcal{I}_{\mathrm{L}} d^2 &= \frac{\alpha^2 \mu}{e^{2\alpha} - e^{\alpha}} \\ \mathcal{I}_{\mathrm{U}} d^2 &= \alpha^2 e^{-\alpha} \mu \\ \mathcal{I}_{\mathrm{slam}} d^2 &= \frac{\alpha^2 \mu}{e^{\alpha} - 1} \end{aligned} \tag{10}$$

For labeling times much shorter than characteristic degradation time of a given gene, $\alpha \ll 1$, and the normalized FIM terms behave as a power function:

$$\mathcal{I}_{\mathrm{slam}} d^2, \mathcal{I}_{\mathrm{L}} d^2 \sim \alpha, \quad \mathcal{I}_{\mathrm{U}} d^2 \sim \alpha^2. \tag{11}$$

However, for labeling times much longer than the characteristic time of degradation τ, $\alpha \gg 1$, the normalized FIM terms vanish exponentially:

$$\mathcal{I}_{\mathrm{L}} d^2 \sim e^{-2\alpha}, \quad \mathcal{I}_{\mathrm{slam}} d^2, \mathcal{I}_{\mathrm{U}} d^2 \sim e^{-\alpha}. \tag{12}$$

In a typical high-throughput experiments, kinetic parameters are monitored for a large set of genes (in the order of thousands), which may have different degradation rates. In this case, every time point in the experiment will be only optimal for a subset of these genes. To illustrate this effect, we simulated read counts for a SLAMseq experiment scheme (with no over dispersion) and fitted the model using various sets of samples. In our *in silico* experiment, we always included the total fraction ($t = 0$ hr), and either one additional time point (labeled and unlabeled fractions) or all time points (2, 4, and 8 hr). The normalization coefficients were set to 1 to mimic the SLAMseq scheme, as discussed earlier, Equation 2.

We fitted the model using the `pulseR` package and computed the 95% confidence intervals (CI) for d using the profile likelihood approach [Uvarovskii and Dieterich, 2017]. Since we assume no overdispersion (Poisson distribution), for high read counts ($\mu = 10000$) the quadratic approximation of the log-likelihood function applies, and the confidence intervals for the rate estimations may be approximated by the Wald intervals, i.e. $(\hat{d} - 1.96\sqrt{(\mathcal{I}^{-1})_{\mathrm{dd}}}, \hat{d} + 1.96\sqrt{(\mathcal{I}^{-1})_{\mathrm{dd}}})$, and hence, they reflect the behavior of the FIM term for d. As expected, the relative CI width is minimal only for a certain subset of the rates, depending on the set of measurements included, see (Figure 2B).

If the degradation rate is very fast in comparison to the experiment time scale, the CI width for these fast genes is defined by the earliest time point in the experiment (see Figure 2B).

Since every labeling time is optimal only for a single degradation rate, it might be beneficial to focus the design on genes with faster rates d, if sample size is limited and no other criteria of optimality are given. The justification follows from the faster decay of the FIM term for $\alpha \gg 1$ (i.e. genes with faster kinetics), Equations 11, 12.

Increasing sample numbers is preferred over higher sequencing depth

Distribution of read count from RNA-seq experiments exhibits overdispersion, and negative binomial distribution (NB) can be used to account for that [Anders and Huber, 2012]. In this section, we explore how presence of overdispersion would affect inference about d. The overdispersion parameter k of the NB distribution describes the level of overdispersion in the data, in which case the variance is defined as $\mathrm{var}(X) = m + m^2/k$ for counts $X \sim NB(m, k)$ with mean m. Smaller values of k correspond to higher overdispersion level, and, for $k \to \infty$, the NB distribution converges to the Poisson distribution, for which $\mathrm{var}(X) = m$. For simplicity, we assume that distributions of read counts in all samples share the same value of k. In addition, we do not consider uncertainty in the overdispersion parameter k when we make inference about d for individual genes, in a way as it is implemented in some packages for differential expression analysis, for example, in DESeq, [Anders and Huber, 2012]. A more advanced quasi-likelihood approach, which accounts for uncertainty in the overdispersion parameter, is discussed in Lund et al. [2012].

In the case of NB distribution, the FIM is not diagonal for the SLAM-seq experiment, see the Extended Methods section. Hence we need to work with the inverse FIM, and the diagonal term for the SLAMseq design is

$$(\mathcal{I}^{-1}_{\mathrm{slam}})_{\mathrm{dd}} = \frac{e^{dt} - 1}{\mu t^2} + \frac{2(1 - e^{-dt})^2}{k t^2}. \tag{13}$$

The presence of overdispersion shifts the optimal time to higher values. But the most important change is that the profile of $\mathcal{I}_{\mathrm{dd}}^{-1}$ is more sensitive to the labeling time t near the optimal point. For higher overdispersion values, the variance of the rate estimator $\hat{d}$ increases faster in the vicinity of the optimum (see Figure 2C). It imposes stricter conditions on the experiment design. The second term in the Equation 13 vanishes for times $t \gg 1$, and the equation coincides with the case of no overdispersion. The contribution of the second term is higher for smaller values of k (higher overdispersion) and for shorter labeling times t, with maximal value at $t \to 0$:

$$\lim_{t \to 0} \frac{2(1-e^{-dt})^2}{kt^2} = \frac{2d^2}{k}. \tag{14}$$

Another limitation, which arises in the over-dispersed model is that increase of the sequencing depth has a limiting effect on the variance. Indeed, only the first term in Equation 13 can be eliminated by increase of sequencing depth:

$$\lim_{\mu \to \infty} (\mathcal{I}_{\mathrm{slam}}^{-1})_{\mathrm{dd}} = \frac{2(1-e^{-dt})^2}{kt^2}. \tag{15}$$

In contrast, repeating the experiment n times affects both terms in $\mathcal{I}_{\mathrm{dd}}^{-1}$, since for n repetitions,

$$\mathcal{I}^{-1} = \frac{1}{n}\mathcal{I}_1^{-1}, \tag{16}$$

where $\mathcal{I}_1^{-1}$ is the inverse FIM for one repetition. That means, that it can be more beneficial to spread the sequencing capacity between several biological replicates. In the limiting case $n \to \infty$, if the total depth of n repetitions is fixed to the initial value such, that $n\mu_1 = \mu$, where μ_1 corresponds to the reduced depth in a single repetition,

$$\lim_{\substack{n\to\infty \\ \mu_1 n = \mu}} \frac{1}{n}\left(\frac{e^{dt}-1}{\mu_1 t^2} + \frac{2(1-e^{-dt})^2}{kt^2}\right) = \frac{e^{dt}-1}{\mu t^2}, \tag{17}$$

which coincides with the FIM term for the Poissonian (non-overdispersed) model, compare to Equation 7.

In the Poissonian case, when $k \to \infty$ and the second term is absent (see Equation 7), adding twice more samples or increasing sequencing depth by two fold results to the same FIM and, consequently, the same approximation of the variance $\mathrm{var}(\hat{d})$. In the logarithmic scale, the Poissonian case is represented by a line, Figure 2D (dashed line), which can be compared to the case of overdispersion (solid line), when an increase in the depth is inefficient starting from a certain value. A similar phenomenon was discussed by [Robles et al., 2012] in the context of differential gene expression analysis by RNA-seq.

Biochemical separation still matters

If one is interested in estimating rates of extreme values by using very short (e.g. TT-seq, Schwalb et al. [2016]) or long labeling times, it may be less efficient to use the protocols, which preserve the ratio of labeled and unlabeled molecules (e.g. SLAMseq). Let us consider a study of fast gene kinetics, in which case, very short labeling times are used. In this case, $dt \ll 1$ for majority of the genes, and the labeled fraction constitutes only a minor proportion of the input SLAMseq sample, because $m_{\mathrm{L}} = \mu(1-e^{-dt}) \approx \mu dt \ll 1$. The input SLAMseq sample mainly contains unlabeled molecules from the non-target slower genes, which results into spending sequencing resources on non-informative material. The same idea holds for very long times, when $dt \gg 1$ and most of unlabeled molecules are subjected to degradation, $m_{\mathrm{U}} = \mu e^{-dt} \ll 1$.

In contrast, conventional experimental setups with a separation step can be used to focus sequencing capacity on relevant molecules. However, the conventional approach suffers from the need to normalize sequencing results from different fractions as it does not preserve the ground truth ratio of labelled and unlabelled molecules. In typical RNA-seq experiments, normalization coefficients assumed to be shared between all the genes in a given sample [Anders and Huber, 2012], but nevertheless, it introduces additional uncertainty into rate estimations and further analysis is required to estimate level of this contribution. In the following derivations, we neglect the uncertainty in estimating the fraction normalization coefficients x_i from the Equation 2.

To illustrate the benefit of the conventional approach, let us consider a set of fast genes $\mathcal{F}$, such that there exists labeling time t, when majority of genes $i \notin \mathcal{F}$ do not contribute to the labeled fractions, i.e. $\mu\left(1-e^{-d_i t}\right) \ll 1$ for

$i \notin \mathcal{F}$, but $\mu\left(1 - e^{-d_i t}\right) \approx 1$ for $i \in \mathcal{F}$. If the sequencing depth of the labeled fraction is approximately the same as for the total sample, then the normalization factor is

$$x_{\mathrm{L}} = \frac{\sum_i \mu_i \left(1 - e^{-d_i t}\right)}{\sum_{i \in \mathcal{F}} \mu_i \left(1 - e^{-d_i t}\right)} \approx \frac{\sum_i \mu_i}{\sum_{i \in \mathcal{F}} \mu_i}, \tag{18}$$

which can be high at short times. Such "zooming" effect can be considered as corresponding increase of the sequencing depth in the SLAMseq experiment by the factor of x_{L} for the labeled fraction. The same idea can be applied to the unlabeled fraction and long labeling times, when the sequencing depth is shared out between the most stable set of genes. Since the normalization factor depends on the rate distribution and expression level in a given system, it is not possible to derive optimal design criteria analytically without imposing additional assumptions.

Since

$$(\mathcal{I}^{-1})_{\mathrm{dd}} = \left(\mathcal{I}_{\mathrm{dd}} - \mathcal{I}_{\mathrm{d}\mu}\mathcal{I}_{\mu\mathrm{d}}/\mathcal{I}_{\mu\mu}\right)^{-1}, \tag{19}$$

and using the fact that $\mathcal{I}_{\mathrm{d}\mu} = \mathcal{I}_{\mu\mathrm{d}}$ and $\mathcal{I}_{\mu\mu} > 0$, the diagonal term of the inverse matrix is bounded

$$\left((\mathcal{I}^{-1})_{\mathrm{dd}}\right)^{-1} \leqslant \mathcal{I}_{\mathrm{dd}}. \tag{20}$$

The diagonal element of the FIM $\mathcal{I}_{\mathrm{dd}}$ represents the upper bound of the information gain in a given design, corresponding to assumption of no uncertainty in parameters other than d. As in the case of the SLAMseq, inference can be improved to a limited extent by increase of sequencing depth, if overdispersion is present in the data, compare to Equation 15:

$$\begin{aligned} \lim_{\mu \to \infty} (\mathcal{I}_{\mathrm{L}})_{\mathrm{dd}} &= \frac{t^2 e^{-2dt} k}{\left(1 - e^{-dt}\right)^2} \leqslant \frac{k}{d^2} \\ \lim_{\mu \to \infty} (\mathcal{I}_{\mathrm{U}})_{\mathrm{dd}} &= t^2 k \end{aligned} \tag{21}$$

It is interesting to note, that for the case of the unlabeled fraction, the bound can be improved by use of longer labeling times (provided very high sequencing depth), which is not the case for the labeled fraction (with the upper bound $\mathcal{I}_{\mathrm{L}} \to k/d^2$ at $t \to 0$).

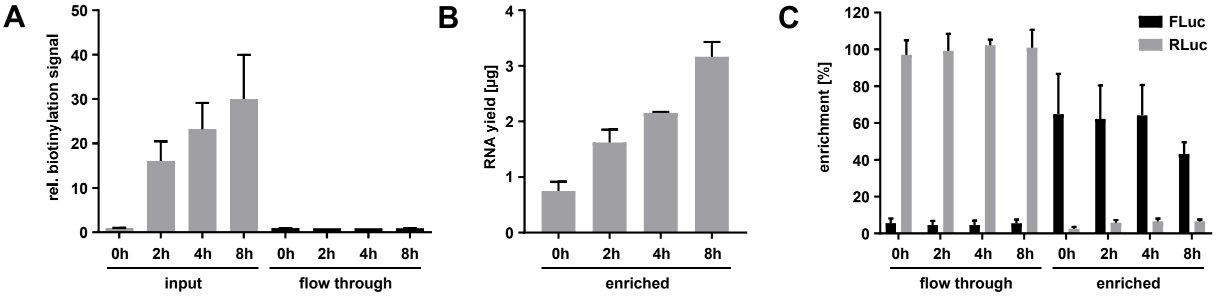

Figure 3: **Purification of labeled and unlabeled RNA fractions.** MCF-7 cells were pulse labeled with 4sU for up to eight hours as indicated. Total RNA was spiked with *in vitro* transcribed 4sU-labeled FLuc and unlabeled RLuc, biotinylated with MTSEA-biotin and subjected to streptavidin purification. (n = 3). **A:** Dot blot-based detection of biotinylation with streptavidin-HRP in input and flow through of streptavidin purification. **B:** The amount of RNA enriched by the streptavidin purification was determined by absorption measurement. **C:** *In vitro* transcribed spike in RNAs 4sU-labeled FLuc and unlabeled RLuc in the flow through and biotin-enriched fraction were measured by RT-qPCR analysis and normalized to a standard curve given in Supplemental Figure 1.

Example from a pulse labeling experiment

MCF-7 cells were pulse labeled with 200 μM 4sU for 2, 4 or 8 hrs. 4sU-labeled and unlabeled RNA were separated by streptavidin purification after MTSEA biotin-XX catalyzed biotinylation of 4sU-labeled RNA [Duffy et al., 2015]. The efficiency of purification was monitored in a dot blot assay that detects biotinylated RNA with streptavidin-HRP

(Figure 3A). This analysis revealed a gradual increase in biotinylation with increasing labeling time. Importantly, biotinylated transcripts were efficiently depleted from the flow trough, as no biotinylation signal could be detected in these samples. Biotin-enriched RNAs are eluted by de-biotinylation with DTT. Therefore, we estimated purification efficiency by the amount of purified RNA determined by $\mathrm{A}_{260\mathrm{nm}}$ absorption measurement. The amount of purified RNA increased gradually with increasing labeling time (Figure 3B) comparable to the biotinylation signal increase in the respective input fractions (Figure 3A). To determine the efficiency and specificity precisely for individual transcripts, we spiked the 4sU-labeled total RNA from MCF-7 with *in vitro* transcribed 4sU-labeled FLuc and unlabeled RLuc that were followed by RT-qPCR analysis using a standard curve for quantification (Supplemental Figure 1). This analysis revealed a purification efficiency of 4sU-labeled FLuc of about 60% (58.56). The specificity was determined by the cross-contamination of RLuc in the biotin-enriched fractions and FLuc in the flow through fractions, which was about 5% for each transcript (RLuc in enriched = 5.32%, FLuc in flow through = 5.01%).

The kinetic model was fitted to the read counts from the sequenced samples for genes with mean read count > 50 in the total samples. Two total samples were collected at 0 hr, labeled and unlabeled fractions at other time points (2, 4 and 8 hrs) in two replicates (see Supplementary Table 1). In the model fitting, we assumed no cross-contamination between fractions and shared normalization coefficients for samples originating from the same time point and fraction.

Having the estimations for expression levels, degradation rates, overdispersion parameter and normalization coefficients, we calculated the FIM diagonal elements $\mathcal{I}_{\mathrm{dd}}$ for the analyzed genes for different time points and fraction types.

In Figure 4A, the value of the diagonal FIM element multiplied by $\hat{d}^2$, i.e. $\mathcal{I}_{\mathrm{dd}}(\hat{\boldsymbol{\theta}}, t)\hat{d}^2$ (compare to Equations 9 and 10), is depicted for both fractions. As mentioned in the previous section, $\mathcal{I}_{\mathrm{dd}}$ can be interpreted as an information gain from the experiment assuming other parameters were known, which represents an upper bound, see Equation 20. In addition, these terms are bounded due to presence of overdisperion in the data, (Equation 21 and dashed lines in Figure 4A), and increase of sequencing depth can not improve these limits.

At short labeling times, the FIM term is higher for the labeled fraction than for the unlabeled one for majority of the genes, (Figure 4A, 2hr), which is a result similar to the SLAMseq case. At longer labeling times, the contribution from the unlabeled fraction increases, and $(\mathcal{I}_{\mathrm{U}})_{\mathrm{dd}} > (\mathcal{I}_{\mathrm{L}})_{\mathrm{dd}}$ for majority of the genes (Figure 4A, 8hr). However, the proportion of genes with high degradation rates d in the unlabeled sample exponentially decreases, since

$$\lim_{t\to\infty} \frac{\mu_{\mathrm{fast}} e^{-d_{\mathrm{fast}} t}}{\mu_{\mathrm{slow}} e^{-d_{\mathrm{slow}} t}} = \lim_{t\to\infty} \frac{\mu_{\mathrm{fast}}}{\mu_{\mathrm{slow}}} e^{-(d_{\mathrm{fast}} - d_{\mathrm{slow}})t} = 0. \tag{22}$$

It results in very low counts and decrease in the $\mathcal{I}_{\mathrm{U}}$ for these fast genes, see Figure 4A, 8hr, reduced values at the right tail of the distribution (blue dots).

Optimal design for such experiments is complicated by the fact, that it depends not only on the degradation rates of some target genes, but on the overall rate distribution in the system being studied. We illustrate a dependency of the $(\mathcal{I})_{\mathrm{dd}}\, d^2$ terms on labeling time for one of the fastest (0.1% quantile) and one of the slowest (99.9% quantile) genes. The normalization coefficients for labeled and unlabeled fractions were adjusted in the same manner as in the previous section, i.e. at every time point t the sequencing depth equals the sequencing depth of the total sample. In the case of low or no overdispersion, use of labeled fraction and shorter labeling times is preferred for estimation of fast genes, and, *vice versa*, longer labeling times and unlabeled fraction is preferred for slow genes, see Figure 4B, dashed lines.

At high values of overdispersion (i.e. low k), the FIM term is bounded $(\mathcal{I}_{\mathrm{L}})_{\mathrm{dd}} d^2 < k$ due to Equation 21. In this case, there may exist values of labeling times at which the terms from the unlabeled fraction $(\mathcal{I}_{\mathrm{U}})_{\mathrm{dd}} d^2$ is larger than maximal $(\mathcal{I}_{\mathrm{L}})_{\mathrm{dd}} d^2$ value, Figure 4B, solid lines. As a protection against such situation in the case of fast genes, use of samples from unlabeled fraction may be a solution.

Although one may have a prior guess about the range of degradation rates in a system, it is unlikely, that there is information about the distribution of the rates and overdispersion level. Hence, such design suggestions are possible only in sequential approach, when an exploratory experiment is done first.

Discussion

In this study, we discuss some aspects of the optimal design of RNA labeling experiments using the results of the asymptotic theory. First, we show that there exists an optimal time point for which the maximum likelihood

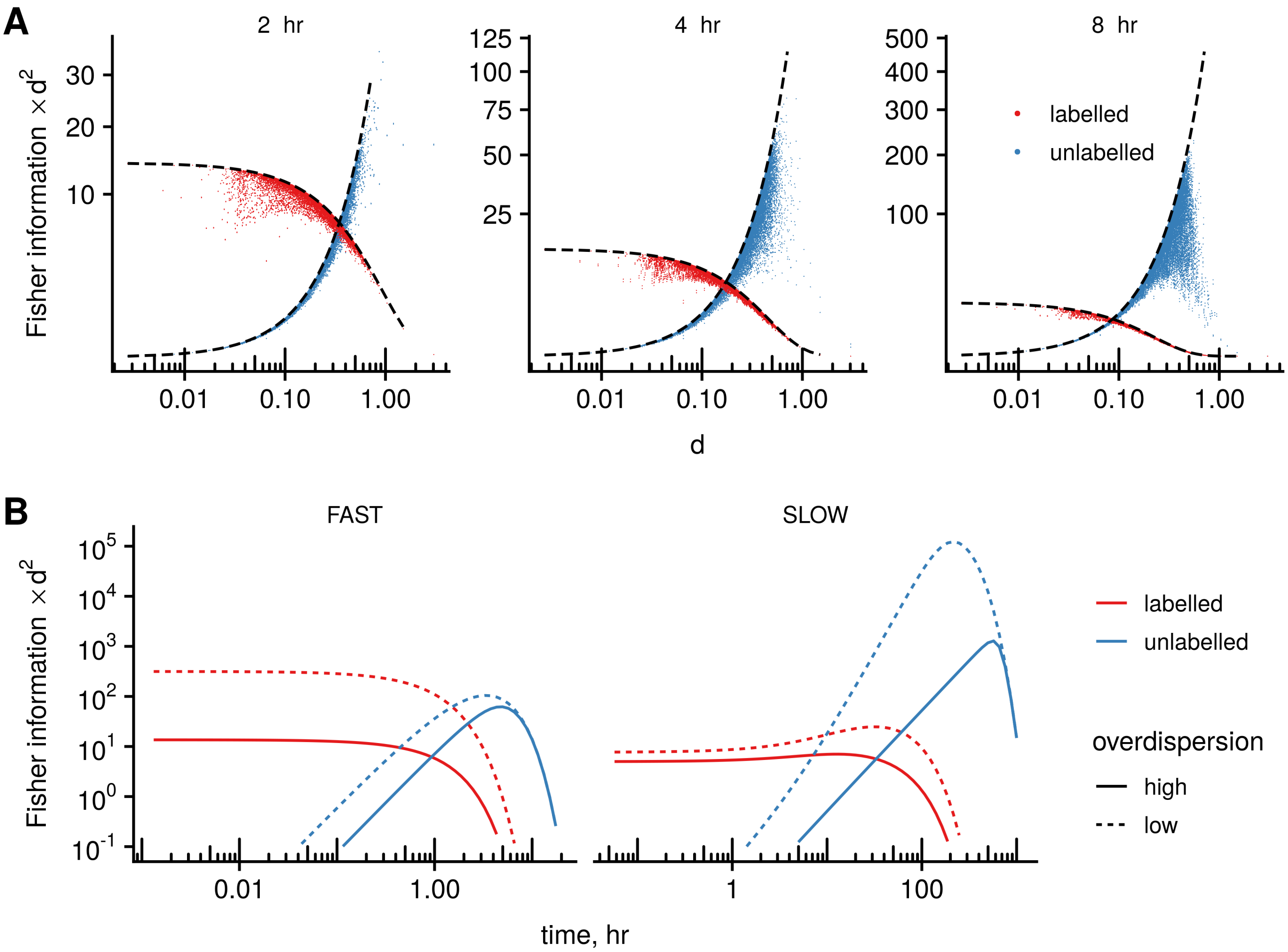

Figure 4: **Application to experimental data from the MCF-7 pulse labeling time course experiment A:** We plot the diagonal term of the FIM computated at estimated parameter values and multiplied by $\hat{d}^2$, $\mathcal{I}_{\rm dd}(\hat{\boldsymbol{\theta}}, t)\hat{d}^2$, to illustrate contributions from labeled and unlabeled fractions to estimations of degradation rates for different experimental points (MCF7 experiment, 2, 4, and 8 hr) and fractions (labeled and unlabeled). The black lines are the limiting values for the $\mathcal{I}_{\rm dd}$ according to Equation 21. **B:** The modified FIM term $\mathcal{I}_{\rm dd}(\hat{\boldsymbol{\theta}}, t)\hat{d}^2$ is computed for a range of labeling times for one of the fastest (at the 0.1% quantile) and one of the slowest (at the 99,9% quantile) genes ($d_{\rm fast} = 0.79\,{\rm hr}^{-1}$, $d_{\rm slow} = 0.019\,{\rm hr}^{-1}$). The normalization coefficient is adjusted in such a way, that sequencing depth (total mean read count) at time t equals the sequencing depth of the total sample.

estimator possess a minimal variance asymptotically. This first result was developed for the case of experiments, which preserve the fraction ratio and hence do not require normalization between fractions (e.g. SLAMseq, TUC-seq, TimeLapse-seq)

In the case of negligible overdispersion, the optimal labeling time for a gene with the characteristic degradation time τ is $t_{\text{slam}} = 1.59\tau$, and smaller labeling times show better rate estimates in comparison to longer times: the variance increases exponentially for times longer than τ and only by a power law for shorter labeling times. This result may explain the observations of a simulation study by Jürges et al. [2018]. Herein, for a given gene with $\tau = 2$ hr, the most precise estimation was at labeling time 3 hr ($t_{\text{optimal}} = 1.59 \cdot 2 = 3.18$ hr), and the worst estimation was observed at the longest time (12 hr).

Moreover, we show that at short labeling times (in comparison to the characteristic time of degradation for a given gene), the labeled fraction contributes most to the Fisher information term corresponding to the degradation rate, and, *vice versa*, at long times the highest contribution is seen for the unlabeled one.

In addition, we show that in the presence of overdispersion, the variance of rate estimates is more sensitive to choices of labeling times different from the optimal, which make it more difficult to optimize conditions for a range of rates. The overdispersion imposes a bound on the asymptotic relative standard deviation for the estimator of the rate ($sd(d)/d$, see Figure 2C), and, from a certain level, increase in sequencing depth is very inefficient (Figure 2D).

Moreover, we discuss possible benefits of use of the conventional experimental approach, especially for estimation of extreme degradation rates, which deviate highly from the general pool. For nucleotide conversion setups with too short or too long labeling times, the majority of reads in a sample originate from the unlabeled or labeled fractions correspondingly. In contrast, the conventional scheme, which involves biochemical fraction separation, allows to concentrate experimental costs only on the relevant material.

Obviously, there are certain limitations to our study. First, the method involving FIM calculation describes only the asymptotic behavior of the estimator. Hence, all the conclusions are only approximate, since we do not investigate the behavior of the likelihood function itself, but only the quadratic approximation of its logarithm using the FIM.

Secondly, we do not consider uncertainty from the shared parameters, such as the overdispersion parameter of the negative binomial distribution and the normalization coefficients for the fractions. Inference on these parameters is based on the whole pool of the genes, and would involve more complex analytic treatment and assumptions on the distribution of rates.

Lastly, cross-contamination between fractions is a highly relevant problem for inference, especially in the absence of external reference molecules (spike ins), which are typically used to assess this phenomenon. However, in the Extended methods section, we show that cross-contamination shifts estimations of fast rates to slower values, and slow rates towards faster values. Previously, Eser et al. [2016] included a global transcriptome-wide cross-contamination term to presented kinetic model, yet future work is needed to assess possible effect sizes on rate estimations.

We hope that our work will encourage further development of the methodology to address the discussed limitations and to improve suggestions on design of metabolic labeling experiments.

Acknowledgments

The authors thank Tobias Jakobi for computational infrastructure support and Etienne Boileau for providing useful comments on the paper draft. We are grateful for the excellent sequencing support by the Cologne Centre for Genomics. IND would like to acknowledge funding by the DFG (NA 1273/1-1). AU and CD acknowledge funding by the Klaus Tschira Stiftung gGmbH and DFG SPP 1784.

Author Contributions

AU designed and performed the research. CD supervised the project. IND designed and performed the MCF7 experiments. AU, IND, CD analyzed the data. AU, IND and CD wrote the manuscript.

Declaration of Interests

The authors declare no competing interests.

References

Simon Anders and Wolfgang Huber. Differential expression of RNA-Seq data at the gene level–the deseq package. *Heidelberg, Germany: European Molecular Biology Laboratory (EMBL)*, 2012.

Marisa A P Baptista and Lars Dölken. Rna dynamics revealed by metabolic rna labeling and biochemical nucleoside conversions. *Nature methods*, 15:171–172, February 2018. ISSN 1548-7105. doi: 10.1038/nmeth.4608.

Herman Chernoff. Locally optimal designs for estimating parameters. *The Annals of Mathematical Statistics*, pages 586–602, 1953.

Michael D Cleary, Christopher D Meiering, Eric Jan, Rebecca Guymon, and John C Boothroyd. Biosynthetic labeling of rna with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mrna synthesis and decay. *Nature biotechnology*, 23:232–237, February 2005. ISSN 1087-0156. doi: 10.1038/nbt1061.

Sebastian de Vries, Isabel S Naarmann-de Vries, Henning Urlaub, Hongqi Lue, Jurgen Bernhagen, Dirk H Ostareck, and Antje Ostareck-Lederer. Identification of ddx6 as a cellular modulator of vegf expression under hypoxia. *Journal of biological chemistry*, pages jbc–M112, 2013.

Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics (Oxford, England)*, 29:15–21, January 2013. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts635.

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.

Erin E Duffy, Michael Rutenberg-Schoenberg, Catherine D Stark, Robert R Kitchen, Mark B Gerstein, and Matthew D Simon. Tracking distinct rna populations using efficient and reversible covalent chemistry. *Molecular cell*, 59:858–866, September 2015. ISSN 1097-4164. doi: 10.1016/j.molcel.2015.07.023.

Philipp Eser, Leonhard Wachutka, Kerstin C Maier, Carina Demel, Mariana Boroni, Srignanakshi Iyer, Patrick Cramer, and Julien Gagneur. Determinants of rna metabolism in the schizosaccharomyces pombe genome. *Molecular systems biology*, 12(2):857, 2016.

Myriam Gorospe, Kumiko Tominaga, Xue Wu, Michael Fähling, and Mircea Ivan. Post-transcriptional control of the hypoxic response by rna-binding proteins and micrornas. *Frontiers in molecular neuroscience*, 4:7, 2011. ISSN 1662-5099. doi: 10.3389/fnmol.2011.00007.

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*, 14:1198–1204, December 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4435.

L E Huang, J Gu, M Schau, and H F Bunn. Regulation of hypoxia-inducible factor 1alpha is mediated by an o2-dependent degradation domain via the ubiquitin-proteasome pathway. *Proceedings of the National Academy of Sciences of the United States of America*, 95:7987–7992, July 1998. ISSN 0027-8424.

Amber Buescher Johnson, Nicholas Denko, and Michelle Craig Barton. Hypoxia induces a novel signature of chromatin modifications and global repression of transcription. *Mutation research*, 640:174–179, April 2008. ISSN 0027-5107. doi: 10.1016/j.mrfmmm.2008.01.001.

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

Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9:357–359, March 2012. ISSN 1548-7105. doi: 10.1038/nmeth.1923.

Anke Liepelt, Jana C Mossanen, Bernd Denecke, Felix Heymann, Rebecca De Santis, Frank Tacke, Gernot Marx, Dirk H Ostareck, and Antje Ostareck-Lederer. Translation control of tak1 mrna by hnrnp k modulates lps-induced macrophage activation. *Rna*, 2014.

Steven P Lund, Dan Nettleton, Davis J McCarthy, and Gordon K Smyth. Detecting differential expression in rna-sequence data using quasi-likelihood with shrunken dispersion estimates. *Statistical applications in genetics and molecular biology*, 11(5), 2012.

Isabel S Naarmann-de Vries, Annika Brendle, Tomi Bähr-Ivacevic, Vladimir Benes, Dirk H Ostareck, and Antje Ostareck-Lederer. Hnrnp k-mediated translational control links nmhc iia to erythroid enucleation. *J Cell Sci*, pages jcs–174995, 2016.

Yudi Pawitan. *In all likelihood: statistical modelling and inference using likelihood.* Oxford University Press, 2001.

Mihaela Pertea, Geo M Pertea, Corina M Antonescu, Tsung-Cheng Chang, Joshua T Mendell, and Steven L Salzberg. Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. *Nature biotechnology*, 33:290–295, March 2015. ISSN 1546-1696. doi: 10.1038/nbt.3122.

Christian Riml, Thomas Amort, Dietmar Rieder, Catherina Gasser, Alexandra Lusser, and Ronald Micura. Osmium-mediated transformation of 4-thiouridine to cytidine as key to study rna dynamics by sequencing. *Angewandte Chemie (International ed. in English)*, 56:13479–13483, October 2017. ISSN 1521-3773. doi: 10.1002/anie.201707465.

José A Robles, Sumaira E Qureshi, Stuart J Stephen, Susan R Wilson, Conrad J Burden, and Jennifer M Taylor. Efficient experimental design and analysis strategies for the detection of differential expression using rna-sequencing. *BMC genomics*, 13(1):484, 2012.

Johannes T Roehr, Christoph Dieterich, and Knut Reinert. Flexbar 3.0 - simd and multicore parallelization. *Bioinformatics (Oxford, England)*, 33:2941–2942, September 2017. ISSN 1367-4811. doi: 10.1093/bioinformatics/btx330.

Jeremy A Schofield, Erin E Duffy, Lea Kiefer, Meaghan C Sullivan, and Matthew D Simon. Timelapse-seq: adding a temporal dimension to rna sequencing through nucleoside recoding. *Nature methods*, 15:221–225, March 2018. ISSN 1548-7105. doi: 10.1038/nmeth.4582.

Björn Schwalb, Margaux Michel, Benedikt Zacher, Katja Frühauf, Carina Demel, Achim Tresch, Julien Gagneur, and Patrick Cramer. Tt-seq maps the human transient transcriptome. *Science*, 352(6290):1225–1228, 2016.

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*, 473:337–342, May 2011. ISSN 1476-4687. doi: 10.1038/nature10098.

Gregg L Semenza. Targeting hif-1 for cancer therapy. *Nature reviews. Cancer*, 3:721–732, October 2003. ISSN 1474-175X. doi: 10.1038/nrc1187.

Rolf Thermann and Matthias W Hentze. Drosophila mir2 induces pseudo-polysomes and inhibits translation initiation. *Nature*, 447(7146):875, 2007.

Alexey Uvarovskii and Christoph Dieterich. pulser: Versatile computational analysis of RNA turnover from metabolic labeling experiments. *Bioinformatics*, 33(20):3305–3307, 2017.

Adriaan Van den Bos. *Parameter estimation for scientists and engineers.* John Wiley & Sons, 2007.

Leonhard Wachutka and Julien Gagneur. Measures of rna metabolism rates: Toward a definition at the level of single bonds. *Transcription*, 8:75–80, March 2017. ISSN 2154-1272. doi: 10.1080/21541264.2016.1257972.

William R Wilson and Michael P Hay. Targeting hypoxia in cancer therapy. *Nature reviews. Cancer*, 11:393–410, June 2011. ISSN 1474-1768. doi: 10.1038/nrc3064.

Boback Ziaeian and Gregg C Fonarow. Epidemiology and aetiology of heart failure. *Nature reviews. Cardiology*, 13:368–378, June 2016. ISSN 1759-5010. doi: 10.1038/nrcardio.2016.25.

Extended Methods

Statistical model

We assume that the read counts follow the negative binomial distribution with the probability distribution function

$$P(X = x) = \frac{\Gamma(k + x)}{x!\Gamma(k)} \left(\frac{m}{m + k}\right)^{x} \left(\frac{m + k}{k}\right)^{-k},$$

where $m = m(\mu, d, \dots, t)$ is the mean read count expected from the kinetic model and depends on the time point t, expression level in a steady-state μ, degradation rate d, and sample normalization, see the next section for details. The negative binomial distribution imposes a relation between variance and mean via the overdispersion parameter k, $\mathrm{var}X = m(m + k)/k$. We assume the same k for all the genes in the data set, which is the simplest model, but more complicated models exist [Anders and Huber, 2012]. In this case, k is a shared parameter, and the model parameters from different genes must be fitted together in one procedure.

The logarithm of the likelihood function depends on the experimental points $X_1, \dots, X_n$ and the vector of all model parameters $\boldsymbol{\theta}$

$$\log \mathcal{L}(\boldsymbol{\theta}, X) = \sum_i \log P(\boldsymbol{\theta}, X_i). \tag{23}$$

The maximum likelihood estimator is then

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log \mathcal{L}(\boldsymbol{\theta}, X) \tag{24}$$

Kinetic model

The solution of the differential equation for the kinetics of synthesis and degradation of the RNA $\frac{dm}{dt} = s - dm$ is

$$m(t) = m_0 e^{-dt} + (1 - e^{-dt})s/d = \mu + (m_0 - \mu)e^{-dt}, \tag{25}$$

where m_0 is the initial amount of RNA, $\mu = s/d$ is the expression level is the steady state for synthesis rate s and degradation rate d.

For definiteness, consider a pulse experiment. The unlabeled fraction is being only degraded (synthesis rate $s = 0$), and the initial RNA level is μ

$$m_{\mathrm{U}} = \mu e^{-dt}. \tag{26}$$

The labeled fraction starts from zero and saturates to the steady state:

$$m_{\mathrm{L}} = \mu(1 - e^{-dt}). \tag{27}$$

In fact, the mean read count is only proportional to the amount of RNA in samples, so without spike-in fragments added to measure absolute concentration, we can estimate the amounts only up to unknown coefficient. For identifiability, we use the read counts in the total samples as a reference (accounting for difference in sequencing depth, as implemented in the `DESeq` package [Anders and Huber, 2012]).

If the ratio of fractions is preserved, as, for example, in the SLAMseq protocol, the counts in the labeled X_{L} and the unlabeled X_{U} fractions are scaled by the same sequencing depth correction x, $\mathbb{E}(X_{\mathrm{U}} + X_{\mathrm{L}}) = x\mu$ and

$$\mathbb{E}X_{\mathrm{U}} = x \cdot \mu e^{-dt} \tag{28}$$

$$\mathbb{E}X_{\mathrm{L}} = x \cdot \mu(1 - e^{-dt}) \tag{29}$$

Assuming that usually samples are sequenced to approximately the same depth, for theoretical derivations we use $x = 1$ for the SLAMseq experiment.

If the fractions were separated by a chemical procedure, the read counts ratio will not coincide with the ratio of labeled and unlabeled molecules. In the case of negligible cross-contamination,

$$\begin{aligned}\mathbb{E}X_{\mathrm{U}} &= x_1 \cdot \mu e^{-dt} \\ \mathbb{E}X_{\mathrm{L}} &= x_2 \cdot \mu(1-e^{-dt}).\end{aligned} \tag{30}$$

In this work, we do not consider cross-contamination, however it can play a significant role, especially for extreme rates (very fast or very slow) even if the overall contaminated material amount is low.

Indeed, if there is a cross-contamination level of γ, the mean read count can be modeled as

$$\mathbb{E}X_{\mathrm{U}} = x_1 \left((1-\gamma)\cdot \mu e^{-dt} + \gamma \cdot \mu \left(1-e^{-dt}\right)\right) \tag{31}$$

Under a wrong model, it leads to a biased degradation rate estimation:

$$\begin{aligned}\mathbb{E}X_{\mathrm{U}} &= x_1 \mu e^{-d^* t} = x_1 \mu \left(e^{-dt} + \gamma\left(1-2e^{-dt}\right)\right) \\ d^* &= -\frac{1}{t}\ln\left(e^{-dt} + \gamma\left(1-2e^{-dt}\right)\right)\end{aligned} \tag{32}$$

and for very slow genes, i.e. $dt \ll 1$, using the first term of the Taylor expansion

$$\frac{d^*}{d} \approx -\frac{1}{dt}\ln(1-\gamma) \gg 1. \tag{33}$$

In contrast, for very fast genes, such as $dt \gg 1$,

$$\frac{d^*}{d} \approx -\frac{\ln\gamma}{dt} \ll 1 \tag{34}$$

Hence, the rate estimations are biased to faster rates in the case of slow genes and toward faster values in the case of fast genes. The same result holds for the labeled fraction.

Fisher information matrix derivation

Let $m(\boldsymbol{\theta}, t)$ is the expected mean read number at time point t and $\boldsymbol{\theta} = (\mu, d)^T$ is a vector of model parameters.

The observed Fisher information matrix (FIM) is defined as

$$I_{\mathrm{ij}}(\boldsymbol{\theta}, X) = -\frac{\partial^2 \log \mathcal{L}(\boldsymbol{\theta}, X)}{\partial\theta_i \partial\theta_j} \tag{35}$$

In the optimal design of experiments, the expected FIM is used

$$\mathcal{I}_{\mathrm{ij}}(\boldsymbol{\theta}) = \mathbb{E}I_{\mathrm{ij}} = -\mathbb{E}\frac{\partial^2 \log \mathcal{L}(\boldsymbol{\theta}, X)}{\partial\theta_i \partial\theta_j}, \tag{36}$$

since we may not have any measurements and are interested in the performance of a given design in average. We refer everywhere only to the expected FIM and name it just FIM, omitting the "expected" term.

We will use here the fact, that the variance of the score function

$$\operatorname{var}_{\boldsymbol{\theta}} S(\boldsymbol{\theta}) = \mathcal{I}(\boldsymbol{\theta}), \tag{37}$$

where

$$S(\boldsymbol{\theta}) = \begin{pmatrix} \partial \log \mathcal{L}/\partial\theta_1 \\ \vdots \\ \partial \log \mathcal{L}/\partial\theta_p \end{pmatrix}.$$

$$S_i(\boldsymbol{\theta}) = X\frac{k}{m(m+k)}\frac{\partial m}{\partial\theta_i} + f(m,k),$$

where $f(m,k)$ is a term not depending on X. Using the fact that var$X = m(m+k)/k$,

$$\mathcal{I}(d) = \mathrm{var}\, S_d = \frac{k}{m(m+k)} \frac{\partial m}{\partial \theta_i} \frac{\partial m}{\partial \theta_j} \tag{38}$$

In the $k \to \infty$ case, when no overdispersion is assumed and the model follows the Poisson distribution,

$$\mathcal{I}_{\mathrm{Pois}}(d) = \mathrm{var}\, S_d = \frac{1}{m} \frac{\partial m}{\partial \theta_i} \frac{\partial m}{\partial \theta_j}$$

By plugging the relations for the RNA amount $m = \mu e^{-dt}$ (the unlabeled fraction) and $m = \mu(1 - e^{-dt})$ (the labeled fraction) in Equation 38,

$$(\mathcal{I}_{\mathrm{L}})_{\mathrm{dd}} = \frac{t^2 e^{-2dt} \mu}{1 - e^{-dt}} \frac{1}{1 + \frac{\mu}{k}(1 - e^{-dt})} \tag{39}$$

$$(\mathcal{I}_{\mathrm{U}})_{\mathrm{dd}} = t^2 e^{-dt} \mu \frac{1}{1 + \frac{\mu}{k} e^{-dt}} \tag{40}$$

For the SLAMseq case,

$$\mathcal{I}_{\mathrm{slam}} = \mathcal{I}_{\mathrm{L}} + \mathcal{I}_{\mathrm{U}} = \begin{pmatrix} \frac{k\mu t^2 e^{-dt}}{\mu e^{-dt} + k} + \frac{k\mu t^2 e^{-2dt}}{\left(\mu\left(1 - e^{-dt}\right) + k\right)\left(1 - e^{-dt}\right)} & \frac{kte^{-dt}}{\mu\left(1 - e^{-dt}\right) + k} - \frac{kte^{-2dt}e^{dt}}{\mu e^{-dt} + k} \\ \frac{kte^{-dt}}{\mu\left(1 - e^{-dt}\right) + k} - \frac{kte^{-2dt}e^{dt}}{\mu e^{-dt} + k} & \frac{ke^{-dt}}{\mu\left(\mu e^{-dt} + k\right)} + \frac{k\left(1 - e^{-dt}\right)}{\mu\left(\mu\left(1 - e^{-dt}\right) + k\right)} \end{pmatrix}$$

The inverse matrix in the general case is

$$\mathcal{I}_{\mathrm{slam}}^{-1} = \begin{pmatrix} \frac{e^{dt} - 1}{\mu t^2} + \frac{2(1 - e^{-dt})^2}{kt^2} & \frac{e^{-2dt}\left(2\mu - 3\mu e^{dt} + \mu e^{2dt}\right)}{kt} \\ \frac{e^{-2dt}\left(2\mu - 3\mu e^{dt} + \mu e^{2dt}\right)}{kt} & \mu + \frac{\mu^2}{k}\left(e^{-2dt} + \left(1 - e^{-dt}\right)^2\right) \end{pmatrix}$$

In the case of low overdispersion, $k \gg 1$, it is simplified to

$$\mathcal{I}_{\mathrm{slam}}^{-1} = \begin{pmatrix} \frac{e^{dt} - 1}{\mu t^2} & 0 \\ 0 & \mu \end{pmatrix}$$

A model with overdispersion imposes a limit on the asymptotic lower bound of the MLE for d, which cannot be improved by increase of sequencing depth (i.e. μ). For a fixed t,

$$\lim_{\mu \to \infty} (\mathcal{I}_{\mathrm{slam}}^{-1})_{\mathrm{dd}} = \frac{2(1 - e^{-dt})^2}{kt^2}$$

Although this limit value decreases with $t \to \infty$, the depth μ must increase exponentially $e^{dt}/\mu(t) \to 0$ in order to satisfy

$$\lim_{t \to \infty} \frac{e^{dt} - 1}{\mu(t) t^2} = 0.$$

Normalization in conventional experiments

The situation is different for the case of biochemical purification, when the labeled, unlabeled and total fractions are sequenced to the approximately same depth. For the next derivation, we assume that samples were normalized externally, e.g. by synthetic spike-ins or exogenous RNA. In addition, we do not consider uncertainty coming from fraction normalization.

The sequencing depth for the total sample is $\sum_i \mu_i$. After labeling for t hours, the concentrations of labeled and unlabeled molecules changes according to Equation 30, and in the case of same depth, the normalization coefficients are

$$\begin{aligned} x_{\mathrm{L}} &\approx \frac{\sum_i \mu_i}{\sum_i \mu\left(1-e^{-d_i t}\right)} \\ x_{\mathrm{U}} &\approx \frac{\sum_i \mu_i}{\sum_i \mu_i e^{-d_i t}}, \end{aligned}$$

where we use the total sample as a reference, i.e. with the normalization coeffiecient 1. At long times, when majority of genes in the labeled fraction achieve saturation, $x_{\mathrm{L}} \approx 1$. Similarly, at short times, $x_{\mathrm{U}} \approx 1$, degradation has a minor effect on the unlabeled fraction.

In contrast, at short times,

$$x_{\mathrm{L}} \approx \frac{\sum_i \mu_i}{\sum_i \mu d_i t} = \frac{1}{\langle d \rangle t}, \tag{41}$$

where $\langle d \rangle = \sum_i \mu_i d_i / \sum_i \mu_i$ is average degradation rate weighted by the steady-state expression level. If there is a small cluster of fast genes $i \in \mathcal{F}$, which dominate the pool of labeled molecules at such times, that $(1 - e^{-d_i t}) \ll 1$ for other genes ($i \notin \mathcal{F}$), and $(1 - e^{-d_i t}) \approx 1$ for $i \in \mathcal{F}$, then

$$x_{\mathrm{L}} \approx \frac{\sum_i \mu_i}{\sum_{i \in \mathcal{F}} \mu_i}. \tag{42}$$

Similar result holds true for the cluster of slow genes $\mathcal{S}$, such that $e^{-d_i t} \approx 1$ for $i \in \mathcal{S}$ and $e^{-d_i t} \ll 1$ for $i \notin \mathcal{S}$, and

$$x_{\mathrm{U}} \approx \frac{\sum_i \mu_i}{\sum_{i \in \mathcal{S}} \mu_i}. \tag{43}$$

Such "zooming" effect of normalization can drastically improve inference about fast genes in comparison to the SLAMseq design, since x_{L} and x_{U} can be interpreted as a corresponding increase in depth in SLAMseq experiment.

Modifying the mean read count in Equations 39 and 40 by the factors x_{L} and x_{U} results in

$$(\mathcal{I}_{\mathrm{L}})_{\mathrm{dd}} = \frac{t^2 e^{-2dt}}{(1 - e^{-dt})} \frac{1}{\frac{1}{x_{\mathrm{L}} \mu} + \frac{1}{k}(1 - e^{-dt})} \tag{44}$$

$$(\mathcal{I}_{\mathrm{U}})_{\mathrm{dd}} = t^2 e^{-dt} \frac{1}{\frac{1}{x_{\mathrm{U}} \mu} + \frac{1}{k} e^{-dt}} \tag{45}$$

As in the SLAMseq case, the overdispersion imposes limits on the terms of the FIM:

$$\lim_{\mu \to \infty} (\mathcal{I}_{\mathrm{L}})_{\mathrm{dd}} = \frac{t^2 e^{-2dt} k}{(1 - e^{-dt})^2} \leqslant \frac{k}{d^2} \tag{46}$$

$$\lim_{\mu \to \infty} (\mathcal{I}_{\mathrm{U}})_{\mathrm{dd}} = t^2 k. \tag{47}$$

It is interesting, that in the case of the unlabeled fraction, this upper bound can be improved by using longer labeling times, which is not true for the labeled one.

Experimental Model and Methods

Tissue Culture Cell Line

MCF-7 cells (ACC-115) were obtained from the Leibniz Institute DSMZ German Collection of Microorganisms and Cell Cultures. Cells were routinely tested for mycoplasma contamination with Venor GeM Classic (Minerva Biolabs). MCF-7 cells were cultured at 37°C and 5% CO2 and maintained in DMEM (Thermo Fisher Scientific) supplemented with 10% fetal calf serum (Merck), 1xMEM non-essential amino acids (Thermo Fisher Scientific) and 1xPenicillin/Streptomycin (Thermo Fisher Scientific).

Tissue Culture

MCF-7 cells were seeded 48 hrs prior to the experiment at a cell density of 0.3×10^5cells/cm^2. Cells were labeled with 4-thiouridine (4sU) (Sigma-Aldrich) at a final concentration of 200 μM for 2, 4 or 8 hrs. Cells were scraped in DPBS and the pellet resuspended in Trizol (Thermo Fisher Scientific).

Isolation of total RNA

Total RNA was isolated using the Trizol method. Briefly, the cell pellet was resuspended in 750 μl Trizol, and incubated 5 min at room temperature before addition of 200 μl chloroform. Samples were centrifuged (20 min, 10.000g, room temperature) and the aqueous phase re-extracted with one volume chloroform: isoamylalkohol (24:1) (5 min, 10.000g, room temperature). The RNA in the aqueous phase was precipitated with one volume isopropanol (30 min, 20.8000g, 4°C), washed twice with 1 ml 80% ethanol in DEPC-H_2O and dissolved in 25 μl DEPC-H_2O (10 min, 55°C, shaking).

In vitro transcription of spike ins

For *in vitro* transcription of linearized plasmids (pBSIIKS-Luc-pA-NB [Liepelt et al., 2014] and pBSIIKS-Renilla-pA [Thermann and Hentze, 2007]), the MEGAscript T7 Transcription Kit (Thermo Fisher Scientific) was used according to the manufacturers instructions. Briefly, the reaction was set up in a total volume of 20 μl containing 1 μg linearized plasmid and 2 μl 10x reaction buffer, 3 μl 40 mM m^{7}GppG-cap analogon (KEDAR), 2 μl 15 mM GTP, 2 μl 75 mM CTP, 2 μl 75 mM ATP, 2 μl enzyme mix and 2 μl 75 mM UTP (for RLuc) or 2 μl 75 mM 4-S-UTP:UTP in a 1:10 ratio (for FLuc). Reactions were incubated 3 hrs at 37°C. Plasmid-DNA was removed by addition of 1 μl Turbo-DNase (15 min, 37°C). *In vitro* transcribed RNA was purified by phenol extraction and Chromaspin-100 (Clontech) purification. RNA was precipitated over night after addition of sodium acetate to a final concentration of 0.3 M and 2.5 volumes 100% ethanol. After centrifugation (30 min, 20.800g, 4°C) the pellet was washed with 1 ml 80% ethanol and dissolved in 40 μl DEPC-H_2O. Concentration was determined by Nanodrop (Thermo Fisher Scientific) measurement and integrity checked by agarose gel electrophoresis.

Biotinylation of RNA

Total RNA was spiked with *in vitro* transcribed 4sU-labeled FLuc and non-labeled RLuc RNAs and biotinylated using MTSEA biotin-XX (Biotium) as described by [Duffy et al., 2015]. Briefly 80 μg total RNA was incubated with 8 ng FLuc and 4.8 ng RLuc (equimolar amounts, 130 amol), 10 mM HEPES pH 7.5, 1 mM EDTA and 5 μg MTSEA biotin-XX (freshly dissolved in DMF) in a total volume of 250 μl. Reactions were incubated 30 min in the dark at room temperature. Biotinylated RNA was recovered by extraction with one volume phenol: chloroform: isoamylalkohol (24:24:1) and separated using Phase-Lock-tubes (5Prime) by centrifugation (5 min, 20.800g, room temperature). RNA was precipitated by addition of 350 μl isopropanol, 25 μl 5 M sodium chloride and 1 μl glycogen (Roche Diagnostics, 20 μg/μl) to assist precipitation (30 min, 20.800g, 4°C). RNA was washed twice with 500 μl 80% ethanol in DEPC-H_2O and dissolved in 25 μl DEPC-H_2O (10 min, 55°C, shaking).

Streptavidin purification

For purification of biotinylated RNAs the method described by [Schwanhäusser et al., 2011] was adapted. 25 μg biotinylated total RNA was adjusted to 100 μl with DEPC-H_2O and filled up with Streptavidin binding buffer (Strep-BB) (20 mM Tris, pH 7.4, 0.5 M sodium chloride, 1 mM EDTA) to 200 μl. RNA was denatured 10 min at 65°C and subsequently placed on ice. 100 μl magnetic streptavidin beads (New England Biolabs) were washed once with 200 μl Strep-BB and resuspended in 100 μl Strep-BB. RNA and beads were incubated 15 min at room temperature on a rotating wheel. Beads were washed three times with 500 μl Strep washing buffer (100 mM Tris pH 7.4, 1 M sodium chloride, 10 mM EDTA, 0.1% Tween 20) prewarmed to 55°C. RNA was eluted three times with 100 μl freshly prepared 100 mM DTT and elution fractions pooled for further analysis. RNA was recovered from total RNA, flow through and eluate by phenol: chloroform: isoamylalkohol (24:24:1) extraction using Phase-Lock-tubes and isopropanol precipitation as described above. The amount of recovered RNA was determined by Nanodrop measurement.

Dot blot-based detection of biotinylation

1 μg biotinylated RNA was applied to nylon membrane (Hybond-N, GE Healthcare) using a dot blot device (Carl Roth). RNA was crosslinked twice at 254 nm using the "Optimal Crosslink" mode of the Spectroline Select XLE-1000 crosslinker. The membrane was blocked 20 min with PBS + 10% SDS and incubated 2 hrs with Streptavidin-HRP (Thermo Fisher Scientific, 1:5000 in PBS + 10% SDS). Prior to detection with SuperSignal West Pico (Thermo Fisher Scientific) the membrane was washed each three times 10 min with PBS + 10% SDS, PBS + 1% SDS and PBS + 0.1% SDS. Images were acquired with the LAS4000 system (GE Healthcare).

Reverse Transcription

1 μl RNA from streptavidin purification was reverse transcribed using the Maxima H Minus First Strand cDNA Synthesis Kit (Thermo Fisher Scientific) with Random Primers according to the manufacturers protocol. For absolute quantification reverse transcription reactions were set up with different amounts of spike in RNAs, ranging from 1600% to 1.56% for FLuc and 400 to 3.12% for RLuc in 1:2 dilutions. Briefly, RNA was mixed in a total volume of 15 μl with 1 μl Random Primer and 1 μl dNTP solution and denatured (5 min, 65°C). Reaction was completed by addition of 4 μl 5xRT buffer and 1 μl Maxima enzyme and incubated 10 min at room temperature followed by 30 min, 50°C and denaturation (5 min, 85°C).

qPCR Analysis

Reverse transcription reactions were diluted 1:10 and used for qPCR analysis on a StepOnePlus instrument (ThermoFisherScientific) with Power SYBR Green PCR Master Mix (Thermo Fisher Scientific) and primers directed against FLuc (forward: CCTTCCGCATAGAACTGCCT, reverse: GGTTGGTACTAGCAACGCAC [de Vries et al., 2013]) and RLuc (forward: GTTGTGCCACATATTGAGCC, reverse: CCAAACAAGCACCCCAATCATG [Naarmann-de Vries et al., 2016]).

Sequencing

Libraries of 2 biological replicate 4sU pulse experiment were sequenced 1x 50bp on an Illumina HiSeq4000. All relevant details on sequencing depth and mapping rates are listed in Supplementary Table 1.

Read processing and counting

Sequencing adapters and low-quality reads were removed from the raw sequencing data with flexbar v3.0.3 [Roehr et al., 2017] using standard filtering parameters. We excluded all reads with more than 1 uncalled base from the output. All remaining reads (>18bp) were then aligned to a custom sequence index including rRNA, tRNA and snoRNA gene loci using bowtie2 with the –very-fast option [Langmead and Salzberg, 2012]. Only reads that did not align to any of the contaminant sequences were considered for further analysis.

Reads were then aligned to the human genome (EnsEMBL 85) and splice sites from the reference annotation with a splice-aware aligner (STAR, v2.5.3a; Dobin et al. [2013]). The BAM files were analyzed with StringTie 1.3.3b [Pertea et al., 2015] and the final read count matrix was prepared with the supplemented python script prepDE.py.

Supplementary Information

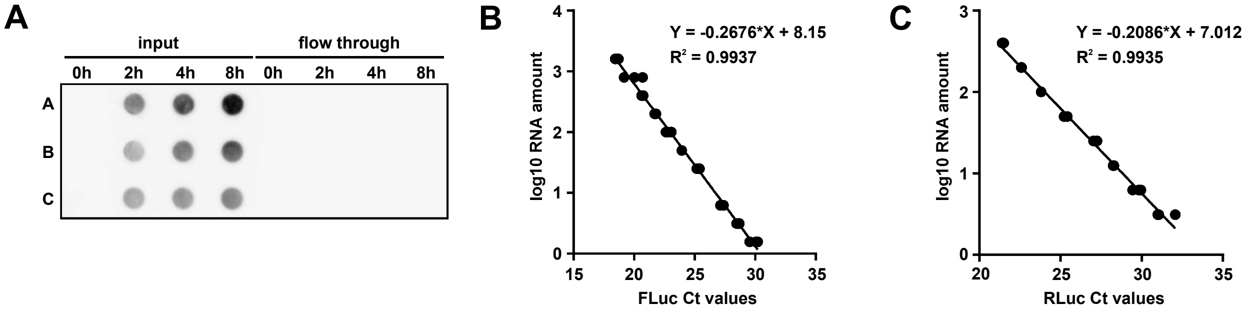

Figure S1: **A:** Dot blot-based detection of biotinylation with streptavidin-HRP in input and flow through of streptavidin purification from three replicate experiments A-C. The quantification of the captured image is shown in Figure 3A. **B:** Standard curve for the absolute quantification of 4sU-labeled FLuc RNA. 1600 to 1.56% of the input used for streptavidin purification was measured by RT-qPCR analysis in 1:2 dilutions. The log10 amount of RNA was plotted against the obtained Ct value and used for linear regression. **C:** Standard curve for the absolute quantification of unlabeled RLuc RNA. 400 to 3.13% of the input used for streptavidin purification was measured by RT-qPCR analysis in 1:2 dilutions. The log10 amount of RNA was plotted against the obtained Ct value and used for linear regression.

Table S1: **Summary of RNA-seq read mapping statistics**

Condition	Sample ID	Uniquely mapped reads %	Uniquely mapped reads number
Rep2 8h pull-down	K002000165_76894	90.4%	31.4
Rep2 0h Total RNA	K002000165_76874	85.1%	29.5
Rep1 8h pull-down	K002000165_76872	89.1%	28.2
Rep2 2h pull-down	K002000165_76882	92.5%	27.3
Rep1 4h pull-down	K002000165_76866	89.3%	26.8
Rep2 4h pull-down	K002000165_76888	91.2%	24.2
Rep1 0h Total RNA	K002000165_76852	79.9%	23.2
Rep1 2h pull-down	K002000165_76860	91.3%	23.2
Rep1 2h flow-through	K002000165_76856	73.7%	21.8
Rep2 2h flow-through	K002000165_76878	77.1%	21.7
Rep2 4h flow-through	K002000165_76884	71.5%	21.0
Rep2 8h flow-through	K002000165_76890	66.0%	18.0
Rep1 8h flow-through	K002000165_76868	63.4%	16.6
Rep1 4h flow-through	K002000165_76862	66.2%	11.4

Table S2: **Read counts for all samples**
Weblink