pulseR: Versatile computational analysis of RNA turnover from metabolic labeling experiments

Alexey Uvarovskii, Christoph Dieterich

Section of Bioinformatics and Systems Cardiology, Klaus Tschira Institute for Integrative Computational Cardiology,Department of Internal Medicine III,
University Hospital Heidelberg, Im Neuenheimer Feld 669, 69120 Heidelberg
German Center for Cardiovascular Research (DZHK), Im Neuenheimer Feld 669, 69120 Heidelberg

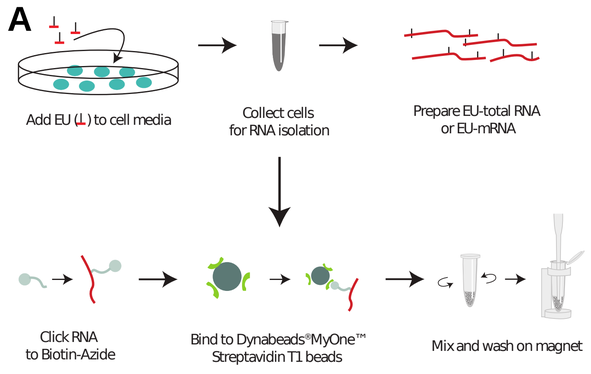
Summary

Gene expression level is defined by rates of RNA synthesis and degradation. Understanding how certain gene levels are regulated by these processes across different experimental conditions helps to gain deeper insights into RNA control mechanisms. Metabolic RNA labelling experiments facilitate to separate newly synthesized (i.e. labelled) from pre-existing RNA. Using this approach, pulse-chase time course experiments can be set up to estimate kinetic rates of RNA turnover from next-generation sequencing data. However, there is no publicly available software for estimating kinetic parameters to date, which is specifically designed to handle fragment count data from RNA sequencing experiments.

We meet this demand by presenting an R package pulseR, which handles different experimental designs and readily incorporates information from exogenous spike ins. Additionally, It is possible to adjust for potential gene-specific and global biases, which may arise during RNA purification, due to different uridine content or due to cross-contamination.

The pulseR package is freely available at https://github.com/dieterich-lab/pulseR under the GPLv3.0 licence. A tutorial can be found under https://goo.gl/xmonsg.

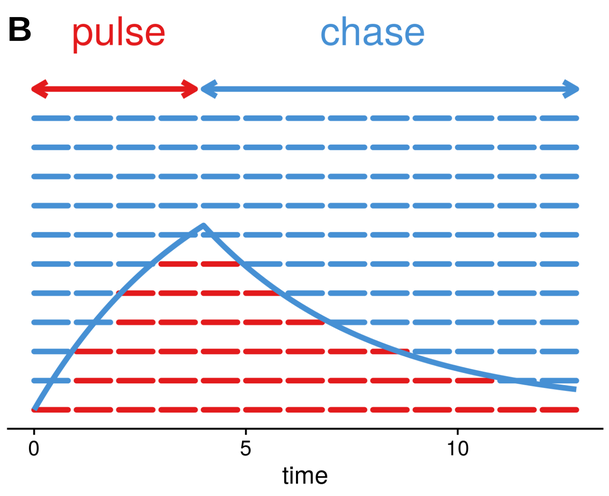
Measurement of RNA turnover



There are two processes involved in RNA metabolism: RNA synthesis and degradation. In order to be able to measure the reaction rates, one needs to distinguish the newly synthesized RNA from the older one. Different approaches were developed, which are based on addition of modified uridine molecules, namely, 4-thiouridine (4sU), 5-etyniluridine (EU) and 5′-bromo-uridine (BrU).

The figure is adapted from Click-iT® Kit Thermofisher manual.

References:
Tani H, Akimitsu N. Genome-wide technology for determining RNA stability in mammalian cells: Historical perspective and recent advantages based on modified nucleotide labeling. RNA Biology. 2012;9 (10):1233-1238. doi:10.4161/rna.22036.



Pulse-chase experiments allow to measure kinetic rates for RNA synthesis and degradation. After the pulse period, when cells are cultured in the presence of the label, the chase period follows with no labelling. Hence, the labelled RNA is being only degraded during the chase time.

The set-up can be varied in terms of time ranges and types of collected fractions (e.g. all 3 fractions or only total and labelled ones)

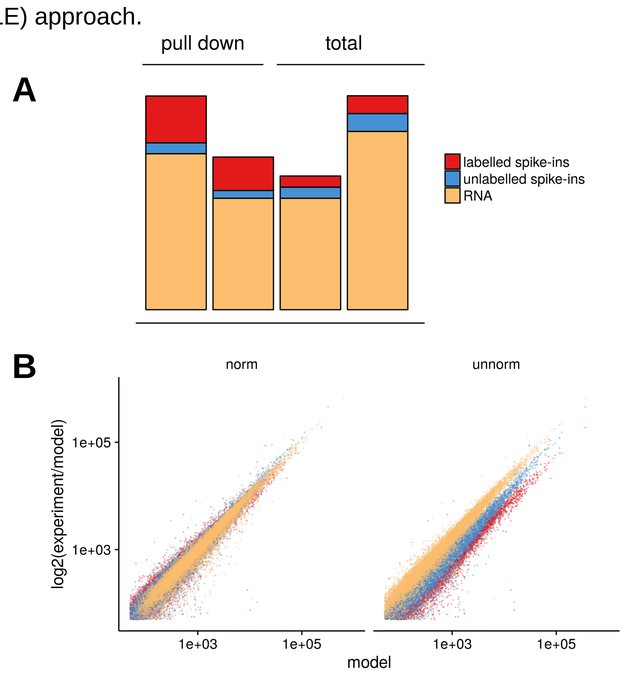
Methods

Kinetic model: RNA dynamics can be described by ordinary differential equations, which have a simple analytic solution if the degradation and synthesis rates are assumed to be constant.

$$\frac{d[RNA]}{dt} = +[synthesis] - [degradation] \cdot [RNA]$$

Stat model: We specifically consider overdispersion in RNA fragment counts by using the negative binomial distribution to model observations from RNA-seq data.

The model is fitted using maximum likelihood estimation (MLE) approach.

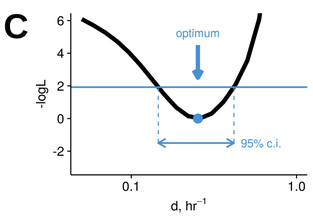


Confidence intervals (CI): The profile likelihood is used to estimate the CI for the model parameters.

In this approach, the parameter p under investigation is fixed to a certain value, while other parameters are left free. The likelihood function L(p) is then a function of one argument and the 95% CI is a set of p such that

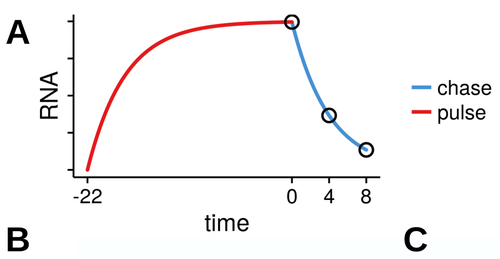
$$2\frac{\log L(p)}{\log L(p^*)} < \chi^2_{1,0.95}$$

where p* is the optimal parameter value and the r.h.s is the 95% quantile for the chi-squared distribution with 1 degree of freedom (approx. 3.84), fig. C.

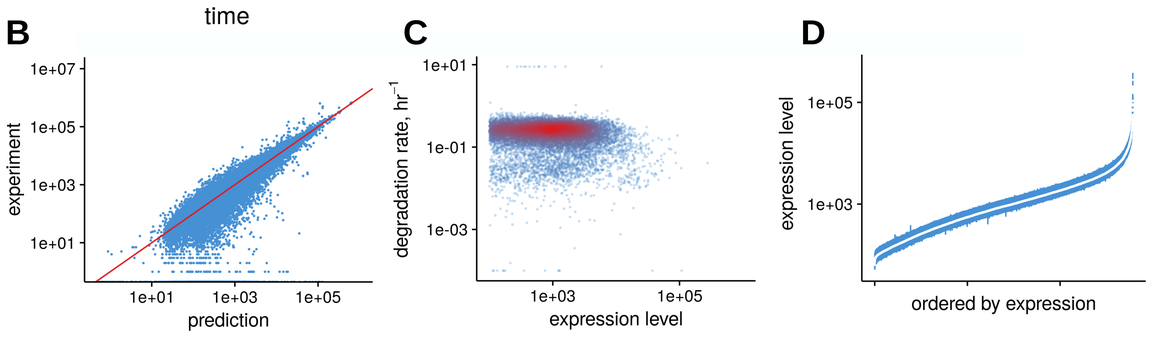


Normalisation: due to difference in sequencing depth and possible cross-contamination (fig. A), it is important to make adjustments in estimations. We use DESeq normalisation if spike ins are provided. Alternatively, we estimate sample relations by MLE. This helps to eliminate sample difference coming from the preparation (fig. B).

Application to H9 cells



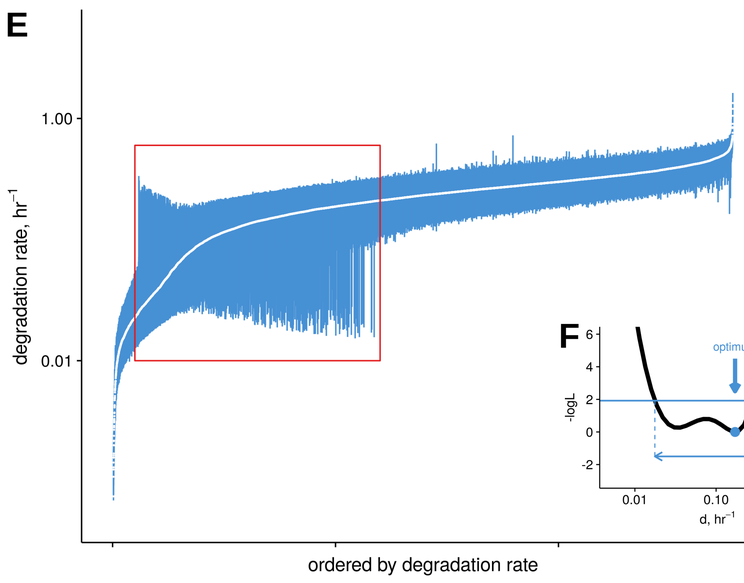
A: We applied the pulseR package to the data from pulse-chase experiment on H9 human embryonic cells. After 22 hours of pulse with EU, the chase setup was sampled at 0, 4 or 8 hr. The data include two fractions, the total and the pull-down RNA.



B: We fitted the model without spike ins, i.e. the relation between samples is derived from the fitting itself. The model predictions align well with the read counts in the samples.

C: Expression level does not show any dependency on the degradation rate. However, there is a subset of slow degradation rate fitted (see below for discussion).

D: The 95% confidence intervals (CI) for the expression level estimated from the profile likelihood do not show any systematic problems on all value ranges.



E: 95% CI for the degradation rate. The genes are ordered by the fitted value.

Large CI are observed for a gene subset with the fitted value in a certain range (red rectangle).

F, G: The diagnostic plots of the profile likelihood show uncertainty in identification of the degradation rate for some genes (fig. C) using data from this experiment.

Workflow

```
library(pulseR)
# put math here
formulas <- MeanFormulas(
  total      = mu,
  labelled   = mu * (1 - exp(-d * 22)) * exp(-d * time),
  unlabelled = mu * (1 - exp(-d * time) * (1 - exp(-d * 22)))
)
# define the fractions
formulaIndexes <- list(
  total_fraction = 'total',
  pull_down      = c('labelled', 'unlabelled'))
lbNormFactors <- list(
  total_fraction = .1,
  pull_down      = c(.1,.00001))
ubNormFactors <- list(
  total_fraction = 10,
  pull_down      = c(10, .3))
opts <- setBoundaries(list(
  mu    = c(.1, 1e8),
  d     = c(1e-4, 9),
  size = c(1,1e6)),
  normFactors = list(lbNormFactors, ubNormFactors)
)
# let conditions be in a data.frame (sample, fraction, time)
pd <- PulseData(counts, conditions, formulas, formulaIndexes,
  groups = ~ fraction + time)
result <- fitModel(pd, initParValues, opts)
```

The typical analysis workflow includes:

- definition of the kinetic equations
- definition of the fraction content, e.g. the pull-down can be contaminated with the unlabelled fraction
- setting lower and upper boundaries for the parameters, as well as fitting options
- creation of the PulseData object and, finally, fitting the model

	pulseR	DRiLL	INSPEcT	DTA	HALO
statistical model	NB	N, BIN	N	–	–
spike−ins	+	–	–	–	–
several time points	+	+	+	–	–
variable design [1]	+	–	–	–	–
non−constant rates	–	+	+	–	–
uridine bias	*	–	–	+	+
RNA processing	*	+	+	–	–
confidence intervals	+	–	–	–	–
language	R	MATLAB	R	R	Java

Table 1. Feature comparison of software packages for parameter estimation in pulse-chase experiments.
+: available, -: not implemented, N: normal, NB: negative binomial, BIN: binomial.
*: must be defined by user. [1]: pulse, chase or combination there of experiments.

For references and further information see:
1. Alexey Uvarovskii, Christoph Dieterich; pulseR: Versatile computational analysis of RNA turnover from metabolic labeling experiments. Bioinformatics 2017 btx368. doi: 10.1093/bioinformatics/btx368
2. Wachutka, Leonhard, and Julien Gagneur. "Measures of RNA metabolism rates: Toward a definition at the level of single bonds." Transcription 8.2 (2017): 75-80.

Acknowledgement

We would like to thank Tobias Jakobi (Dieterich Lab) for the computing support, David Vilchez and Seda Koyuncu (CECAD Cologne), Janine Altmüller and Marek Franitza (CCG Cologne) for providing the experimental data.

a.uvarovskii@uni-heidelberg.de
christoph.dieterich@uni-heidelberg.de

https://dieterichlab.org