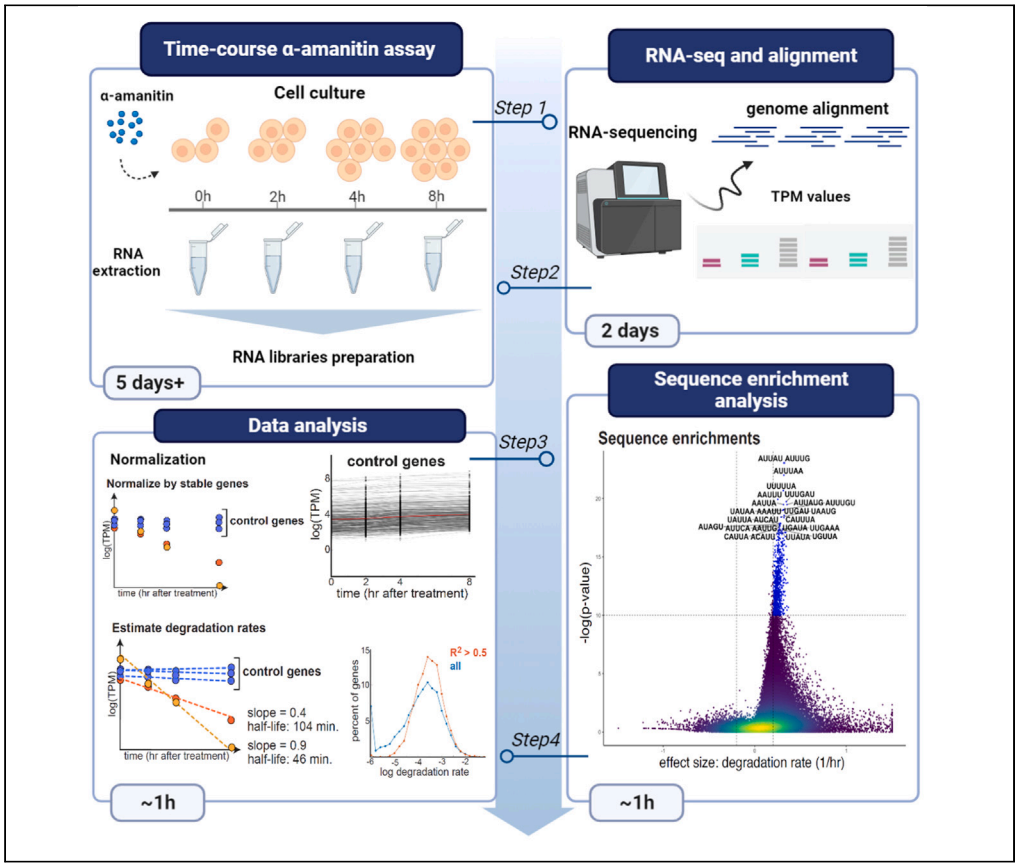
STAR Protocols

CellPress
OPEN ACCESS

Protocol

Calculating RNA degradation rates using large-scale normalization in mouse embryonic stem cells



Data normalization is critical to the process of estimating RNA degradation by analyzing RNA levels when transcription is blocked. Here, we present a protocol for measuring mRNA degradation rates, optimized for mouse embryonic stem cells, using α-amanitin inhibitor. We describe steps for a time course α-amanitin treatment, RNA-seq, and alignment; we then detail procedures for analyzing data and sequence enrichment. Our method relies on large-scale normalization of stable transcripts in genomic RNA-seq measurements, providing reliable readouts.

Publisher's note: Undertaking any experimental protocol requires adherence to local institutional guidelines for laboratory safety and ethics.

Juliane Oliveira Viegas, Lior Fishman, Eran Meshorer, Michal Rabani

eran.meshorer@mail.huji.ac.il (E.M.)
michal.rabani@mail.huji.ac.il (M.R.)

Highlights

Time course α-amanitin treatment blocks transcription to infer RNA decay rates

Large-scale data normalization by stable transcripts provides reliable readouts

Optimized for α-amanitin treatment in mouse embryonic stem cells

Downstream analysis of decay rates predicts associated short sequence elements

Viegas et al., STAR Protocols 4, 102534
September 15, 2023
https://doi.org/10.1016/j.xpro.2023.102534

STAR Protocols

CellPress
OPEN ACCESS

Protocol

Calculating RNA degradation rates using large-scale normalization in mouse embryonic stem cells

Juliane Oliveira Viegas,[1,3] Lior Fishman,[1,3] Eran Meshorer,[1,2,*] and Michal Rabani[1,4,*]

[1]Department of Genetics, The Institute of Life Sciences, The Hebrew University of Jerusalem, Edmond J. Safra Campus, Jerusalem 9190401, Israel

[2]The Edmond and Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem, Edmond J. Safra Campus, Jerusalem 9190401, Israel

[3]Technical contact: juliane.viegas@mail.huji.ac.il; lior.fishman@mail.huji.ac.il

[4]Lead contact

*Correspondence: eran.meshorer@mail.huji.ac.il (E.M.), michal.rabani@mail.huji.ac.il (M.R.)
https://doi.org/10.1016/j.xpro.2023.102534

SUMMARY

Data normalization is critical to the process of estimating RNA degradation by analyzing RNA levels when transcription is blocked. Here, we present a protocol for measuring mRNA degradation rates, optimized for mouse embryonic stem cells, using α-amanitin inhibitor. We describe steps for a time course α-amanitin treatment, RNA-seq, and alignment; we then detail procedures for analyzing data and sequence enrichment. Our method relies on large-scale normalization of stable transcripts in genomic RNA-seq measurements, providing reliable read-outs.

For complete details on the use and execution of this protocol, please refer to Viegas et al.[1]

BEFORE YOU BEGIN

The protocol below describes the steps to grow mouse embryonic stem cells (ESCs), inhibit RNA polymerase using α-amanitin treatment in a time-course manner, and calculate RNA degradation rate using large-scale normalizations. The conditions are optimized for undifferentiated mouse ESCs but can be used for differentiated cells as well as any other cell type.

The method includes steps:

1. Growing mouse ESC;
2. α-amanitin treatment and harvesting;
3. RNA extraction, library preparation and sequencing;
4. Large-scale normalization of a-amanitin treatment time course;
5. RNA decay analysis;
6. Associating short 3′ UTR sequences with mRNA decay rates.

Preparation of buffers and culture media

Culture media and buffers are prepared following the recipes described in the materials and equipment section. A complete list of materials used in this study is shown in the key resources table.

Computational requirements

Software: Linux (kernel version 5.4 or 5.10: e.g., in Ubuntu 20.04 or above); Windows 10 or MacOS (High Sierra or above), including installation of MATLAB (MathWorks, v. R218b) and R (R Core Team, v. 4.2.2).

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
Dulbecco's modified Eagle's medium, high glucose (DMEM)	Sigma-Aldrich	Cat#D5671
Fetal bovine serum (FBS)	Biological Industries	Cat#04-007-1A
Penicillin-Streptomycin solution (10 mg/mL)	Biological Industries	Cat#03-031-1B
L-Glutamine solution (200 mM)	Biological Industries	Cat#03-020-1B
Sodium pyruvate solution (100 mM)	Biological Industries	Cat#03-042-1B
2-Mercaptoethanol (14.3 M)	Sigma-Aldrich	Cat#M3148
Gelatin solution, 2% in H_2O	Sigma-Aldrich	Cat#G1393
PBS X10	Biological Industries	Cat#02-023-5A
Non-essential amino acids (NEAA)	Biological Industries	Cat#01-340-1B
Critical commercial assays		
RNeasy Mini Kit	QIAGEN	Cat#74106
KAPA Stranded mRNA-Seq Kit	Kapa Biosystems, Inc.	Cat#07962142001
RNAse-Free DNase set	QIAGEN	Cat#79254
Deposited data		
Raw and analyzed data	Viegas et al.[1]	GSE179002
Experimental models: Cell lines		
Mouse R1-ESC	Prof Eran Meshorer Lab, Hebrew University of Jerusalem	ATCC Cat#SCRC-1011 https://www.atcc.org/products/scrc-1011
Software and algorithms		
FastQC (version 0.11.9)	Andrew[2]	https://www.bioinformatics.babraham.ac.uk/projects/download.html
STAR (version 2.4.0.1)	Dobin et al.[3]	https://github.com/alexdobin/STAR
R Studio (Core Team, v.4.2.2)	R	https://www.r-project.org
MATLAB (v. R218b)	MATLAB	https://www.mathworks.com
Other		
Agilent TapeStation	Agilent Technologies	https://www.agilent.com/en/product/automated-electrophoresis/tapestation-systems
Illumina NextSeq500 sequencer	Illumina	

MATERIALS AND EQUIPMENT

2i medium		
Reagent	Final concentration	Amount
DMEM medium	N/A	87.74 mL
Fetal Bovine serum (FBS)	10% (v/v)	10 mL
Sodium Pyruvate (100 mM)	1 mM	1 mL
Non-essential amino acids (NEAA)	1/100 of the stock	1 mL
Penicillin/Streptomycin stock	20 U/mL/20 μg/mL	50 μL
LIF	N/A	100 μL
2-mercaptoethanol (2-ME)	N/A	70 μL
CHIR99021 10 mM	3 μM	30 μL
PD0325901 10 mM	1 μM	10 μL
Total	**N/A**	**100 mL**

Note: Store at 4°C for up to 1 month.

Gelatin 0.2%		
Reagent	Final concentration	Amount
Gelatin solution, 2% in H_2O	0.2%	10 mL
Nuclease-free water	N/A	90 mL
Total	**N/A**	**100 mL**
Note: Store at RT for up to one month.		

PBS 1×		
Reagent	Final concentration	Amount
PBS ×10	1×	10 mL
Nuclease-free water	N/A	90 mL
Total	**N/A**	**100 mL**
Note: Store at RT for up to one month.		

α-Amanitin preparation

α-Amanitin comes as powder. In a chemical hood with proper handling materials, dissolve all powder with sterile water to the final concentration of 1 mg/mL, make aliquots of 10 μL and store in −20°C. α-Amanitin powder is stable for over a year, but it is recommended to use promptly after dissolving in water.

⚠ CRITICAL: α-amanitin is considered a hazardous substance very toxic by inhalation, in contact with skin and if swallowed. Always wear proper protective equipment and use it only in a fume hood with filter tips.

STEP-BY-STEP METHOD DETAILS

Growing mouse ESC

Timing: 3 days

This section describes how to grow mouse embryonic stem cells (ESCs).

Note: Use an appropriate number of replicates. Treat each biological replicates independently changing tips for each well.

Note: We use 6-well plates; all the volumes described here are relative to this size of well.

Note: All the 2i medium should be warmed up to 37°C in a water bath, before adding to the cells. Prepare the volume of 2i medium required for the experiment before starting the experiment.

Alternatives: This protocol is for mouse ESC but other types of adherent cells can be used, as well as in different conditions.

1. Seeding day.
   a. Wash each well of the ESC plate with 1 mL of 1× PBS, and aspirate.
   b. To detach the cells from the plate, add 0.5 mL Trypsin and incubate the cells for 5 min at 37 °C at 5% CO_2.
   c. Deactivate the Trypsin by adding 4× the volume added of Trypsin with 2i medium. Then collect the cells in a centrifuge tube.
   d. Centrifuge the cell suspension for 5 min at 400 g.
   e. Discard the supernatant and resuspend the cells with 5 mL of 2i medium.

f. Count the cells with a cell-counter and prepare the cell suspension of 1.5 × 10^5 cells/well to seed for the experiment.

Note: Make sure you have sufficient replicates for the four time-points (0 h, 2 h, 4 h, 8 h) of α-amanitin-inhibition treatment which will be done on the harvest day (day 3).

g. Coat new plates with 0.2% gelatin for 5–10 min at RT and aspirate.
h. Seed 1.5 × 10^5 cells/well onto gelatin-coated plates, add up to 2 mL of 2i media to each well and rock the plates to distribute the cell suspension evenly.
i. Incubate the cells 10–12 h in a humidified incubator (5% CO_2; 37°C).

2. **Day 1-day 2:** Check the cells under the light microscope and change the medium with fresh 2i medium.

Note: Medium change is done by removing the old medium and adding another 2 mL of fresh pre-heated (37°C) 2i medium in each well.

α-amanitin treatment and harvesting

Timing: approximately 10 h

This section describes how to perform transcriptional inhibition using α-amanitin.

Note: Check your cells under microscope and try to estimate the confluence of your plates. Should be around 70%–80% plate coverage. As the colonies begin to merge the chances for differentiation increase.

3. α-amanitin treatment:
   a. 1–2 h before the treatment, change to fresh 2i medium your control cells only (time-point 0 h).
   b. Prepare 2i medium with 2 μg/mL α-amanitin enough for all the plates, except your controls (without α-amanitin).
   c. Change to fresh 2i medium with α-amanitin added to your cells (2 h, 4 h and 8 h time-points, Figure 1A).
   d. Incubate the cells with 2i+α-amanitin medium at 37°C at 5% CO_2.

Note: Proceed immediately to the next step to harvest the control cells (time-point 0 h).

4. Harvesting: (0 h, 2 h, 4 h and 8 h).
   a. Wash the cells with 1× PBS.
   b. Add trypsin in each well and incubate for 5 min at 37°C.
   c. Deactivate the trypsin by adding 4× the volume added of trypsin with 2i medium. Resuspend the cells very gentle and collect them to centrifuge tube.
   d. Centrifuge the cells suspension for 5 min at 400 g.
   e. Discard the supernatant and resuspend the cells with 1 mL of PBS, transfer the cell suspension to a new collection tube (1.5 mL).

Note: Incomplete removal of the cell-culture medium will inhibit lysis and might dilute the lysate, both effects may reduce the RNA yield.

   f. Centrifuge the cells for 5 min at 400 g. and discard the supernatant.
   g. Add buffer RLT (lysis buffer from QIAGEN kit) and snap freeze the cells.
   h. Store the cells at −80°C until all the samples will be harvested and you will be ready to proceed.

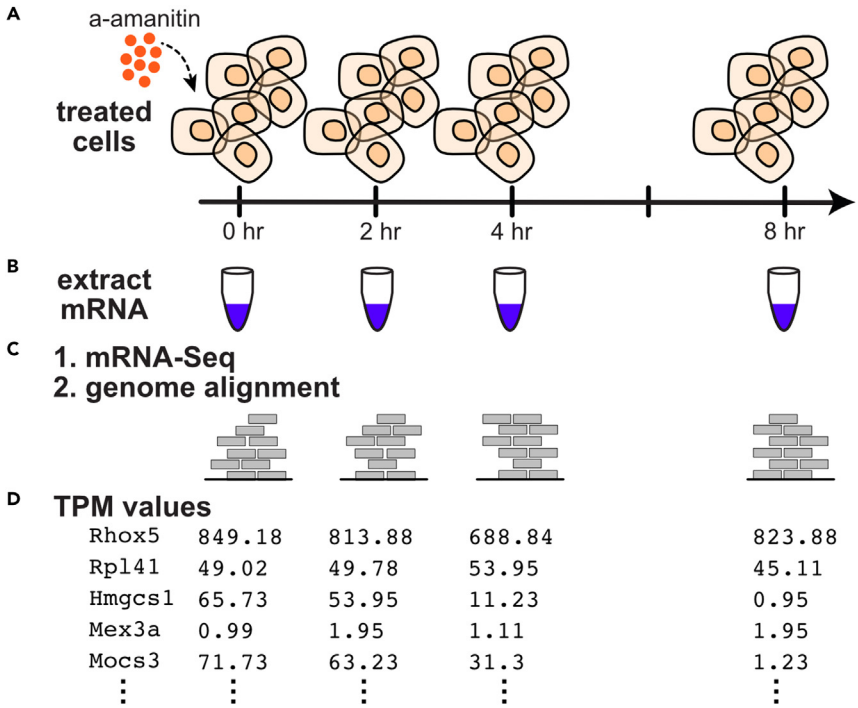


Figure 1. Cell treatment with α-amanitin and RNA-seq

(A) Treatment of cells with α-amanitin to inhibit transcription, followed by temporal samples (0 h, 2 h, 4 h, 8 h).

(B) RNA is extracted from each temporal sample.

(C) RNA is used for preparation of mRNA-seq library and sequencing. Resulting sequencing reads are aligned to the genome.

(D) Expression levels of genes between temporal samples are estimated from alignment by calculating TPM values.

RNA extraction and library preparation

Timing: approximately 2 days

This section describes how to extract and sequence RNA from temporal samples.

5. RNA extraction:
   a. If frozen, remove samples from −80°C and thaw on ice.
   b. In order to homogenize the cells, you can use a syringe with a needle. Pass the lysate through a 20-gauge needle for at least 5–10 times or until a homogeneous lysate is achieved.
   c. Proceed with the purification of total RNA (Figure 1B) by performing the protocol from RNeasy Mini Kit (QIAGEN).

   Note: We used the purification of total RNA from Animal cells using spin technology and the on-column DNase digestion in order to eliminate genomic DNA contamination, using RNAse-Free DNase set (QIAGEN).

   d. To quantify the amount of RNA and assess the quality of total RNA in each sample, use NanoDrop-2000 (Thermo Fisher Scientific), or equivalent. Dilute all samples to the minimal concentration obtained (typically around 100–250 ng/μL), and use 500 ng of each sample for library preparations.

   Alternatives: The concentration of RNA could also be determined by measuring the absorbance in a spectrophotometer. Agilent bioanalyzer or fluorometric quantification can also be used.

Pause point: RNA extracts can be stored at −80°C until further use.

6. Library preparation, sequencing and data processing.
   a. Prepare the indexed RNA libraries from 500 ng of total RNA extracted and purified at step 5. Note that while we used 500 ng, the starting amount can range from 100 ng to 4 μg of total RNA.
   b. Use KAPA Stranded mRNA-Seq kit by Kapa Biosystems, and follow according to the manufacturer's instructions (https://rochesequencingstore.com/wp-content/uploads/2022/07/KAPA-Stranded-mRNA-Seq-Kit-Technical-Data-Sheet.pdf).

   Alternatives: Other types of RNA-seq kits for library preparation can be use, for instance, Illumina Stranded mRNA prep.

   c. Determine the average size of the libraries by Tape Station (Agilent Technologies).
   d. Pool all samples and sequence on Illumina NextSeq500 sequencer or other equivalent instrument.
   e. Data processing (Figures 1C and 1D):
      i. Initial quality control (QC) checks can be done with FASTQC software. Each sample will generate a QC report, and you should examine the data for low per base quality, for unusual sequence content, if the sequences contain reads of uniform or variable lengths, then decide whether to trim the reads or not. Trimmomatic can be used to remove low-quality reads, trim adaptor sequences, and eliminate poor-quality bases.
      ii. Align the reads with STAR alignment pipeline or another preferred aligner. We used STAR (Spliced Transcripts Alignment to a Reference), version 2.7.1a. Aligning reads using STAR is a two-step process, which involves creating a genome index and mapping reads to the genome. As reference genome we used GRCm38/mm10. The following command can be used:

```
> STAR-2.7.1a/bin/Linux_x86_64/STAR--runThreadN4--genomeDir/ems/elsc-labs/meshorer-e/lab--shared/star_index_mouseN--readFilesIn$file.fastq--quantModeGeneCounts--outFileNamePrefix ./${file} --outSAMtype BAM SortedByCoordinate
```

Note: We excluded the mitochondrial chromosome from the alignment, to avoid possible biases due to different drug activity on mitochondrial RNA polymerases. In case drug used is expected to act similarly on mitochondrial RNA polymerases, then mitochondrial RNAs could be retained in the analysis.

      iii. The result of STAR alignment is an output folder containing read counts and bam file containing sequence alignment. You will see counts.csv, which are counts for all of your samples in a tab delimited file and tpm.csv which are normalized by TPM table. We used gencode.vM24.primary assembly.annotation.gtf.

Alternatives: You can choose to normalize your values by FPKM, which will only control for the number of sequencing reads per sample, but not total expression. By our experience, predicted degradation rates based on FPKM normalization are highly similar, but could result in slight biases in estimating highly stable genes, and lead to a larger number of control genes (see below).

Large-scale normalization of a-amanitin treatment time course

Timing: approximately 1 h

This section describes how to select control genes and normalize temporal samples.

7. Select a subset of stable genes for normalization.
   a. Fit an exponential decay model to TPM values relative to time of sample collection (Figure 2A).
      i. Set a minimal expression threshold on TPM values (we used a minimal value of 1e-2).
      ii. For each gene, use linear regression to estimate a gene-specific slope, which describes the change in expression levels over time, under the assumption of (1) an exponential decay model (first order reaction) and (2) a constant slope value (over time).
      Under these assumptions, we can write:

$$\frac{dX}{dt} = -\beta X \rightarrow X_t = X_0 \cdot e^{-\beta t} \rightarrow \log X_t = -\beta t + log\, X_0$$

      Where X is RNA level of a gene (TPM values), t is time of sample collection (after α-amanitin treatment) and β is the slope.
      To find β, we fit a generalized linear regression model (matlab glmfit function) to TPM values, which assumes a poisson distribution of the response variable (TPM values) and a logarithmic link function.

   Alternatives: fit a standard linear regression to log-transformed TPM values, which should produce equivalent results. This analysis assumes a normal distribution of the response variable (logTPM values).

   b. Set lower and upper bounds on estimated slopes (β). We used a lower bound of −1.5 and an upper bound of 1.5 (maximal degradation; equivalent to a half-life of 28 min). Positive values represent a decrease in expression, while negative values represent an increase in expression.
   c. Calculate a standard r-squared value (coefficient of determination) for the fit between model predictions and measured TPM values.
      R-squared calculation is done using the following formula:

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}} = 1 - \frac{\sum_i (\log TPM_t - (-\beta t + log\, X_0))^2}{\sum_i (\log TPM_t - \overline{\log TPM})^2}$$

      Where TPM_t is the TPM value of the gene at time t, β and $log\, X_0$ are fitted by the regression model, and $\overline{\log TPM}$ is the average of all temporal TPM values for the gene.
   d. Select a subset of control transcripts with a minimal degradation rate in wild-type examples by requiring:

$$(\text{slope} < 0.05)\ \text{AND}\ (\log 2(X_0) > 1)\ \text{AND}\ (\text{r} - \text{squared} > 0.5)$$

   e. These transcripts represent a subset of genes with a significant initial expression but minimal decrease in their mRNA levels during the course of the α-amanitin treatment, and will be used for normalization.
8. Normalizing the TPM values by the stable genes control set.
   a. For each time point, calculate the average logTPM value across the set of control genes, and subtract that value from all logTPM values of all genes, and add an arbitrary non-zero value (we used 2, Figure 2B).

The final formula for a gene ***X***, at time ***t*** is:

$$normalized\ logTPM_{x,t} = logTPM_{x,t} - logTPM_{controls,t} + 2$$

mRNA decay analysis

Timing: approximately 1 h

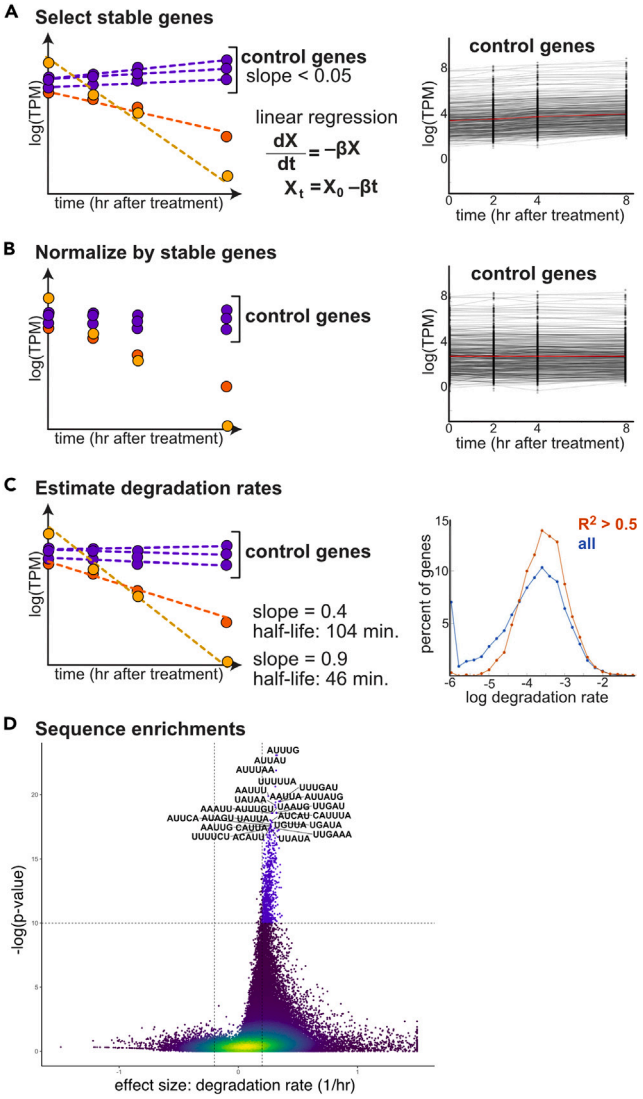


Figure 2. Temporal sample normalization and degradation analysis

(A) Left: example of fitted regression models (dashed lines) to raw TPM data (y-axis; log(TPM)) in temporal samples (x-axis; hours) of stable control genes (blue) and degraded mRNAs (yellow, orange). The slope fitted to the control genes is < 0.05, showing their stability. Right: raw TPM data (y-axis) in temporal samples (x-axis; hours) of 349 control genes selected for normalization of undifferentiated mouse ESCs.

(B) Left: example of normalized expression (y-axis) in temporal samples (x-axis; hours) after normalization by control genes (blue). The slope of the control genes is ~0 after normalization. Right: normalized expression (y-axis) in temporal samples (x-axis; hours) of 349 control genes of undifferentiated mouse ESCs.

(C) Left: example of fitted regression models (dashed lines) to normalized expression (y-axis; log(TPM)) in temporal samples (x-axis; hours) of stable control genes (blue) and degraded mRNAs (yellow, orange). The slope fitted to each gene estimates its degradation rate (shown as half-life). Higher values represent faster degradation. Right: distribution of all degradation rates (x-axis; log2) fitted to genes in undifferentiated mouse ESCs. Blue: all 5,294 expressed genes. Red: only 3314 genes with r-squared value > 0.5 for the fitted regression model.

(D) Volcano plot showing the effect size (difference in the standard mean of each of the two distributions, x-axis) and p-value (-log10(p-value), Kolmogorov-Smirnov FDR < 1% , y-axis; -log) for the association of short 5–8 nucleotide sequences with differences in degradation rates of genes. Negative effect size values (left side of plot) represent sequences associated with stable genes, while positive effect size values (right side of plot) represent sequences associated with unstable genes. Colors represent density (yellow = high density; dark blue = low density). P-value threshold (horizontal line) is 10^{-11}. Effect size threshold (vertical lines) is plus/minus 0.2. Significant sequences are marked in light blue and annotated on plot. All identified sequences are associated with unstable genes in undifferentiated mouse ESCs.

This section describes how to calculate mRNA degradation rates from normalized RNA-seq data.

9. Calculate degradation rate based on normalized TPM values.
   a. Fit an exponential decay model to normalized TPM values relative to time of sample collection (Figure 2C).
      i. For each gene, use linear regression to estimate a gene-specific slope, as described above (step 7.i.ii).
      Slope directly relates to degradation rate and half-life:

$$\text{degradation rate} = \beta$$

$$\text{half life} = \frac{\log(2)}{\beta}$$

      Resulting rate (1/time) and half-life (time) are given in the same time units as the input times. In this example, times are given in hours, and therefore half-life is given in hours and degradation rate in 1/h.
   b. Set lower and upper bounds on estimated degradation rate (β). We used a lower bound of 0.01 (minimal degradation; equivalent to half-life of 69 h) and an upper bound of 4 (maximal degradation; equivalent to a half-life of 10 min).

Optional: Calculate a standard r-squared value for the fit between model predictions and normalized TPM values. This value will estimate the percent of variation in a gene's normalized TPM values that is explained by the model. The higher the value the better the model explains the data.

Note: Download our degradation rate package, and run the following code from a linux/unix terminal to perform the analysis in steps 7–9:

```
> make run_dg_estimate TPM_FILE=example_ko_tpm.txt OUTPUT_DIR=example_ko
```

TPM_FILE is a tab delimited text file containing TPM values, where each row represents a gene and each column represents a sample (times 0, 2, 4 and 8 and replicates). First column should include gene ids followed by TPM values as numbers. First row should include the time value for each column as numbers. The resulting half-life values will be given in the same time unit as the column titles in this file. We provide two matrices as example within the code distribution. The results will be placed in the OUTPUT_DIR directory within the run folder.

Associating short 3′ UTR sequences with mRNA decay rates

Timing: approximately 1 h

This section describes how to calculate associate short 3′ UTR sequences with differences between genes in calculated mRNA degradation rates.

10. Prepare a reference set of 3′ UTR sequences.
    a. Download mouse 3′ UTR sequences. We used the ensembl biomart GRCm38, version 99.[4]
    b. For each gene, filter the annotated 3′ UTRs to keep a single longest annotated 3′ UTR sequence. Remove from the analysis 3′ UTR sequences which are shorter than 10 nucleotides.
11. Identify short sequences that are over-represented within the 3′ UTR sequences of genes with either faster or slower degradation rates.
    a. Calculate the set of all short sequences between 5–8 nucleotides long (k-mers). We used the *ape* and *k-mer* R packages.
    b. For each k-mer,

i. Find the set of all genes with this k-mer in their 3′ UTR (**positive set**). It is recommended to include only genes with r-square > 0.5 for the fitted degradation model in this set.
ii. Find the set of all genes without this k-mer in their 3′ UTR (**negative set**). It is recommended to include only genes with r-square > 0.5 for the fitted degradation model in this set.
iii. Use a one-sided Kolmogorov-Smirnov test to calculate p-values for two hypotheses: (1) degradation rates in the positive set are *larger* than the negative set, and (2) degradation rates in the positive set are *smaller* than the negative set. Select the minimal p-value.
iv. Calculate an effect size by the standardized mean difference, defined as:

$$\theta = \frac{\mu_1 - \mu_2}{\sigma}$$

where μ_1 is the mean of the first population, μ_2 is the mean of the second population and σ is the standard deviation (based on both populations).

c. Use a False Discovery Rate correction of all p-values (2 p-values per k-mer) and set a threshold for significance (e.g., corrected p-value < 0.05).
d. Set a threshold for significance on the absolute value of the effect size (we require a minimal absolute effect size of 0.2).
e. Select k-mers with a p-value below the significance threshold and an effect size above the significance threshold as candidate regulatory sequences (Figure 2D).

Note: Download our degradation rate package, and run the following code from a linux/unix terminal to perform the complete analysis in steps 7–11:

```
> make run_complete_analysis input_tpm_file=example_ko_tpm.txt out_dir=example_ko ensembl_
organism=mmusculus_3UTR_ensembl99
```

TPM_FILE is a tab delimited text file containing TPM values calculated in step 6. The name of the organism should be specified in order to correctly analyze 3′ UTR sequences. Our package contains 3′ UTR sequence information for the mouse, human and zebrafish genomes. The results will be placed in the out_dir directory within the run folder.

Note: You can also run the k-mer analysis (step 11) separately. Download our k-mer analysis package, and use the following code to perform the analysis in this section from a linux/unix terminal:

```
> make run_ks_test_job path_to_kmer_matrices=3utr/kmer_matrices PARAMETERS_FILE=parameters.
tsv output_path=3utr/kmer_out term=half_life

> make ks_tests_plots ks_test_output_folder=3utr/kmer_out term=half_life output_pa-
th=3utr/kmer_out/plots
```

The kmer_matrices directory contains 3′ UTR sequence information of the relevant organism. Our package contains 3′ UTR sequence information for the mouse, human and zebrafish genomes. Instructions for adding information of other organisms are provided online. Parameter file contains the half-life estimations from step 9.
The results will be placed in the output_path directory within the run folder.

EXPECTED OUTCOMES

This protocol provides a methodology and tools for a reliable quantification of mRNA degradation rates in undifferentiated mouse ESCs. We perform a transcriptional shutoff by α-amanitin, followed by collecting 4 samples for mRNA sequencing along a time course. We provide an approach for large-scale normalization by a subset of stable transcripts in the data, which allows to compare temporal samples and analyze the relative decrease in mRNA levels for each gene, leading to reliable estimates of mRNA degradation. Using this quantification, we identify short (5–8 nucleotide)

sequences that are over-represented within the 3′ UTR sequences of genes with either faster or slower degradation rates. We suggest these sequences as candidate regulatory elements of mRNA degradation.

LIMITATIONS

We present a protocol for a reliable quantification of mRNA degradation rates. While this protocol provides reliable estimates for mRNA degradation in undifferentiated mouse ESCs, there are still some limitations in its application.

First, this protocol relies on transcriptional shutoff by α-amanitin. Inhibiting transcription is a significant stress for the cell, and in response cells may change mRNA stability, localization or translation.[5] Thus, half-lives obtained with such inhibitors may not reflect unperturbed cellular degradation rates for some genes.

Second, our analysis relies on hundreds of transcripts which remain stable following α-amanitin treatment in undifferentiated mouse ESCs. Therefore, estimated rates are in fact relative to this set. Other systems might have a smaller number of such transcripts, and could require adjustment of thresholds for selection of the normalization reference set.

Finally, the resolution and range of mRNA degradation rates is different in different biological systems. For example, an average half-life is of several minutes for yeast mRNAs,[6] and a few hours for mammalian mRNAs.[7] Thus, mRNA levels of some transcripts could quickly reach lower levels below detection threshold. In such cases, linear regression could be biased and not reflect degradation rates accurately. As the resolution and range of time course samples directly affects the resolution and range of estimated degradation rates, adjusting and calibrating sampling times for each system could minimize such effects. Later samples beyond 10 h after treatment could be collected to measure slower degradation rates. Earlier and more frequent sampling will allow to dissect smaller differences between degradation rates.

TROUBLESHOOTING

Problem 1

Cell detachment or reduced cell attachment (in step 2).

Potential solution

Make sure you have the plates well coated for better adherence, before you seed the cells. Partial coverage will cause poor or unequal cell attachment. Also, check before the α-amanitin treatment, the cell confluency does not exceed 80% of the plate for mESC or 90% for other adherent cell types. Very dense confluences will cause cell detachment.

Problem 2

No control genes selected (in step 7).

Potential solution

The most likely reason is that no genes were fitted to a linear model with a slope below the current threshold of 0.05. Adjusting to a higher threshold could solve the problem.

Problem 3

Low regression coefficients (absolute values) of linear regression models fitted to temporal measurements (in step 9).

Potential solution

This indicates that expression levels of most genes do not follow a linear model, as expected by an exponential decay (first order reaction).

A common reason for globally low regression coefficients is a technical problem with one of the temporal samples. Several factors can help to evaluate the quality of sequencing samples, and identify samples that might need to be removed from the analysis. First, a low total number of reads in one sample compared to other temporal samples. Second, a sample with low numbers of mapped reads compared to other temporal samples. Third, a sample with a low correlation of expression values (TPM) across all genes to other temporal samples.

In these cases, it is recommended to remove the low-quality sample from the regression analysis.

Problem 4

Negative slopes of linear regression models fitted to temporal measurements after normalization (in step 9).

Potential solution

A negative slope in our model indicates that the mRNA levels of a gene increased. However, the expression levels of genes in a transcriptional shut-off experiment are not expected to increase if the data was correctly normalized. Therefore, negative slopes could suggest a problem with selection of genes for the control set used for normalization.

Selection of the control set for normalization could be optimized. Adjusting the threshold for selecting the control genes to a higher value could reduce such biases, since the most strongly increasing genes would be normalized.

Problem 5

Short regulatory sequences not found (in step 11).

Potential solution

Before concluding that there are no short regulatory sequences associated with degradation rates, a few optional adjustments could be tested.

- Change the range of k-mer lengths. Minimal length could be reduced from 5 to 3 nucleotides. Maximal length could be increased beyond 8 nucleotides (but running times will become significantly longer).
- Use an alternative statistical test, such as a t-test (if underlying distribution is normal), or the Mann-Whitney test (which should be more sensitive to detecting differences in mean between two samples).
- Require a lower threshold on the effect size.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Michal Rabani (michal.rabani@mail.huji.ac.il).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The source code for the analyses, together with the example data-files, is freely available and can be found on Zenodo, for degradation rate calculation[8] and for kmer analysis.[9]

ACKNOWLEDGMENTS

This work was supported by the Israel Ministry of Science (MOST 0004272 to E.M.), the European Innovation Council (EIC) Horizon Europe (RT-SuperES 101099654 to E.M.), and the European

Research Council (ERC) Horizon 2020 grant (852451 to M.R.). E.M. is the incumbent of the Arthur Gutterman Professor Chair for Stem Cell Research. J.V. is supported by the European Union's Horizon-2020 Marie Skłodowska Curie EpiSyStem ITN Network (765966).

AUTHOR CONTRIBUTIONS

J.O.V., E.M., and M.R. wrote the manuscript. M.R. and L.F. conceived the computational approach and wrote the code for bioinformatics and statistical analyses. E.M. and M.R. supervised the project.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Viegas, J.O., Azad, G.K., Lv, Y., Fishman, L., Paltiel, T., Pattabiraman, S., Park, J.E., Kaganovich, D., Sze, S.K., Rabani, M., et al. (2022). RNA degradation eliminates developmental transcripts during murine embryonic stem cell differentiation via CAPRIN1-XRN2. Dev. Cell *57*, 2731–2744.e5. https://doi.org/10.1016/j.devcel.2022.11.014.

2. Andrew, S. (2010). FastQC - A Quality Control Tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

3. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/BIOINFORMATICS/BTS635.

4. Howe, K.L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., Bhai, J., et al. (2021). Ensembl 2021. Nucleic Acids Res. *49*, D884–D891. https://doi.org/10.1093/NAR/GKAA942.

5. Bensaude, O. (2011). Inhibiting eukaryotic transcription. Which compound to choose? How to evaluate its activity? Transcription *2*, 103–108. https://doi.org/10.4161/trns.2.3.16172.

6. Wang, Y., Liu, C.L., Storey, J.D., Tibshirani, R.J., Herschlag, D., and Brown, P.O. (2002). Precision and functional specificity in mRNA decay. Proc. Natl. Acad. Sci. *99*, 5860–5865. https://doi.org/10.1073/pnas.092538799.

7. Rabani, M., Levin, J.Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N., et al. (2011). Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. Nat. Biotechnol. *29*, 436–442. https://doi.org/10.1038/nbt.1861.

8. Fishman, L. (2023). Rabanilab/RNA_deg_ESC:version 1. https://doi.org/10.5281/zenodo.8176716.

9. Fishman, L. (2023). Rabanilab/cont_kmer_analysis: Version 1. https://doi.org/10.5281/zenodo.8176720.