Poly(A) Tail Regulation in the Nucleus

DISSERTATION
zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

eingereicht an der
Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin

von
B.Sc. Jonathan Alles

Präsident (komm.)
der Humboldt-Universität zu Berlin

Prof. Dr. Peter Frensch

Dekan der Lebenswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

Prof. Dr. Dr. Christian Ulrichs

Gutachter/innen

1. Prof. Dr. Nikolaus Rajewsky
2. Prof. Dr. Markus Landthaler
3. Prof. Dr. Elmar Wahle

Tag der mündlichen Prüfung
21.03.2022

Dedicated to my parents Ingrid & Ralf

Abstract

The RNA metabolism involves different steps from transcription to translation and decay of messenger RNAs (mRNAs), and those different layers are each regulated and interconnected to precisely control mRNA and protein abundance. Most mRNAs have a poly(A) tail attached to their 3'-end, which protects them from degradation and stimulates translation. Complete removal of the poly(A) tail, a process termed deadenylation, is thereby the rate-limiting step in RNA decay and is dynamically regulated to control transcript stability and translation. mRNA decay has so far been mostly investigated in the cytoplasmic context, and it is unclear if and to what extent RNA deadenylation and decay occur in the nucleus. Investigating the function of poly(A) tails in a genome-wide context was so far limited by the lack of simple methods for measuring poly(A) tails for each gene, since previous approaches involved complicated experimental protocols and customization of sequencing hard- and software.

A novel method for genome-wide determination of poly(A) tail length, termed FLAM-Seq, was hence developed, enabling genome-wide analysis of complete RNAs, including their poly(A) tail sequence based on third generation sequencing. FLAM-Seq analysis of cell lines, organoids and *C. elegans* samples uncovered a strong correlation between poly(A) tail and 3'-UTR length and many genes for which alternative isoforms of the same gene were associated with significant differences in poly(A) tail length. Investigating the nucleotide content across poly(A) tails showed that cytosines were significantly enriched in poly(A) tails.

Investigating poly(A) tails of unspliced RNAs from FLAM-Seq data revealed the genome-wide synthesis of poly(A) tails with a length of more than 200 nt. This could be validated by splicing inhibition experiments which uncovered potential links between the completion of splicing and poly(A) tail shortening. Measuring RNA deadenylation kinetics using metabolic labeling experiments hinted at a rapid shortening of tails within minutes. The analysis of subcellular fractions obtained from HeLa cells and a mouse brain showed that initial deadenylation is a nuclear process. Nuclear deadenylation is gene specific and poly(A) tails of lncRNAs retained in the nucleus were not shortened.

To identify enzymes responsible for nuclear deadenylation, RNA targeting Cas-systems, siRNAs and shRNA cell lines were used to perturb expression of PAN2-PAN3, CCR4-NOT and PARN deadenylases. Despite efficient mRNA knockdown, subcellular analysis of poly(A) tail length by did not cause any molecular phenotypes on nuclear poly(A) tail length that could be linked to individual deadenylase complexes.

Zusammenfassung

Der Ribonukleinsäure (RNS) Stoffwechsel umfasst verschiedene Schritte, beginnend mit der Transkription der RNS über die Translation bis zum RNA Abbau. Die verschiedenen Ebenen sind verbunden und haben jeweils bestimme Kontrollmechanismen, welche an Ende die Produktion und Verfügbarkeit von Proteinen kontrollieren. Poly(A) Schwänze befinden sich am Ende der meisten der Boten-RNS und am Ende mancher nicht-kodierender langer RNS. Poly(A) Schwänze schützen die RNA vor Abbau und stimulieren deren Translation. Die Deadenylierung von Poly(A) Schwänzen ist dabei der limitierende Schritt für den Abbau von RNS und Deadenylierung kann dynamisch reguliert werden was die Stabilität von Transkripten beeinflusst. Bisher wurde RNS Abbau meist im Kontext von cytoplasmatischen Prozessen untersucht, ob und wie RNS Deadenylierung und Abbau in Nukleus erfolgen ist bisher unklar. Die bisher verfügbaren Methoden um Poly(A) Schwänze für jedes Gen zu messen waren bisher limitierend, und erforderten komplizierte Eingriffe in die verfügbare Sequenziertechnologie.

Eine neue Methode zur genomweiten Bestimmung von Poly(A) Schwanzlänge wurde deshalb entwickelt, welche die Analyse kompletter RNS Moleküle inklusive der Poly(A) Länge und Sequenz ermöglicht. Die Methode wurde FLAM-Seq genannt. FLAM-Seq wurde verwendet um Zelllinien, Organoide und *C. elegans* RNS zu analysieren und es wurde eine signifikante Korrelation zwischen 3'-UTR und Poly(A) Länge gefunden. Für viele Gene wurden 3'-UTR Isoformen identifiziert, welche mit signifikanten Unterschieden in den assoziierten Poly(A) Profilen einhergingen. Weiterhin wurde Cytosin als das zweihäufigste Nukleotid in Poly(A) Schwänzen identifiziert.

Die Untersuchung von Poly(A) Schwänzen von nicht-gespleißten RNS Molekülen zeige, dass deren Poly(A) Schwänze eine Länge von mehr als 200 nt hatten. Diese Analyse wurde durch eine Inhibition des Spleiß-Prozesses validiert, wodurch auch potenzielle Zusammenhänge zwischen Spleißen und Deadenylierung gefunden wurden. Methoden zur Markierung von RNS, welche die zeitliche Auflösung der RNS Prozessierung ermöglicht, deutete auf eine Deadenylierung der Poly(A) Schwänze schon wenige Minuten nach deren Transkription hin. Die Analyse von subzellulären Fraktionen aus HeLa Zellen und einem Maus Gehirn zeigte, dass diese initiale Deadenylierung ein Prozess im Nukleus ist. Dieser Prozess ist gen-spezifisch und Poly(A) Schwänze von bestimmten Typen von Transkripten, wie nuklearen langen nicht-kodierende RNS Molekülen waren nicht deadenyliert.

Um Enzyme zu identifizieren, welche die Deadenylierung im Zellkern katalysieren, wurden verschiedene experimentelle Methoden wie RNS-abbauende Cas Systeme, siRNAs oder shRNA Zelllinien verwendet um die Genexpression der Enzym(-komplexe) PAN2-PAN3, CCR4-NOT und PARN zu reduzieren. Trotz einer effizienten Reduktion der Boten-RNS Expression konnten keine molekularen Phänotypen identifiziert werden welche die Poly(A) Länge im Zellkern beeinflussen.

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig verfasst habe und sämtliche Quellen, einschließlich Internetquellen, die unverändert oder abgewandelt wiedergegeben werden, insbesondere Quellen für Texte, Grafiken, Tabellen und Bilder, als solche kenntlich gemacht habe.

Diese Arbeit oder Teile davon wurden bei keiner anderen wissenschaftlichen Einrichtung eingereicht, angenommen oder abgelehnt. Ich besitze keinen Doktorgrad. Die dem Promotionsverfahren zugrunde liegende Promotionsordnung habe ich zur Kenntnis genommen. Weiterhin erkläre ich, dass keine Zusammenarbeit mit gewerblichen Promotionsberaterinnen/-beratern stattgefunden hat und dass die Grundsätze der Humboldt-Universität zu Berlin zur Sicherung guter wissenschaftlicher Praxis eingehalten wurden.

Berlin, 18.08.2021

List of publications

Legnini, I.*, **Alles, J**.*, Karaiskos, N., Ayoub, S. & Rajewsky, N. FLAM-seq: full-length mRNA sequencing reveals principles of poly(A) tail length control. *Nat. Methods* **16**, 879–886 (2019). * Equal contribution

Alles J.*, Karaiskos N.*, Praktiknjo S.D.*, Grosswendt S., Wahle P., Ruffault P.L., Ayoub S., Schreyer L., Boltengagen A., Birchmeier C., Zinzen R.P., Kocks C. and Rajewsky N. Cell fixation and preservation for droplet based single-cell transcriptomics. *BMC Biology.*15:44 (2017) * Equal contribution

Karaiskos N., Wahle P., **Alles J.**, Boltengagen A., Ayoub S., Kipar C., Kocks C., N., Zinzen R.P. The Drosophila embryo at single-cell transcriptome resolution. *Science* 358 (6360): 194-199 (2017)

Acknowledgements

I would first like to thank my supervisor Prof. Dr. Nikolaus Rajewsky for giving me the opportunity to work on challenging and exciting projects, for continuous support and motivation and opening many doors for my personal development beyond scientific work.
I would like to thank all previous colleagues and members of the Rajewsky Lab for creating a great working atmosphere and a lot of fun. I would also like to thank all collaborators on projects which could not be presented in this thesis.
I would like to thank Dr. Ivano Legnini, with whom I collaborated intensely on the work presented in this thesis, for great discussions, constructive feedback, and mutual support in driving the projects forwards. I also would like to thank Dr. Nikos Karaiskos and Salah Ayoub for working together on the FLAM-Seq publication, as well as Jonathan Fröhlich and Dr. Agnieszka Rybak-Wolf for kindly providing organoid and *C. elegans* RNA.
I would like to thank Margaretha Herzog, Anastasia Boltengagen, Gwendolin Thomas, Salah Ayoub, Sarah Tagliaferro and Marie Schott for technical support and keeping the lab running.
I would also like to thank Claudia Quedenau from the MDC Genomics Core Facility for help with PacBio sequencing.
I would like to thank Maddalena Pacelli for great support and contributing to the second part of this thesis.
I would like to thank the spatial transcriptomics team for continuously working on a very challenging project and many intense discussions.
I would like to thank Prof. Dr. Markus Landthaler and Dr. Marina Chekulaeva for being members of my thesis advisory committee, providing feedback and new ideas for my projects.
I would like to thank Alex Tschernycheff and Dr. Grietje Krabbe for a lot of administrative support.
I would like to thank the MDC-NYU exchange program for funding and the exciting opportunity to work in New York. I would especially like to thank Dr. Carlos Carmona Fontaine and Logan Schachtner for welcoming me in New York and an exciting collaborative project on T cell biology.
I would like to thank 10x Genomics for grating me the opportunity to work at their Stockholm R&D site. In particular I would like to thank Dr. Caroline Gallant and Dr. Marlon Stoeckius for supervision and guidance.
Finally, I would like to thank my friends and family for their support and encouragement.

Author Contributions

The work presented in this thesis conducted at the Max Delbruck Center for Molecular Medicine in Berlin, Germany. The work was performed and documented according to Good Scientific Practice. Relevant literature sources were cited accordingly. A basic level of understanding of (molecular) biology ("textbook knowledge") was assumed for the reader and general concepts of biology (e.g. RNA, nucleotides, …) were not explicitly explained. The reader may here refer to standard literature such as Principles of Biochemistry (Lehninger, A. L., Nelson, D. L., & Cox, M. M. (2013). *Lehninger principles of biochemistry*. New York: Worth Publishers.) or Molecular Biology of RNA (Elliot D., Ladomery M. (2016) *Molecular Biology of RNA* Oxford University Press).

The first part of the thesis is related to development of the FLAM-Seq method, which was published in Nature Methods, until chapter 4.1.6. was performed in collaboration with Ivano Legnini, Nikos Karaiskos and Salah Ayoub from the Max Delbruck Center. The experimental FLAM-Seq protocol was developed and optimized by Jonathan Alles, Ivano Legnini and Salah Ayoub, involving optimization of reaction conditions, purification steps optimization of primers and reagents. Ivano Legnini prepared sequencing libraries from HeLa S3, organoids, iPS cells and *C. elegans* samples. *C. elegans* RNA was provided by Jonathan Fröhlich and brain organoid RNA by Agnieszka Rybak-Wolf. Nikos Karaiskos implemented the seed extension algorithm for poly(A) quantification and proposed nucleotide frequency counts after 5'- or 3'-alignments of poly(A) tails. Ivano Legnini wrote a script for appending poly(A) tail length to alignment files and performed PAT-assays for validation of poly(A) tail length.

The second part of this thesis, starting from chapter 4.1.7., was performed in collaboration with Ivano Legnini and Maddalena Pacelli. Ivano Legnini developed a first concept for identification of unspliced reads from FLAM-Seq datasets, which as been refined and extended. Ivano Legnini further generated shRNA-inducible cell lines for performing PAN3, CNOT7 and PARN knockdowns as well as cloning of guide RNAs for Cas13b and CasRx transfections. Maddalena Pacelli performed biochemical fractionation experiments in mouse brains and library preparation.

All other experiments and analysis were performed by Jonathan Alles. All figures in the 'Results' chapter were produced by Jonathan Alles, except Figures 4A, 4D and 6A, which were adopted from Legnini et al. 2019 as denoted in the legend.

Table of Content

List of Figures

Abbreviations

°C	degree Celsius
µJ	micro Joule
4E-T	Eukaryotic translation initiation factor 4E transporter
4sU	4-Thiouridine
5-EU	5-Ethynyl Uridine
5-mC	5-Methylcytosin
A,T,G,C	Adenine, Thymin, Guanine, Cytosine
ActD	Actinomycin D
AGO	Argonaute
Air1p/Air2p	Arginine methyltransferase-interacting RING finger protein ½
AK2	Adenylate Kinase 2
ALYREF	Aly/REF Export Factor
ARE	AU-rich elements
AU-rich	Adenosine / Uridine-rich
BCAP31	B-cell receptor-associated protein 31
BLAST	Basic Local Alignment Search Tool
BTF3	Basic Transcription Factor 3
BTG2	BTG Anti-Proliferation Factor 2
CAF1	Carbon Catabolite Repressor Protein (CCR4)-Associative Factor 1
CAGE	Cap Analysis of Gene Expression
CAMKII	Ca2+/calmodulin-dependent protein kinase II
CBC	Cap Binding Complex
CCR4	C-C Motif Chemokine Receptor 4
CCS	Circular Consensus Sequence
CD47	Leukocyte Surface Antigen CD47
CDK9	Cyclin Dependent Kinase 9
cDNA	complementary DNA
CDS	Coding Sequence
CF I/IIm	Cleavage Factor I/IIm
c-Fos	Fos Proto-Oncogene
Cid1 (gene)	Polyadenylate-binding protein-interacting protein 7
circRNA	Circular RNA
CNOT1	CCR4-NOT transcription complex subunit 1
CNOT7	CCR4-NOT transcription complex subunit 7
CNOT8	CCR4-NOT transcription complex subunit 8
CPA	Cleavage and Polyadenylation Complex
CPEB	Cytoplasmic Polyadenylation Element-Binding protein
CPEs	Cytoplasmic Polyadenylation Element
CpG	Cytosine-Guanosine
CPSF	Cleavage Polyadenylation Specificity Factor
CPSF160	Cleavage Polyadenylation Specificity Factor 160 kDa subunit
CPSF2	Cleavage Polyadenylation Specificity Factor 100 kDa subunit
CPSF73	Cleavage Polyadenylation Specificity Factor 73 kDa subunit
C-rich	Cytosine-rich
CstF	Cleavage Stimulation Factor
Cstf-77	Cleavage Stimulation Factor 77 kDa subunit
Ct value	Cycle Threshold value
CTD	C-terminal Domain
CTTN	Src substrate cortactin
CUT	Cryptic Upstream Transcript
DBP5	DEAD box protein 5
DCP1/DCP2	mRNA-decapping enzyme subunit 1/2
DCt value	Delta Ct value
DDCt value	Delta Delate Ct value
DIS3L2	DIS3-like exonuclease 2
Dis3p	Exosome complex exonuclease RRP44 Exosome complex exonuclease RRP44
DMEM	Dulbecco's Modified Eagle Medium
DMSO	Dimethylsulfoxide
DNA	Desoxyribonucleic Acid
dNTPs	Deoxynukleosidtriphosphate
Dox	Doxycycline
DRB	5,6-Dichlorobenzimidazole 1-β-D-ribofuranoside
dT, dC, dG, dA	Desoxythymidin, Desoxycytidine, Desoxyguaninen, Desoxyadenine
DTT	Dithiothreitol

dTTP	Desoxythimidine triphosphate
dUTP	Desoxyuridine triphosphate
ECL	Amersham™ ECL Select™
eIF4E	Eukaryotic translation Initiation Factor 4 E
eIF4G	Eukaryotic translation Initiation Factor 4 G
EJC	Exon Junction Complex
ELAV	Embryonic Lethal Abnormal Visual Protein
ENE	Expression and Nuclear retention Element
ER	Endoplasmic Reticulum
ERCC	External RNA Controls Consortium
eRNAs	Enhancer RNA
ESE	Exonic Splice site Enhancer
EtOH	Ethanol
f (filter value)	variable f
FAM	Fluorescein
FBS	Fetal Bovine Serum
FC	Fold Change
FDR	False Discovery Rate
FISH	Fluorescence In-Situ Hybridization
FLAM-Seq	Full-Length mRNA and poly(A) tail sequencing
g	Gravitation force; Unit for Acceleration
GAPDH	Glycerinaldehyd-3-phosphat-Dehydrogenase
GC-rich	Guanosine/Cytosine-rich
GEO	Gene Expression Omnibus
GI-tailing	Guanosine-Inosine-tailing
GLD-2	Defective in germ line development protein 2
GO term	Gene Ontology Term
GW182	Glycine-tryptophan protein of 182 kDa
HDE	Histone Downstream Sequence Element
HEK (cell line)	Human Embryonic Kidney cell line
hFip1	Factor interacting with PAP
hnRNP	Heterogeneous nuclear ribonucleoprotein
HOTTIP	HOXA Distal Transcript Antisense RNA
HOX	Homeobox protein
HRP	Horseraddish-Peroxidase
HuR	Hu-antigen R
IAA	Iodoacetamide
IDR	Intrinsically Disordered Region
IEG	Immediate Early Gene
IL-2	Interleukin-2
IMP4	Interacting with MPP10 protein 4
iPS cells	Induced Pluripotent Stem cells
JRE	Janus Response Element
Kbp	Kilobasepair
KD	Knockdown
L4 stage	Larval stage 4
lin-41	Abnormal cell lineage protein 41
lncRNA	Long non-coding RNA
Log2	Logarithm base 2
Lsm1-7	Sm-like
M, mM	Molar, milli Molar
m6A	N6-Methyladenosine
MALAT1	metastasis associated lung adenocarcinoma transcript
Mbp	Mega basepairs
Mex67p	mRNA export factor MEX67
Min, sec, s, h	minute, second
miRNA	microRNA
Mlp1	Muscleblind Like Splicing Regulator 2
MMLV RTase	Moloney Murine Leukemia Virus Reverse Transcriptase
MPG buffer	Magnetic Porous Glass buffer
MT-CO1	Cytochrome C Oxidase subunit 1
MT-CO2	Cytochrome C Oxidase subunit 2
mTOR	mechanistic Target of Rapamycin
Mtr4p	mRNA transport regulator MTR4
N (variable)	Number (of replicates)
Nab2	Nuclear polyadenylated RNA-binding protein NAB2
NEXT	nuclear exosome targeting complex
ng/µL	nanograms per microliter
nm	nanometer

NMD	Nonsense-Mediated Decay
Not	Negative regulator of transcription
NPC	Nuclear Pore Complex
nt	nucleotides
NUFIP2	Nuclear fragile X mental retardation-interacting protein 2
ORFs	Open Reading Frame
PABP2	Poly(A) Binding Protein 2
PABPC1	Cytoplasmic Poly(A) Binding Protein 1
PABPN1	Nucleoplasmic Poly(A) Binding Protein 1
PAL-Seq	poly(A)-tail length profiling by sequencing
PAP	Poly(A) Polymerase
Pap1p	Poly(A) Polymerase 1
PAPD1	PAP-associated domain-containing protein 1
PARN	Polyadenylate-specific ribonuclease
PAS	Polyadenylation Signal
PAT Assay	Poly(A) Tail-Length Assay
PAXT	poly(A) tail exosome targeting
P-bodies	Processing Bodies
Pcf11	Pre-mRNA cleavage complex II protein Pcf11
PCR	Polymerase Chain Reaction
PD	Pulldown
PDE12	2',5'-phosphodiesterase 12
PI3K	Phosphoinositid-3-Kinasen
PI4,5P2	Phosphatidylinositol-4,5-bisphosphate
PKC	Protein Kinase C
PNPase	Polyribonucleotide nucleotidyltransferase
PPD	PABPN1 and PAP mediated decay
pre-mRNA	pre-messenger RNA
PROMPT	Promoter upstream transcripts
qPCR	quantitative PCR
r	Pearson correlation coefficient
RBP	RNA-Binding Protein
RBP7	Retinoid-binding protein 7
RISC	RNA-Induced Silencing Complex
RNA	Ribonucleic Acid
mRNA	messenger RNA
RNAP	RNA polymerase
RNA-Seq	RNA sequencing
RNGTT	mRNA-capping enzyme
RNMT	mRNA cap methyltransferase
RNP	Ribonucleoprotein
RPB1	DNA-directed RNA polymerase II subunit RPB1
RPL37	Large ribosomal subunit protein eL37
RPS28	Small ribosomal subunit protein eS28
rRNA	Ribosomal RNA
Rrp6p	Ribosomal RNA-processing protein 6
RT primer	Reverse Transcription primer
RT	Room Temperature
RT-PCR	Reverse Transcription Polymerase Chain Reaction
SAGE	Serial Analysis of Gene Expression
scaRNA	Small Cajal body specific RNA
SCD	Stearoyl-CoA desaturase
sd	standard deviation
dsDNA	double stranded DNA
Ser2	Serine Position 2
SF1	Splicing Factor 1
SF3B1	Splicing Factor 3b Subunit 1
shRNA	Small Hairpin RNA
siRNA	Small Interfering RNA
SLAM-seq	SH-Linked Alkylation for the Metabolic Sequencing of RNA
SLBP	Stem Loop Binding Protein
SL-RNA	Splice Leader RNA
SMRT	Single Molecule Real Time
SN	Supernatant
snoRNA	Small Nucleolar RNA
snRNA	Small Nuclear RNA
snRNP	Small Nuclear Ribonucleoprotein
S-phase	Synthesis phase
SR protein	Serine-Arginine protein

ssDNA	Single Stranded DNA
STAR	Spliced Transcript Alignments to a Reference
Star-PAP	Speckle targeted PIP5K1A-regulated poly(A) polymerase
Strep-HRP	Streptavidin-Horseradish Peroxidase
T1 / T2 / WZ	Parameter names
$t_{1/2}$	half life time
TAP	Transporter associated with antigen processing
TBP	TATA-Box Binding Protein
TCR	T-Cell Receptor
TDP43	Transactive response DNA binding protein 43 kDa
TE	Translation Efficiency
TENT4A	Terminal nucleotidyltransferase 4A/B
TFII	General transcription factor II-I
TFIID	General transcription factor II-I D
TIR1	Transport inhibitor response 1
TNF-α	Tumor Necrosis Factor alpha
TOP	Terminal Oligopyrimidine
TR	Telomer RNA
TRAMP	Trf4/Air2/Mtr4p Polyadenylation complex
TREX	TRanscription and EXport
Trf4p	Topoisomerase 1-related protein TRF4
Trf5p	Topoisomerase 1-related protein TRF5
tRNA	Transfer RNA
TSO	Template Switch Oligo
TSS	Transcription Start Site
TTP	Tristetraprolin
TUT	Terminal Uridylytransferase
U2AF35	U2-Auxiliary Factor 35
U2AF65	U2-Auxiliary Factor 65
UHRR	Universal Human Reference RNA
µL / mL	micro liter / milli liter
UMI	Unique Molecular Identifier
UTR	Untranslated Regions
UV	Ultra Violet
vol	Volume
XIST	X-inactive specific transcript
Xrn1	5'-3' exoribonuclease 1/2
Yra1	RNA annealing protein YRA1
ZBP1	Zipcode-binding protein 1
ZC3H14	Zinc finger CCCH domain-containing protein 14
ZCCHC8	Zinc finger CCHC domain-containing protein 8
ZFC3H1	Zinc finger C3H1 domain-containing protein
ZMW	Zero Mode Waveguide

1 Introduction

Gene regulation is the central biological process translating the information encoded in an organism's DNA into proteins which exert most cellular functions. The importance of regulating how, when and where protein is produced in a cell becomes obvious in all domains of biology: cells grow, divide, and react to their environment, which requires for a regulatory layer to produce the right cellular components at the right time. Increasing complexity evolved with multicellular organisms, where cells and cell types share the same DNA sequence but develop into highly diversified building blocks of tissues and organs. This transformation from a single zygote into the complex appearance of higher mammals has fascinated the scientific community since its beginnings. From the foundational work of Max Delbrück, Nikolaj Timofejew-Ressowski and Karl Günther Zimmer the 1930s [1], which framed the concept of molecular genetics, we have drastically expanded our understanding of the molecular principles governing gene expression and the complicated regulatory networks enabling the sheer endless diversity of life.

Francis Crick coined the "central dogma of molecular biology", illustrating the flow of genetic information from DNA which is transcribed into RNA and in turn translated into proteins [2]. Regulation of these steps is universal and highly specialized for different species. Messenger RNA (mRNA) plays a pivotal role in regulating gene expression and numerous cellular mechanisms operate on the mRNA level. Besides its role as carrier of information, different classes of non-coding RNAs exert functions besides serving as templates for protein production: ribosomal RNAs (rRNA) are the major structural and enzymatic components of ribosomes, the macromolecular machineries responsible for protein production. Transfer RNAs (tRNA) connect amino acids to mRNA sequence during translation through the usage triplet codons ('genetic code'). Small nuclear RNAs (snRNA) are essential for gene splicing, micro RNAs (miRNA) are important mediators of post-transcriptional gene regulation and long non-coding RNAs have finally important regulatory roles for instance in inactivating one female X-chromosome [3].

Eukaryotic mRNAs are typically modified with a poly(A) tail at their 3'-end after completion of transcription. Poly(A) tails are important dynamic regulators of protein output as well as RNA stability and impact mRNA fate from biogenesis to degradation. Poly(A) tail function has been studied since the late 1960s, yet only recently the systems biology community developed the first methods to study poly(A) tails for each gene and individual molecules in high throughput. These approaches enabled a holistic perspective on poly(A) tail and their impact

on gene regulation, although previous methods had different technical limitations. The first part of this work introduces FLAM-Seq, a novel technological approach for sequencing full-length mRNAs, including their poly(A) tails in high-throughput. FLAM-Seq was used to investigate poly(A) tail length profiles of different biological model systems which revealed important regulatory aspects between poly(A) tail length and for instance towards 3'-UTR sequence and alternative polyadenylation. The second part investigates poly(A) tail dynamics in the nucleus, uncovering an immediate poly(A) tail shortening ('deadenylation') step right after mRNA synthesis. Deadenylation is an established mechanism and precedes mRNA decay, yet it has mostly been studied in the cytoplasm, illustrating the need for a deeper understanding of nuclear RNA processing for control over gene expression.

The first part of the introduction describes the key steps through which an mRNA is processed throughout its lifetime along with the connections between different regulatory layers. The second part introduces the biology of poly(A) tails covering evolutionary aspects of polyadenylation, poly(A) tail biogenesis in the nucleus and nuclear RNA processing, as well as the impact of poly(A) tails on translation and relevance for mRNA decay. The last part of the introduction describes the technological state-of-the-art for experimentally measuring poly(A) tails with a focus on high-throughput sequencing technology, which is relevant for understanding the foundation for development of the FLAM-Seq method.

1.1 The mRNA life cycle

1.1.1 mRNA biogenesis, maturation and splicing

Eukaryotic RNA is produced in the nucleus by transcription of DNA by RNA polymerases (*Figure* 1). Transcription initiation requires concerted action of different nuclear factors and processes which enables production of mRNA: Epigenetic modifications of histones and DNA primary sequence define whether DNA is accessible for ('pioneering') transcription factors and the transcription machinery. Acetylation of lysine residues in histones enables for instance binding of bromodomain proteins which opens chromatin conformation [4], while methylation of DNA CpG dinucleotides is typically repressing transcription [5]. Chromatin and 3D genome architecture impact interactions of gene regulatory elements, such as enhancers and promoter sequences, thereby fine-tuning gene expression for instance in different cell types [6]. Finally, the expression of transcription factors is indispensable for recruiting the transcription machinery to specific genes and establishing the regulatory code which specifies cell type and context dependent gene expression [7]. The sheer number of around 1400 identified transcription factors further underpins the relevance of tightly regulating mRNA production [7].Eukaryotic transcription is catalyzed by a set of DNA-dependent RNA polymerases: RNA polymerase I (RNAPI) mainly transcribes rRNA, RNA Polymerase III (RNAPIII) transcribes tRNA and 5S rRNA, and RNA polymerase II (RNAPII) transcribes mRNA and different classes of non-coding RNAs as for instance lncRNAs, miRNAs and snRNAs [8].

RNAPII transcription initiation involves binding of promoter- and context-dependent transcription factors along with a conserved set of general transcription factors (TFII proteins) which are required for RNAPII initiation: As part of the core transcription-initiation complex, the TBP protein is first recruited to the core promoter sequences by binding the -30 TATA box or as a component of the TFIID complex which interacts with promoters without TATA boxes. This leads to recruitment of TFIIA and TFIIB, which stabilize the interactions between the assembled TFII complex and the genomic DNA. Subsequently, the TFIIF – RNAPII complex associates with the promoter and the C-terminal domain (CTD) of the RPB1 subunit of RNAPII is phosphorylated. The RNAPII CTD is a low-complexity protein domain and composed of 52 heptad repeats in human [9] which are each targeted by different post-translational modifications such as phosphorylation. The resulting CTD phosphorylation code is indicative of RNAPII transcription status and important for recruiting RNA processing factors, for instance required for co-transcriptional capping. CTD phosphorylation at serine 2 (Ser2) by TFIIH and the pTEFb complex enable processive transcription elongation. Transcription of complete human genes

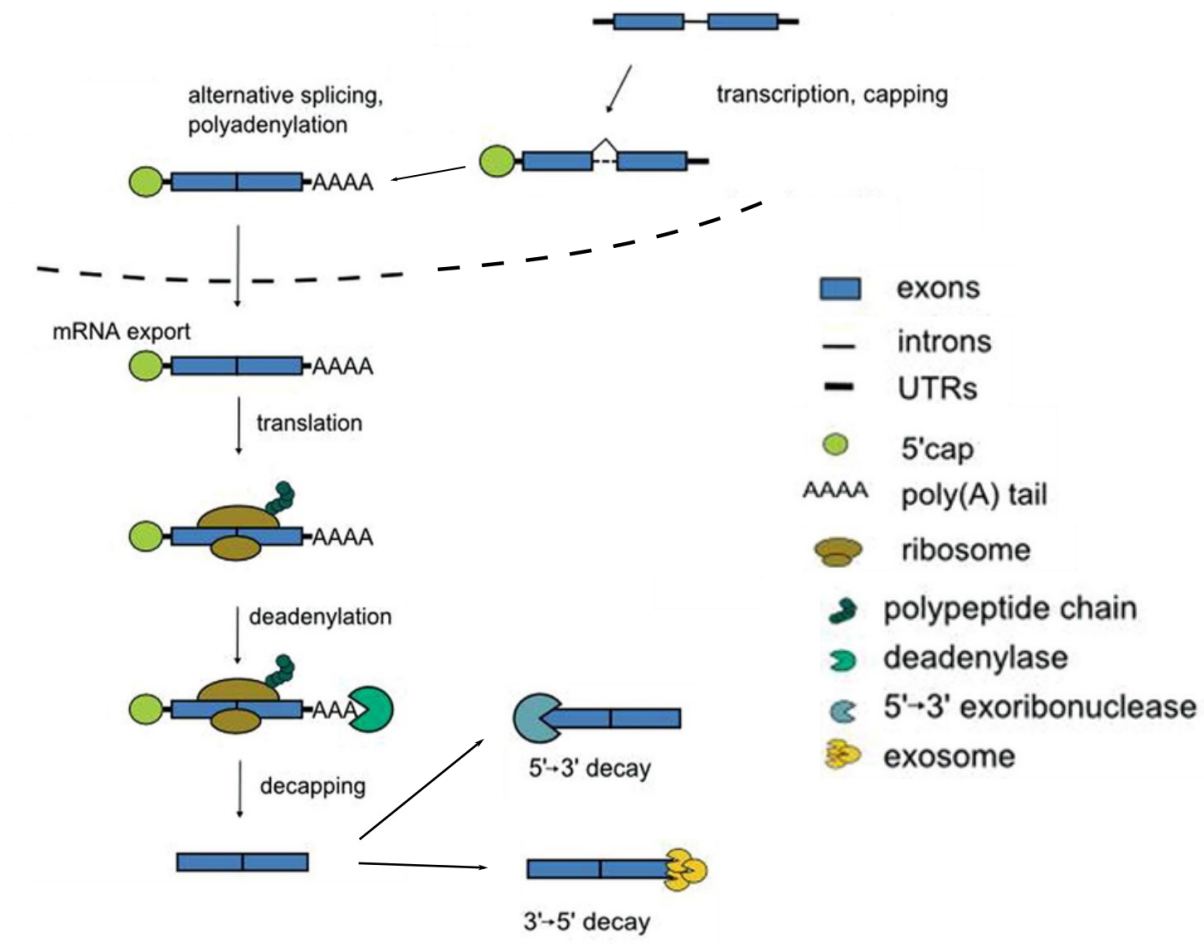

Figure 1 **Overview of eukaryotic RNA processing** (modified from Nolte et al. 2015*)*

typically operates in the order of minutes with average elongation rates of 1.5–4.3 kbp per minute [10–12]. A notable example of regulation through transcription rates is the DMD (‘dystrophin’) gene with a length of 2.3 Mbp which requires 16 hours for its transcription suggesting a strong delay between RNA and protein production [13].

Transcription termination requires pausing and release of RNAPII from the DNA template sequence: Upon transcription of conserved polyadenylation signals (PAS) at the 3’-end of a transcription unit, PAS sequences are bound by the cleavage and polyadenylation complex (CPA). RNAPII CTD interacts with the CPA, which triggers cleavage and definition of 3’-end of the nascent RNA transcript (described in detail below) and additionally a slowdown of RNAPII transcriptional activity [14]. After cleavage, the nuclear 5’-3’-exonuclease Xrn2 degrades the downstream transcript and RNAPII is released from DNA upon contacting Xrn2. Other factors reducing RNAPII transcription rates and facilitating dissociation in this context include chromatin structure and formation of R-loops through RNA:DNA hybridization of the nascent pre-mRNA [14].

Transcription can be inhibited by different small molecules with different mechanisms of action [15]: Actinomycin D intercalates GC-rich DNA regions and prevents progression of transcription for all RNA polymerases with little specificity for RNAPII-mediated mRNA transcription. DRB on the other hand selectively inhibits CDK9 leading to stalled transcription complexes but has off-target specificity for other kinases. α-amanitin is a direct inhibitor for targeting the RNAPII active site and traps RNAPII in a non-processive conformation. Transcription inhibitors are important molecular biology tools for instance for measuring transcription rates [11,12] and are also used as chemotherapeutics for treating various cancers [16].

Transcribed pre-mRNA is co-transcriptionally capped at its 5'-end, typically with a 7-methylguanine cap, which protects from exonucleolytic degradation and prevents induction of innate immune responses (recognition as 'self') which is triggered for instance by uncapped viral RNAs [17]. Upon transcription initiation, the guanylytransferase RNGTT is recruited by phosphorylated RNAPII CTD which catalyzes guanylation of the pre-mRNA 5'-end. The methyltransferase RNMT then methylates the guanine N-7 position and the cap methyltransferases CMTR1/CMRT2 methylate the 2'-OH positions of the first two bases. After synthesis, the cap is bound by the heterodimeric Cap Binding Complex (CBC), which promotes nuclear pre-mRNA processing including splicing and export [18].

Most genes in higher eukaryotes are organized as split genes containing exons (expressed regions) which comprise mature messenger RNA and introns which are removed from pre-mRNA during the splicing process. Human genes have an average length of 28 kbp and have on average 9 exons with a mean length of 170 bp. Introns are significantly longer, on average 5000 bp per human intron [19,20]. Gene architecture is highly species dependent: Of the ca. 6000 genes in the baker yeast *Saccharomyces cerevisiae*, only ca. 4% contain introns, compared to around 90% of the genes in the human genome [21]. Prokaryotes generally lack intronic sequences.

Introns are spliced out from pre-mRNA by the splicing process, which is catalyzed by the spliceosome, which is a large and dynamic RNA-protein (RNP) complex [22]. Pre-mRNA contains conserved splice site sequence elements which guide definition of intron boundaries: Most intronic sequences start with a highly conserved GU dinucleotide at the 5'-splice site and end with an AG dinucleotide at the 3'-splice site with both motifs embedded into more complex sequence contexts. Around 25 nt upstream of the 3'-splice site, a conserved branch point adenosine nucleotide and surrounding motif are followed by a pyrimidine rich element ('Polypyrimidine Tract') [23]. 5'-splice sites are first recognized by U1 snRNP, an RNP composed

of U1 snRNA and Sm proteins, which associate to pre-mRNA through RNA-RNA interactions. The 3'-splice site is defined by binding of U2AF65, U2AF35 and SF1 proteins. U2 snRNP interacts with the branch point sequence and facilitates recruitment of the U5-U4/U6 tri-snRNP complex, which completes assembly of the major spliceosome. Release of U1 and U4 snRNP triggers conformational changes towards the activated complex. The activated complex then catalyzes the first step of the splicing reaction, which is the nucleophilic attack of the branch point adenosine 2'-OH group at the 5'-splice site, leading to formation of a lariat intermediate. In the second step of the splicing reaction, the 3'-OH of the upstream exon attacks the 3'-splice site which leads to transesterification of the two adjacent exons. The lariat intron is then released, and the spliceosome is disassembled after completion of the splicing reactions. Mammalian spliceosomes further deposit exon-junction complexes (EJCs) 24 nt upstream of the splice junctions. EJCs are important elements of RNA quality control and involved in nonsense-mediated decay (NMD) and generally act as translational activators [24].

The spliceosome can be targeted by several small molecules which modulate different steps of the spliceosome cycle [25]. A common target for inhibition of splicing is SF3B1, a protein component of the SF3B complex with stabilizes U2 snRNP binding to the branch point sequence in pre-mRNA. SF3B1 has been found mutated in several cancers [26] and as such is of interest as a pharmacological target. Compounds such as spliceostatin A [27] or pladienolide B [28] have been identified as potent splicing inhibitors with anti-proliferative properties, which are commonly used in molecular biology to inhibit splicing.

How splice sites are recognized largely depends on intron length: For short *S. cerevisiae* introns (mean length of 230 nt [29]), splicing occurs in 'intron definition' mode where spliceosome components assemble and interact over 5'- and 3'-splice sites of the same intron, whereas for long human introns splices sites are defined over a given exon ('exon definition'), which increases the fidelity of splice site recognition in long intronic sequences which may contain a number of random or highly degenerate cryptic splice sites [30].

Through the modular gene architecture, exons can be selectively included in an mRNA by alternative splicing, which is common in eukaryotes [31]. For each exon, a regulatory code of *cis*- and *trans*-acting factors determines whether an exon is recognized by the spliceosome and spliced or skipped and excluded. RNA binding proteins (RBPs), such as SR proteins, regulate this process, as they typically enhance splice site recognition by binding to sequence elements termed exonic splice site enhancers (ESEs). Conversely, binding of RBPs to exonic or intronic splice site silencers promotes exon skipping, which is for instance mediated by hnRNP proteins.

Alternative splicing is not limited to inclusion or exclusion of exons, also 5'- and 3'-splice site choices for individual exons are modulated by similar mechanisms [32,33]. Transcription kinetics play an important role in regulating alternative splicing: slower transcription elongation leads to increased exon inclusion in budding yeast, which is explained by a longer time window for exon definition through the splicing machinery ('kinetic coupling model') [34]. Albeit mechanistically appealing, this model could not be generalized to human systems. Altering transcription elongation rates in stem cells and HEK cell lines was indeed found to have profound impact on alternative splicing yet the effects were not proportional to the changes in elongation rates [35,36].

The splicing process has initially been hypothesized to occur post-transcriptionally, after complete transcription of a pre-mRNA [37]. Subsequent investigation of splicing dynamics with respect to transcription elongation has yet led to the current view that splicing occurs for the majority of genes co-transcriptionally, such that introns are spliced shortly after being transcribed by RNAPII [38–41]. Recent studies applying full-length RNA sequencing yet claim that complete splicing occurs post-transcriptionally for up to 40% of human genes [42,43], with many exons not being spliced in their linear order of transcription.

Kinetics of splicing reactions have been studied using inducible splicing reporters, high-resolution microscopy techniques or next generation sequencing, yielding estimates which appear highly dependent on model system and experimental setup [44]. Splicing of human introns occurs on average within 5-15 min after synthesis [45,46]. Splicing in the budding yeast *S. cerevisiae* is much faster, being completed within less than 90 seconds after transcription [47,48]. Global splicing rates yet need to be understood in context of individual gene architectures: Slower processing has been attributed to shorter, highly expressed genes [46] and splicing efficiency is decreased towards the 3'-end of a transcript [38]. Terminal exons have distinct kinetic properties with increased RNAPII pausing, which is proposed to facilitate completion of splicing in budding yeast [49]. Terminal intron splicing has also been linked to mRNA 3'-end formation, since mutations in polyadenylation sites can suppress terminal intron splicing, which highlights the complex mechanistic interplay of mRNA 3'-end maturation [50,51]. Transcripts with retained introns have long been regarded as targets for rapid degradation, although regulated intron retention for nuclear transcripts was shown to be widespread during neuronal development [52] or induction of cellular signaling pathways [53,54].

Some RNAs do not follow the linear order of exon splicing: circular RNAs (circRNAs) are a class of non-coding RNAs which are covalently linked at their 5'- and 3'-ends and have

regulatory functions, for instance on miRNA targeting [55,56]. circRNAs are produced in a back-splicing reaction in which the 5'-splice site is joined with an upstream 3'-splice site in a process that competes with canonical RNA splicing [57].

In the roundworm *Caenorhabditis elegans,* many pre-mRNA 5'-exons are trans-spliced to splice-leader RNAs (SL-RNAs) which provide an alternative route for capping of mRNAs [58]. Both back- and trans-splicing illustrate the plasticity of the splicing machinery in co-transcriptionally regulating gene expression.

Factors involved in transcription and splicing are at least in part localized to subcellular structures termed nuclear speckles which are membrane-less organelles [59]. Despite a general lack of understanding why nuclear speckle form, there is evidence for post-transcriptional splicing occurring within speckles [60].

1.1.2 3'-UTRs, (alternative) polyadenylation and RNA export

Completion of mRNA transcription requires cleavage and polyadenylation of the nascent transcript producing the mature, polyadenylated mRNA 3'-end. The cleavage reaction is catalyzed ca. 20 nt downstream of a highly conserved AAUAAA polyadenylation signal (PAS) where the CPSF complex assembles. Different PAS variants are found for some genes while other completely lack polyadenylation signals [61], hinting at different converging pathways for CPSF assembly. The PAS is bound by the CPSF components CPSF30 and Wdr33 [62], assisted by several RNA binding factors such as hFip1. This provides scaffolding function for CPSF73, which catalyzes the cleavage reaction adjacent to a CA dinucleotide [63]. Poly(A) Polymerase (PAP) then appends the poly(A) tail to the cleaved 3'-end [64]. Besides cleavage and polyadenylation, the third function of the CPSF complex is its phosphatase function towards RNAPII CTD which is thought to facilitate transition from transcription elongation to termination [65].

The PAS is embedded in additional sequence elements which assist in directing the specific polyadenylation site for a transcript. U/GU rich sequence elements upstream and downstream of the PAS and cleavage site are bound by RNA-binding complexes CstF, CF Im, and CF IIm which interact with CPSF to increase specificity in PAS selection.

The architecture of the *S. cerevisiae* 3'-end processing machinery is similar to human, yet with a less conserved PAS motif and surrounding sequence elements and differences in auxiliary components of the cleavage and polyadenylation complex [66].

mRNA open-reading frames (ORFs), which encode proteins, are embedded in 5'- and 3'-untranslated regions (UTRs), which have regulatory functions. 5'-UTRs contain sequence elements relevant for mRNA translation: examples are TOP tracts, which govern protein production by the mTOR pathway in response to nutrients [67] or IRES elements which drive cap-independent translation for instance for certain viral RNAs or under stress conditions [68].

3'-UTR are on the other hand major regulators of post-transcriptional gene regulation and influence RNA stability, localization, protein expression and binding of miRNAs and RNA binding proteins [69]. 3'-UTRs are conserved between species, which early on pointed at relevant regulatory roles [70]. Human 3'-UTRs have an average length of 1000 nt, which is significantly longer than those in mouse (850 nt), fruit fly (270 nt), and baker yeast (150 nt) [71,72]. First hints of how 3'-UTRs impact gene expression came from experimental evidence that deletion of AU-rich sequence elements (AREs) from the c-Fos 3'-UTR activates oncogenic pathways by inducing constitutive c-Fos expression [73]. This observation linked 3'-UTR embedded sequence motifs to RNA stability. AREs are established post-transcriptional regulatory motifs which require *trans*-acting factors such as RBPs to cause functional effects. Binding of TTP to AREs is linked to recruitment of RNA decay factors such as Xrn1, the exosomes complex [74–77] and a crystal structure of the CCR4-NOT component NOT1 binding to TTP has been determined[78]. Binding of the RBP HuR on the other hand stabilizes mRNAs by outcompeting destabilizing factors as TTP [79,80]. The regulatory mode of a given 3'-UTR sequence hence needs to be interpreted in context of expressed *trans*-acting factors.

The discovery of RNA-interference and miRNA-mediated post-transcriptional regulation of gene expression further highlights the role of 3'-UTRs as platforms for controlling transcript fate: miRNAs in a complex with AGO proteins (RISC complex) bind to conserved target sites, mostly found within 3'-UTRs, which causes deadenylation and decay of target genes [81]. Opposite functionality has been observed for some miRNA binding sites which cause destruction of the miRNA instead [82].

RBPs and miRNAs do not act independently on individual motifs and the outcome of a given 3'-UTR can involve cooperative action of multiple regulatory sites [83] as for instance shown for two highly conserved miRNA binding sites in the *C. elegans lin-41* 3'-UTR [84].

Many biological systems, such as oocytes or neurons, are highly polarized and active transport of cellular components is required their homeostasis. This also includes transport of mRNAs, which are in many cases locally translated, for instance in neurites or neuropil [85,86]. 3'-UTRs are thereby instrumental in determining RNA localization: *bicoid* mRNA is for instance

actively transported to anterior poles in *Drosophila melanogaster* egg cells, which involves binding of transport components to a stem-loop secondary structure within the *bicoid* 3'-UTR [87]. Other motifs such as 'zipcode' 3'-UTR sequences elements are bound by ZBP1 and target for instance beta-actin mRNA to the cellular periphery [88]. In most cases, the *cis*-regulatory sequences involved in transport are yet more diverse and involve multiple binding sites or structural elements [89] to determine mRNA localization.

Usage of different polyadenylation sites for a gene can result in isoforms with different 3'-UTRs. This process is termed alternative polyadenylation (APA), which may lead to longer, extended 3'-UTRs by cleavage at a more distal PAS or shorter 3'-UTRs through proximal PAS usage. Extended 3'-UTR sequences have been shown to be enriched in miRNA binding sites and destabilizing *cis*-regulatory elements such as AREs [90]. Longer 3'-UTRs are hence regarded as less stable, although this is likely a simplification as genome-wide studies on isoform specific RNA stability observed only a minor trend supporting this notion [91]. The fraction of genes producing alternative 3'-UTR isoforms is larger for higher eukaryotes and highly tissue-dependent, with up to 70% of yeast and mammalian genes producing APA isoforms [92–95]. Polyadenylation sites also occur within introns and intronic cleavage occurs for instance in tumors which leads to inactivation of tumor suppressor genes [96]. Another example is a feedback mechanism for controlling expression of the CPA component Cstf-77, which triggers intronic cleavage of its own mRNA [97]. Global differences in PAS choice have been observed for a multitude of biological processes: 3'-UTRs of proliferative cells are typically shorter [90], which is also observed for oncogenes in cancer cell lines [98]. 3'-UTRs change in length throughout development [99], and in particular neuronal systems are characterized by expression of genes with extensively long 3'-UTRs [100,101].

Alternative polyadenylation is important in specifying post-transcriptional processing of different RNA isoforms: As described above, APA can modulate RNA stability, localization, RBP recruitment, miRNA binding and other features by inclusion or exclusion of *cis*-regulatory sequence elements. Selection of individual polyadenylation sites for cleavage can be attributed to *cis*-regulatory motifs adjacent to the PAS, for instance U-rich / GU-rich elements which bind auxiliary CPA components such as the CstF complex. Sequence elements determine the affinity for recruiting those *trans*-acting factors and the availability of CPA components in return also affects PAS selection. One example here is Fip1, a component of the CPSF complex relevant for recruiting Poly(A) Polymerase: its depletion leads to differentiation of stem cells along with deregulation of APA patterns [102].

Similar to the proposed 'kinetic coupling' model for control over alternative splicing, RNAPII transcription rates have been linked to PAS selection. Slow RNAPII elongation favored proximal PAS site choice, probably through the longer time available for binding for CPA assembly [103]. Overall, the mechanisms controlling APA are conceptually similar to regulation of alternative splicing, although it remains unclear in how far differential stability of 3'-UTR isoforms or active regulation of PAS site choice contribute to steady state isoform expression levels.

Several examples have been identified which highlight the functional importance of APA: Immunoglobin M heavy chain protein isoforms are modulated by PAS site choice upon B cell differentiation. Inclusion of a terminal exon, which encodes a domain for membrane-anchoring, is regulated by CstF-64, which binds downstream of the cleavage site [104]. APA further impacts proper cellular trafficking of proteins, independent of mRNA localization: CD47 protein is shuttled to the membrane if translated from a CD47 mRNA isoform with a long 3'-UTR. In this case mRNA binds the RBP SET which associates the mRNA and associated ribosomes to the ER transport machinery [105].

3'-end processing is mechanistically coupled to transcription and splicing: polyadenylation site choice can already be influenced during transcription initiation at promotor sites by RBPs which associate with paused RNAPII: One example is ELAV, which is an RBP expressed in the central nervous system and enriched at promoter sites of genes undergoing strong 3'-UTR extensions during development. ELAV thereby suppresses proximal cleavage of the nascent mRNA [106,107].

Mutations in the PAS further inhibit splicing of terminal introns *in vitro* [50], and Poly(A) Polymerase-U2AF65 interactions were shown to promote splicing efficiency [51]. Vice versa, CPA factors such as CPSF2 have been identified in exon binding and modulation of alternative splicing [108]. U1 snRNP further suppresses PAS recognition [109] which is important for prevention of premature transcript cleavage and a mechanism to define the directionality of transcription at bidirectional promoters [110].

Capping, splicing and polyadenylation leaves mature, nuclear transcripts which are exported to the cytoplasm through the nuclear pore complex (NPC). Preparation of export begins co-transcriptionally by loading of RNA export adapters to nascent transcripts [111]. The formation of a compacted nuclear mRNP (mRNA-protein) complex is an important step in nuclear RNA processing. Export adapters comprise SR proteins, which simultaneously affect splice site choice, and the TREX complex. SR proteins are dephosphorylated upon completion of splicing

which then enables association and binding of the export receptor proteins TAP and p15 [112]. TAP and p15 interact with FG repeat proteins, which are components of the nucleoplasmic site of the NPC. Channeling of mRNA through the nuclear pore is facilitated by the RNA helicases DBP5 [113]. Upon reaching the cytoplasmic compartment, export factors dissociate and are reimported into the nucleus.

1.1.3 Translation and cytoplasmic RNA decay

Genetic information stored in mRNA sequences is translated into proteins by the ribosome. Translation competent mRNA is typically in a 'closed-loop conformation', in which 5'- and 3'-ends are in direct proximity, which is suggested to increases ribosome re-initiation after completing one round of translation [114]. The 5'-cap is bound by the initiation factor eIF4E, while the 3'-poly(A) tail is bound by cytoplasmic poly(A) binding proteins (PABPC1). eIF4G bridges the interaction between eIF4E and PABPC1, establishing contact between 5'- and 3'-mRNA ends. Recent data investigating mRNP conformation *in vivo* yet suggest that the close-loop conformation could be of transient nature [115].

Translation begins with formation of the 48S pre-initiation complex which consists of translation initiation factors, the 40S small ribosomal subunit and Methionyl-tRNA, which recognizes the AUG start cordon. The pre-initiation complex scans the 5'-UTR starting at the 5'-cap (cap dependent translation) to identify the start codon, which is typically embedded in the Kozak consensus ribosome binding site. After definition of the start codon, the pre-initiation complex is remodeled and the 60S ribosomal subunit complexes assembly of an elongation competent ribosome. The mRNA is translated into a growing amino acid chain through amino-acyl-loaded tRNAs which link codons to amino acids ('genetic code'). The first 'pioneering' round of translation is important as it entails replacement of the Cap Binding Complex (CBC) by eIF4E initiation factor, displacement of exon-junction complexes (EJC), and exchange of the nuclear poly(A) binding protein (PABPN1) by the cytoplasmic PABPC1 [116]. mRNA quality control mechanisms as nonsense-mediated decay (NMD) act during the first round and prevent translation of possibly toxic protein sequences resulting from aberrant or unspliced mRNAs. NMD detects premature stop-codons, which are likely to occur in introns of unspliced mRNA, through recognition of prematurely stalled ribosomes. NMD in mammals makes use of downstream deposited exon-junction complexes (EJCs) to decide whether a stop-codon is premature [117].

Steady state RNA expression levels are determined both by RNA production (transcription) and degradation rates. Regulation of RNA decay is an active process, and indispensable in gene expression control generating a dynamic range of RNA abundance ranging from one to several thousand expressed mRNAs per gene [118]. Mammalian mRNA half-lives thereby range from 20 minutes to more than 24 hours, with a median of around 4 hours [53,119,120]. This is significantly longer than for yeast mRNAs with an average half live of 23 minutes [121]. Proto-oncogenes or

cytokines are often instable, which enables tight control over their expression in response to external stimuli [122].

RNA decay is regulated by *cis*-regulatory sequences, such as AU-rich elements (AREs), which are found in 20% of all human transcripts [123]. Regulatory motifs can have context dependent stabilizing or destabilizing effects, depending on availability of RNA binding proteins as TTP or HuR. Motifs regulating RNA stability are not limited to 3'-UTRs since destabilizing elements have also been identified within coding regions, for instance for the c-myc transcript [124]. Changes in RNA decay rates are actively regulated by extracellular stimuli: activation of the MAPK signaling pathway triggers phosphorylation of TTP, which prevents TTP from binding to AREs and which stabilizes inflammatory genes as such as TNF-α [125,126]. The JNK responsive element (JRE) is present in 5'-UTRs of IL-2 mRNA and stabilizes mRNA upon activation. Other pathways involved in modulation of mRNA decay have been described, for instance TCR/CD28 activation, PKC or PI3K pathways [127]. Reporter screens for 3'-UTR sequence elements which impact RNA stability in early *Xenopus* development further uncovered U-rich stabilizing and G-rich destabilizing motifs [128,129]. The ENE element has been described as a U-rich motif found in a viral lncRNA which sequesters the poly(A) tail into a triple-helix structure that prevents its deadenylation and decay [130].

mRNA decay requires several processing steps and different enzymes: first the poly(A) tail is removed by deadenylase enzyme complexes as CCR4-NOT, PAN2-PAN3 or PARN (the detailed mechanisms of deadenylation-dependent decay are discussed in the next chapter). The mRNA is subsequently decapped by the DCP1/DCP2 complex [131], which exposes the 5'-end for degradation by the 5'-3'-exonuclease Xrn1. Decapping is stimulated by binding of the Lsm1-7 complex to mRNA tails with short oligo(A) overhangs [132,133]. Xrn1 is a monomeric enzyme and involved in the decay of cytoplasmic mRNA, non-coding RNAs and NMD-targets [134]. mRNA decay from the 3'-end is catalyzed by the exosome which is a multi-subunit complex with distinct molecular compositions depending on its subcellular localization. While the cytoplasmic exosome mediates the bulk of RNA turnover, the nuclear exosome is important for RNA quality control and rRNA maturation [135]. 5'- and 3'-end mediated decay mechanisms can operate in parallel on the same RNA [136]. RNA decay can also in some cases occur independently of deadenylation (deadenylation-independent decay): *S. Cerevisiae* ribosomal protein RPS28 for instance binds to a hairpin structure within its own 3'-UTR which recruits decapping factors triggering 5'-decay [137].

RNA decay factors are also partially compartmentalized into P-bodies (processing bodies, also Dcp- or GW bodies), which contain components of the decapping complexes, deadenylases and exonucleases, as well as structural RBP components which facilitate formation of RNA-protein networks contributing to phase separation of P bodies [138].

1.1.4 lncRNA processing

Long non-coding RNAs (lncRNAs) are a class of RNA molecules longer than 200 nt without encoding proteins. lncRNAs have important functions in regulating gene expression: The lncRNA XIST is a key regulator of X chromosome inactivation to ensure proper dosage compensation in female cells [3]. lncRNA HOTTIP coordinates expression of HOX genes, which are important regulators specifying the position of body parts during development. More than 100,000 human lncRNA genes have been annotated [139], which by far exceeds the number of around 20,000 annotated protein-coding genes [140].

lncRNAs are RNAPII transcripts but expressed at much lower levels than mRNAs from protein-coding genes. lncRNA expression also has a higher tissue specificity [141]. lncRNAs are yet less efficiently spliced [38] and have higher degradation rates [142]. Many lncRNAs are actively retained in the nucleus which requires for mechanisms that prevent their export [143]. Two regions have for example been identified in MALAT1 lncRNA which are bound by nuclear protein factors such as RNPS1 that localize MALAT1 to nuclear speckles [144]. Another proposed mechanism relates to U1 snRNP, which binds U1 motifs in a number of lncRNAs and tethers them to chromatin [145], where some lncRNAs impact gene expression and epigenetic regulation.

1.1.5 Systems biology perspectives on exploring mRNA biology

In summary this chapter described the foundations of eukaryotic gene expression which is relevant for understanding the role of polyadenylation and poly(A) tail length on RNA metabolism. mRNAs (and lncRNAs) are transcribed, capped, spliced and polyadenylated in the nucleus. Mature mRNPs are then exported to the cytoplasm where mRNAs are translated into proteins, stored, and eventually degraded. Most importantly, individual processing events are interconnected, polyadenylation site choice may for instance be influenced by splicing and vice versa. This enabled development of complex regulatory layers and adaptations to changing environments, but on the other hand complicates interpretations when studying individual processes. Many regulatory principles, such as negative feedback regulation or the interplay of *cis*-regulatory sequence motifs and *trans*-regulation by expressed proteins are recurring at different levels of RNA metabolism.

Systems biology attempts a holistic perspective on understanding mRNA processing, mostly by measuring steady-state RNA abundance after perturbation of different pathway components. As described above, many processing steps are transient and require accurate time-resolved measurements for instance in understanding splicing kinetics. Other properties as intrinsically hard to measure on a genome-wide level, as for instance nuclear decay. One central problem so far had been the lack of methods to sequence complete mRNA molecules for each gene, which is important for instance in understanding splicing and poly(A) tail biology. This work contributes to this problem by developing a method which enables full-length mRNA sequencing to elucidate the expression patterns of individual mRNA isoforms and hence contributes to a holistic, systems-level understanding of RNA biology. Of relevance is here the investigation of the poly(A) tails, which are central players of gene regulation and introduced in the next section.

1.2 Poly(A) tails controlling gene expression

Essential properties of mRNA poly(A) tails were discovered in the 1960's and 1970's: it became apparent that stretches of adenosines were appended to a large fraction of the cellular RNA pool [146]. Those are added enzymatically and independent of a genomic DNA template [147]. Studies in viruses first hinted at a protective role of poly(A) tails against exonucleolytic enzymes and RNA decay [148]. Dynamic regulation of poly(A) tail length with the age of the mRNA has also been discovered early on through time-resolved labeling experiments [149]. Poly(A) tails are by now understood as essential elements in regulating RNA maturation, processing and translation as well as stability and decay.

1.2.1 Evolutionary perspective on poly(A) tail diversification

Poly(A) tails are considered a hallmark of eukaryotic mRNAs, although polyadenylation is involved in bacterial mRNA decay as well, which suggests evolutionary conserved mechanisms for regulating RNA decay. *E. Coli* poly(A) polymerase (PAP) has been discovered in the 1960, but bacterial polyadenylation generally does not lead to steady-state poly(A) tails of noticeable length as for eukaryotes. Prokaryotic poly(A) tails are also not synthesized directly on nascent transcripts. Adenylation by PAP and deadenylation by the *E. Coli* exonucleases RNaseII and PNPase are competitive processes and impact transcript stability since short bacterial poly(A) tails act as a platform for recruiting decay enzymes, which catalyzed degradation of the transcript body [150].

mRNA decay pathways differ between bacteria and eukaryotes, but certain elements of bacterial poly(A) turnover can be found in cellular organelles: mRNA encoded by the mitochondrial genome are transcribed as polycistronic transcripts and processed into individual mRNAs by endonucleases. This results for many mitochondrial mRNAs in incomplete stop codons, which are completed by addition of a poly(A) tail. Mitochondrial poly(A) tails are synthesized by mitochondrial poly(A) polymerase (mtPAP, PAPD1), resulting in a steady-state poly(A) tail length of around 50 nt [151]. Poly(A) tails of mitochondrial mRNAs are removed by the deadenylase PDE12 [152].

Yeast mtRNAs are not polyadenylated and their 3'-ends are defined by distinct motifs, such as an AU-rich dodecamer sequence [153] in baker yeast (*S. cerevisiae*) or a C-rich terminal sequence in fission yeast (*S. pombe*) [154]. The motifs are bound by RNA binding proteins that may confer stability and protection from decay [155]. Transient polyadenylation, as observed for bacteria, has also been reported for chloroplast mRNAs and plant mitochondrial RNAs [156].

Differences in global steady-state poly(A) tail length were also observed between different species and their developmental stages or tissues: yeast poly(A) tails have median poly(A) tail length per gene of ca. 30 nt, the plant *Arabidopsis thaliana* has an average tail length of 50 nt, whereas human and mouse cancer and fibroblast cell lines have average tail lengths ranging from 80 to 110 nt [157]. *C. elegans* has a median poly(A) length of 82 nt [158] and the slime mold *Dictyostelium discoideum* of 65 nt [159]. Developmental stage specific poly(A) tail length profiles have been observed during zebrafish, frog and fruit fly development [157,160].

In summary, poly(A) tails evolved from transient decay intermediates in bacteria to stabilizing elements of mammalian mRNAs. Poly(A) tail length distributions are thereby specific for different species, throughout development and different mechanisms operate for instance in cellular organelles, which are evolutionary related to prokaryotes.

1.2.2 Poly(A) tail synthesis and function in the nucleus

Eukaryotic poly(A) tails are synthesized upon 3'-end cleavage of the nascent transcript by nuclear poly(A) polymerases (PAP). The polyadenylation process is highly conserved between human and yeast. Human poly(A) tail synthesis is triggered by tethering of PAP to the cleaved mRNA 3'-end through CPSF160, a scaffold protein component of the cleavage and polyadenylation complex (CPA), and other auxiliary RNA binding CPA components as hFip1[161]. Tethering facilitates distributive synthesis of 12 nt poly(A) tails, which can be bound by nuclear poly(A) binding protein (PABPN1) [162]. PABPN1 then stimulates processive poly(A) tail synthesis until a length of ca. 250 nt is reached (Figure 2). The upper tail length is thereby defined by binding of additional PABPN1 molecules to the nascent, elongating poly(A) tail, which results in a torus-like conformation which becomes sterically instable at around 250 nt of synthesized poly(A) tail [163]. This instable conformation causes dissociation of PAP from CPSF and termination of the processive polyadenylation reaction [164]. Synthesis of poly(A) tails of around 250 nt in length has further been demonstrated by radioactive labeling of newly transcribed RNAs in mammalian cell culture systems [149,165]. It is however unclear if there are gene-specific differences in the length of the synthesized poly(A) tails.

Stimulation of polyadenylation in yeast is less dependent on poly(A) binding proteins: yeast poly(A) polymerase Pap1p is efficiently stimulated by the cleavage and polyadenylation factor CPF [166]. The yeast nuclear poly(A) binding protein Nab2, is required for limiting poly(A)

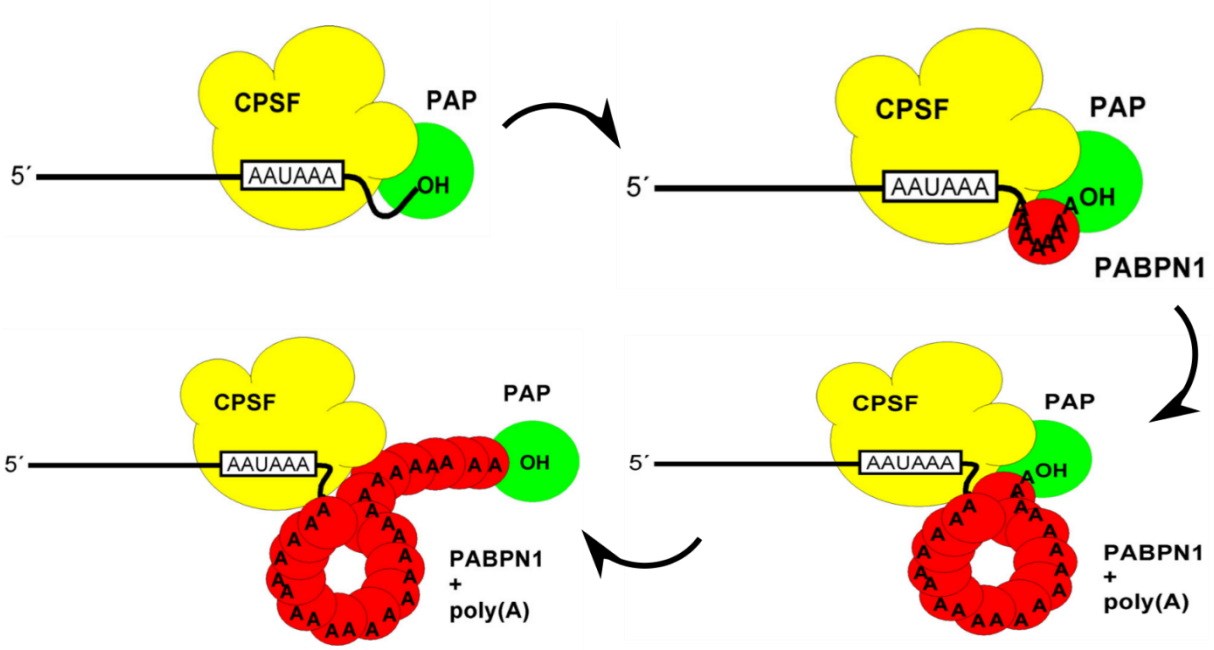

Figure 2 **Model for poly(A) tail length control during synthesis of poly(A) tails** (modified from Kühn et al. 2009)

tail length to 70-80 nucleotides [167,168] and has additional important functions in connecting polyadenylation to RNA export [169]. The metazoan homologue of Nab2, the zinc finger protein ZC3H14, is also involved in polyadenylation for instance in neurons [170]: mutations in ZC3H14 cause reduction in hippocampal poly(A) length and induce intellectual dysfunctions in mice [171].

Three canonical human nuclear poly(A) polymerases (PAPs) have been identified which synthesize poly(A) tails [172]. PAP-α and PAP-γ are ubiquitously expressed and contain a catalytic, a RNA-binding and a C-terminal domain which is specifying interactions for instance with U1-70k or U2AF65 splicing factors [51,109]. PAP-β is specifically expressed in testis [173].

Star-PAP (TUT1) is a non-canonical poly(A) polymerase which has been identified in nuclear poly(A) tail synthesis. In contrast to canonical PAPs, Star-PAP is capable of direct pre-mRNA binding, and forms specific CPA complexes which have different target sequence preferences [172]. Star-PAP is activated by the phosphatidyl-inositol-4,5-biphosphate (PI4,5P2) signaling pathway in response to oxidative stress which in turn alters polyadenylation site usage [174]. PAPs may hence regulate gene expression beyond the synthesis of poly(A) tails, for instance by modulating RNA isoform expression.

The nuclear poly(A) binding protein PABPN1 is essential for regulating efficient poly(A) tail synthesis, correct tail length functions and impacts different RNA maturation steps: knockdown of PABPN1 leads to a significant reduction in the poly(A) tail length of newly synthesized RNAs. Although poly(A) tail synthesis is not completely abolished in absence of PABPN1, poly(A) tail length profiles of newly synthesized mRNAs are considerably shorter [175].

PABPN1 is molecularly distinct from the cytoplasmic poly(A) binding protein PABPC1. It is composed of one RNA binding RRM domain, with a preference for binding stretches of 12 adenosines [176] and an unstructured C-terminal domain, which is likely facilitating multivalent protein-protein interactions. PABPN1 has been shown to suppress PAS site usage by binding A-rich sequences in pre-mRNAs, leading to 3'-UTR extensions [177]. Beyond the role of PABPN1 in regulation of alternative polyadenylation site choice, PABPN1 is suggested to promote splicing efficiency of terminal introns, although the exact mechanism is not clear [178].

Mutations in the PABPN1 gene can lead to alanine extensions in the N-terminal domain. Those mutated PABPN1 proteins can form toxic aggregates which can be a cause of oculopharyngeal muscular dystrophy, a rare genetic muscular disease around the eyelids [179]. PABPN1 mutations also imply defects in pre-mRNA processing such as an increased usage of intronic polyadenylation sites [180]. PABPN1 is further involved in nuclear, polyadenylation-dependent regulation of lncRNA turnover by the nuclear exosome [181] and interacts with the PAXT complex, which targets polyadenylated, non-coding transcripts as SHNG RNAs for decay by the exosome [182].

The *Drosophila melanogaster* PABPN1 ortholog PABP2 is involved in cytoplasmic CCR4-NOT mediated RNA decay [183], hinting at potential cytoplasmic roles of nuclear poly(A) binding proteins. This observation is important since PABPN1 may be exported with mRNA and in part coat poly(A) tails during the first 'pioneering' round of translation, where RNA surveillance mechanisms such as nonsense-mediated decay trigger recruitment of the CCR4-NOT complex to deadenylate aberrant RNAs [184].

Poly(A) tails play an important role for production of telomer RNA (TR), the RNA component of human telomerase, which provides the template for reverse transcription of telomeres at chromosome ends. TR is a non-polyadenylated RNAPII transcript, but the exact 3'-end processing pathway protecting it from decay is unclear [185]. TR precursors have poly(A) tails which are likely bound by PABPN1. PABPN1 then recruits the nuclear deadenylase PARN, which is required for deadenylation and stabilization of TR. This process competes with recognition of short poly(A) tails by the nuclear exosome and decay of the TR transcript, which

illustrates the concept stabilization of polyadenylated transcripts and deadenylation and decay triggered by short poly(A) tails to balance RNA abundance [186,187].

Histone mRNAs, which encode the proteins that package DNA into chromatin, are an important example of non-polyadenylated transcripts with a distinct biogenesis pathway and 3'-end processing [188]. Histone mRNAs contain a conserved stem-loop structure at their 3'-ends, along with a histone downstream sequence element (HDE), downstream of the cleavage site. The stem-loop is bound by stem-loop binding protein (SLBP) and the HDE by U7 snRNP. SLBP and U7 snRNP trigger recruitment of cleavage and polyadenylation complex components, such as CPSF73, which catalyze the cleavage at the histone 3'-ends. The stem-loop bound by SLBP conveys similar functions as a poly(A) tail in mediating 5'- and 3'-end interactions during translation and control over decay. Histone genes lack introns and artificial insertion of an intron leads to partial formation of a polyadenylated histone mRNAs [189], which again highlights the relevance of splicing for stimulation of polyadenylation.

1.2.3 Polyadenylation and nuclear RNA quality control

Pervasive genomic transcription produces numerous non-coding RNA species which are not exported to the cytoplasm and must be degraded in the nucleus. Examples are unstable, polyadenylated RNAPII transcripts such as cryptic, unstable transcripts (CUTs) or promoter upstream transcripts (PROMPTs) [190,191], but also aberrantly spliced pre-mRNAs which are retained from export. Nuclear RNA decay is thereby mostly mediated by the nuclear exosome. In yeast, the nuclear exosome is involved in turnover of 50% of intron containing genes [192], illustrating the extent of nuclear decay also during pre-mRNA processing [193]. Similar to the cytoplasmic exosome, the nuclear exosome is composed of a 9 subunit core structure without catalytic activity and Dis3p and Rrp6p subunits which catalyze 3'-5'-RNA decay [194]. The nuclear exosome is typically recruited to target RNAs by RNA-binding cofactors which are involved in different nuclear RNA surveillance pathways. Defects in nuclear exosome components lead to accumulation of polyadenylated RNA in the nucleus, illustrating the tight coupling between nuclear polyadenylation control and decay. Many short-lived cryptic transcripts, mostly from pervasive transcription of intergenic regions, become only detectable upon inactivation of the nuclear exosome [190,191].

The yeast nuclear poly(A) binding protein Nab2p thereby has a central role in protecting mRNA from decay by the nuclear exosome since Nab2p depletion leads to a massive decrease in global mRNA levels through nuclear decay [195]. Mutations in the yeast *pap1* gene, which is encoding canonical yeast poly(A) polymerase, further result in reduced polyadenylation activity, but also

in increased clearance of polyadenylated transcripts by the nuclear exosome [196]. This observation also points towards cellular sensors for polyadenylation status which define nuclear RNA stability.

Several adaptor complexes target the nuclear exosome to specific RNAs: the yeast TRAMP complexes mark nuclear transcripts for exosome mediated decay. The complexes are composed of the non-canonical poly(A) polymerases (PAPs) Trf4p and Trf5p, the RNA binding proteins Air1p and Air2p as well as the helicase Mtr4p [197]. Mtr4p is the central hub connecting the TRAMP complex to the nuclear exosome. Different compositions of TRAMP subunits were reported which may guide RNA target specification [198]. TRAMP complexes target a broad range of RNAs within the nucleus, which is involving (pre-)mRNA, rRNA and a broad class of non-coding transcripts that result from pervasive transcription [199]. Trf4 and Trf5 polymerase subunits selectively add short oligo(A) tails to target transcripts, which serve as initiation site for the helicase Mtr4p. Mtr4p unwinds possibly structured RNAs and recruits the nuclear exosome [200]. Mtr4p has a central role in marking RNAs for decay by the nuclear exosome and is involved in several other nuclear decay pathways. How decay of regular, polyadenylated mRNAs is prevented is not fully understood, although protection of RNA with long poly(A) tails through PABPN1 or kinetic competition with RNA export could impact the balance between RNA maturation and decay [201]. TRAMP subunits are to some degree conserved in mammals but complemented by other pathways to diversify RNA regulation. Human and mouse TRAMP proteins mostly localize to nucleoli and function in rRNA processing [202,203]

Exosome-mediated nuclear decay can also be triggered by polyadenylation through canonical poly(A) polymerases: Intronless RNA reporters, which are not processed by the spliceosome, are hyperadenylated by canonical poly(A) polymerases and decayed by the nuclear exosome [204]. This 'PABPN1 and PAP mediated decay' (PPD) pathway has also been shown to target heterogenous groups of coding and non-coding RNAs, with evidence that splicing kinetics impact susceptibility for decay [205].

Related nuclear pathways for exosome recruitment are NEXT (nuclear exosome targeting complex) and PAXT (poly(A) tail exosome targeting). The NEXT [202] complex is composed of hMTR4 helicase, which mediates exosome recruitment, the adaptor protein ZCCHC8 and the RNA-binding protein RBM7. RBP7 is thereby directly loaded onto nascent RNAPII transcripts. The assembled NEXT complex then recruits the exosome for decay of newly synthesized transcripts, as for instance enhancer RNAs (eRNAs) or PROMPTs [206].

The PAXT complex on the other hand diversifies nuclear decay towards polyadenylated transcripts [182]: The zinc-finger protein ZFC3H1 is an adaptor between PABPN1 and hMTR4, which targets a number of polyadenylated RNAs for decay by the exosome.

Packaging of mature nuclear mRNAs for export by binding of export adapters and the interaction with the nuclear pore are required for translocation to the cytoplasm. Polyadenylation is thereby mechanistically linked to export. Synthetic addition of a poly(A) sequence at the end of a transcript facilitates for instance export of otherwise retained non-polyadenylated transcripts [207].

Export requires formation of densely packed mRNP complexes which can be shuttled through the nuclear pore. In yeast, nuclear poly(A) binding protein Nab2p is capable of binding both poly(A) tails and A-rich sequences within the transcript. Nab2p dimerizes and can hence form interactions across a transcript, which leads to compaction of the mRNA particle [208] and contributes to the compact, elongated structures observed in purified yeast mRNP complexes [209]. Nab2p also interacts with Mlp1, which binds to the nuclear pore complex. Mlp1 mediates retention of unspliced mRNAs through interactions with 5'-splice sites [210]. Disruption of Mlp1 and Nap2p interactions lead to nuclear accumulation of mRNA [211]. Pcf11, which is a component of the CF I complex and required for assembly of the 3'-cleavage and polyadenylation machinery (CPA), recruits yeast export adapter Yra1, which is an important step in producing export-competent mRNPs. Assembly of export adapters on newly synthesized RNAs has also been proposed as being required for disassembly of the CPA after completion of cleavage and polyadenylation: mutants of yeast export adapter Mex67p are unable to remove CF I components from polyadenylated RNA. Retention of CPA components on mature mRNA then triggers hyperadenylation of poly(A) tails and their decay [212]. mRNA binding of the PAXT pathway component ZFC3H1 further competes with the RNA export factor AlyREF, which impacts the balance between mRNA export and nuclear decay in mammals [213].

Release of 3'-end matured mRNA from the sites of transcription is an essential step in gaining export-competence. Depletion of the poly(A) binding protein Pab1 in yeast leads to retention of transcripts at transcription sites, and the same phenotype is observed for deletions of the nuclear deadenylase Pan2-Pan3 complex [214], which suggest that deadenylation and maturation of poly(A) tails may be required for transcript release from sites of transcription.

The duality of poly(A) tails in mediating protection from decay and stimulation of the exosome appears contradictory and can only be understood in terms of the kinetics of each process. Maturation of newly transcribed mRNA to an export-competent mRNP yet appears as key step

in preventing nuclear decay and could be a distinctive feature compared to cryptic transcripts which are degraded right away.

1.2.4 Regulation of mRNA translation through poly(A) tails

Poly(A) tails are essential for efficient mRNA translation. The classic closed-loop model proposes interactions of the 5'-cap with the 3'-end of a mRNA[215], which are mediated through poly(A) binding proteins (PABPs) and translation initiation factors (eIFs). eIF4E binds both to the mRNA 5'-cap and eIF4G which provides a scaffold for mediating interactions of eIF4E and PABP. eIF4G-PABP interactions enhance the cap-binding affinity of eIF4E, which in summary leads to synergistic effects on translation efficiency [216]. Poly(A) tails have been described as "translational enhancers", in the sense that poly(A) tail length is correlated with protein output *in vitro* which is attributed to increased translation initiation [217,218]. The poly(A) tail is thereby capable of recruiting the 40S ribosomal subunit independently of the 5'-cap [219,220], although efficient stimulation of ribosome recruitment requires synergistic action between poly(A) tail and 5'-cap [221].

Coupling between poly(A) tail length and translation output has been observed *in vitro* and *in vivo* during development in different model systems: During maturation of *Xenopus* oocytes, poly(A) tails of several mRNAs are selectively extended [222] or shortened [223] which modulates protein production of the affected transcripts. Similar mechanisms were observed in the slime mold *Dictyostelium discoideum* [224], mouse oocytes [225] or during *Drosophila* development[226]. During early development, regulation of poly(A) tail length is a mechanism to directly control mRNAs translation rates for given mRNAs.

The coupling of tail length and translation rates is not universal: genome-wide analysis of poly(A) tail length [157,158,227] did not show strong correlations between median poly(A) tail length per gene and translation efficiencies. On the contrary, poly(A) tails of highly expressed genes were found to be on average shorter in mammalian cell lines, yeast, *C. elegans* and in mouse liver. As a consequence, a "coupling" regime has been proposed in which poly(A) tail length is correlated to translation rates, as observed in early development, and as well as an "uncoupled" regime in which this relationship is absent [157]. The availability of cytoplasmic poly(A) binding proteins (PABPC1) has in this context been proposed as essential in regulating the transition between those regulatory modes: In "coupled" systems, PABPC1 is limiting, and long poly(A) tails compete more efficiently for PABPC1 binding, which increases mRNA translation. On the contrary, increasing PABPC1 concentrations throughout development

abolish the competitive advantage of long poly(A) tails leading to decoupling of translation rates and tail length [228].

Specification of protein output by changes in poly(A) tail length is of particular importance for understanding the precise temporal regulation of gene expression required during early development and oocyte maturation. During the first meiotic cycles, protein production of a number of genes is dependent on transcripts which are stored in a deadenylated form [229]. Efficient translational repression is mediated by maskin, an adapter protein which binds to Cytoplasmic Polyadenylation Element Binding protein (CPEB). Maskin competes with translation initiation factor eIF4G for binding to eIF4E[230] at the 5'-cap and exclusion of eIF4G then prevents ribosome assembly. This and the lack of poly(A) tail leads to translational repression of the target genes.

Throughout the meiotic cycles, mRNAs encoding meiosis regulators such as Cyclin B2, are polyadenylated and selectively translated. This happens in three waves for different sets of genes and drives progression from Prophase I to Metaphase II [231]. Re-adenylated mRNAs contain Cytoplasmic Polyadenylation Elements (CPEs) within their 3'-UTRs. Those are bound by CPE-binding proteins (CPEBs) in the cytoplasm [232]. Phosphorylated CPEB1 then recruits the cytoplasmic non-canonical poly(A) polymerase GLD-2 which extends the poly(A) tails [233]. The CPSF cleavage and polyadenylation complex is further required to define the polyadenylation site, which highlights a cytoplasmic role also for components of the cleavage and polyadenylation machinery. Different CPEBs confer specificity for distinct gene sets whose translation is required for different steps in meiosis [234].

A second example of cytoplasmic polyadenylation occurs in neurons: Both aurora kinase and CaMKII are activated through signaling events, and both in turn phosphorylate CEPB. Phosphorylated CEBP leads to localized polyadenylation and translation of synaptic transcripts, as for instance α-CaMKII itself [235,236].

Despite the role of cytoplasmic mRNA re-adenylation during early development and in specialized cases, such as polarized neuronal systems, cytoplasmic extension of poly(A) tails has not been reported as a general mechanism, for instance in counteracting deadenylation, although the role of cytoplasmic adenylation in differentiated systems is a matter for future investigations.

1.2.5 Deadenylation-dependent RNA decay

Canonical mRNA decay requires deadenylation of poly(A) tails, which typically precedes decapping by the Dcp1/Dcp2 complex. In yeast, deadenylation leaves poly(A) tails of around 10 nucleotides. This oligoadenylate tail is then bound by the Lsm1-7 complex which recruits Dcp1/Dcp2 for rapid decapping [237] and degradation in 5'- to 3'-direction by the exonuclease Xrn1[134] or from the 3'-end by the exosome complex[135].

Different deadenylase complexes were identified with Pan2-Pan3 and Ccr4-Not complexes being most relevant for deadenylation-dependent mRNA decay. The Pan2-Pan3 (Pabp1-dependent poly(A) nuclease) complex is a heterotrimer and composed of two Pan3 and one Pan2 subunit [238,239]. Pan2 deadenylase capacity was first identified in yeast since Pan2p deletion causes increased steady state poly(A) tail length [240]. Pan2p contains an exonuclease domain and a WD40 domain which mediates protein-protein interactions. Pan3 on the other hand does not have catalytic deadenylation activity and interacts with Pan2 through its C-terminus [241], and with its pseudokinase domain with poly(A) binding proteins [242]. In yeast, the Pan3-Pab1 interaction is highly relevant since Pan2-Pan3 deadenylation activity depends on Pab1 [243] and is possibly regulated through phosphorylation [244]. Human PAN3 is further recruited by GW182, a component of the miRNA-induced silencing complex (miRISC) to deadenylate mRNAs targeted for decay. Depletion of PAN2-PAN3 in human cell lines has been reported to impact trimming of long poly(A) tails, with little impact on global steady state poly(A) length distributions [245], such that the major deadenylase function has been attributed to CCR4-NOT. For some genes, yeast Pan2 deadenylates mRNAs directly after transcription, which hints at a role of yeast Pan2 in nuclear deadenylation [246].

CCR4-NOT is a large, multi-subunit complex with multiple functions in gene expression. Not1 (human CNOT1) is a scaffold-protein, which is bound by Not2 (CNOT2), Not5 (CNOT3), and Caf40 (CNOT9). The Not1 subunit is further involved in transcriptional repression of estrogen receptor expression, illustrating the broad role of CCR4-NOT in mRNA metabolism. CCR4-NOT deadenylase function is mediated by Caf1 and Ccr4 subunits, which are diversified into Caf1a/Caf1b (CNOT7, CNOT8), CCR4a/CCR4b (CNOT6C, CNOT6L) orthologs in human. Ccr4 and Caf1 thereby have different functions: while Caf1 is not able to deadenylate poly(A) tails bound by poly(A) binding protein Pab1, Ccr4 displaces Pab1 and efficiently trims Pab1 bound tails. While Ccr4 provides a universal deadenylation capacity, Caf1 is suggested to enhance deadenylation of instable transcripts with poly(A) tails less densely bound by Pab1 [247]. Similar roles have been attributed to human CCR4 and CAF1 subunits, which also harbor the

main deadenylation capacity and interact with human PABPC1 in a similar manner [245]. Similar to yeast Pan2-Pan3, CCR4-NOT has been implied in nuclear deadenylation of poly(A) tails for instance after induction of serum response genes [248]. Both Pan2-Pan3 and Ccr4-Not were found to mainly localize to the cytoplasm which also defines the compartment of the majority of mRNA deadenylation and decay [249,250].

Poly(A) tails adopt a helical conformation which aids substrate recognition by both complexes beyond the nucleotide preference for adenosines. Mutations or genetic ablation of Ccr4 or Caf1 cause different phenotypes, including defects in cell cycle and growth control [251,252] and physiological implication on bone growth [253]. On a molecular level, depletion of the human CCR4-NOT CAF1 subunits leads to genome-wide lengthening of poly(A) tails, while PAN2-PAN3 depletion has modest effects on bulk steady state poly(A) tail length, but increasing tail length is only observed for very long tails upon PAN2-PAN3 knockdown [245].

Other deadenylase complexes were identified with diverse and in part little understood functions [254]. The deadenylase PARN is important during early development and deadenylates maternal mRNAs during oocyte maturation [255]. In the cytoplasm, PARN interacts with CPEB for coordinated poly(A) tail deadenylation of cell cycle regulators during meiosis [256]. PARN has dedicated nuclear roles for instance in telomerase RNA maturation [257] or control of nascent mRNA processing during genotoxic stress [258]. PARN is also recruited by the nonsense-mediated decay (NMD) machinery [259], which illustrates its broad implication in different aspects of RNA metabolism and RNA quality control.

The deadenylase PDE12 is involved in degradation of double stranded RNAs, which are occurring for instance in cases of viral infection. PDE12 thereby removes oligoadenylate tails linked by a 2'-5'-phosphodiester bonds, which are generated by the interferon induced RNA decay pathway [260]. PDE12 is also required for removal of mitochondrial poly(A) tails [152] (s. 1.2.1).

A connection of polyadenylation and circadian biology has been discovered with nocturnin, a deadenylase which is rhythmically expressed in photoreceptors, liver and other tissues [261,262]. Nocturnin knockout leads to profound changes in lipid metabolism and obesity in mice [263].

Most mRNAs are degraded through a deadenylation-dependent pathway. Deadenylation has been identified as the rate-limiting step in mRNA decay[264]. Gene specific deadenylation rates integrate both basal deadenylation and *cis*-regulatory effects, for instance from AU-rich elements (AREs), which recruit destabilizing RBPs. RBPs can impact deadenylation rates by

recruiting deadenylase complexes. Genome-wide deadenylation rates have been determined using next-generation sequencing in combination with metabolic labeling and poly(A) tail length measurements [265] revealing that deadenylation rates encompass several orders of magnitude for different genes and are predictive of mRNA half-lives. Crr4-Not was identified as the major deadenylase enzyme complex in yeast, capable of completely removing poly(A) tails. The Pan2-Pan3 was yet unable to deadenylate poly(A) tails shorter than 25 nt [249]. Different models have been proposed how deadenylation is orchestrated by different deadenylases: The "biphasic" model states that long tails are first trimmed by PAN2-PAN3 until a length of around 110 nt is reached, upon which CCR4-NOT deadenylates the remaining tail, which is followed by decapping and exonucleolytic decay [250]. Other studies suggest that CCR4-NOT comprises the main deadenylase activity and that deadenylation patterns of PAN2-PAN3 and CCR4-NOT are largely overlapping [245].

Poly(A) binding proteins (PABPs) have an important role in modulating deadenylation. Poly(A) binding proteins can both protect mRNAs and increase their decay by recruiting the deadenylation machinery. Excess yeast Pab1 inhibits deadenylation [249], while depletion of Pab1 causes reduced deadenylation and translation rates [266] and leads to decapping prior to complete deadenylation [267]. In this context, Pab1 was shown to recruit deadenylase complex Pan2-Pan3 through interactions of its C-terminal domain with Pan3 [242] and Crr4-Not [247]. Whether the stoichiometry of PABP binding itself impacts deadenylation rates is unclear, but steady-state poly(A) tail length was not found to be correlated to PABPC1 binding [268], which indicates that poly(A) tails are not saturated by PABP binding.

Several proteins have been identified which stimulate deadenylation and hence impact mRNA turnover: the anti-proliferative protein BTG2 is transiently expressed after diverse signals, such for instance growth factors. BTG2 induces a global increase in deadenylation rates by recruiting the CCR4-NOT complex[269], thereby providing a general switch for tuning RNA decay. RNA binding proteins (RBPs) with destabilizing properties operate through similar mechanisms: TTP or PUF3 recruit CCR4-NOT upon binding of *cis*-regulatory motifs such as AREs which leads to deadenylation and decay of RNAs. RBP binding affinities and specificity for certain motifs thereby determine deadenylation kinetics [270].

Besides addition and removal of non-templated adenosines, poly(A) tails can be modified with other ribonucleotides which can have an impact on deadenylation and decay kinetics. Genome-wide analysis of poly(A) tail sequences identified guanosines most enriched towards the 3'-ends of longer poly(A) tails [227]. The terminal nucleotidyl transferases TENT4A and TENT4B

were identified in attaching "mixed" tails to the 3'-end of messenger RNA. Guanosines in tails act as a barrier for deadenylases, which deadenylate tails containing guanosines with drastically reduced efficiency. This can be explained by disruption of the helical poly(A) tail conformation and different substrate specificities for guanosines compared to adenosines [271,272]. Canonical poly(A) polymerases (PAPs), responsible for synthesis of poly(A) tails on nascent mRNAs, is also capable of incorporating non-A nucleotides into poly(A) tails *in vitro* [273].

Addition of terminal uridines is another mode of regulating RNA stability. Tails containing uridines have been detected in a number of species from yeast to human and are typically a mark of RNA degradation. Occurrence of uridines was initially observed upon miRNA-directed mRNA cleavage [274] and on non-coding RNAs such as U6 snRNA [275]. Yeast Cid1 was shown to uridylate mRNAs independent of previous deadenylation. Uridylation enhances decapping, likely through binding of Lsm1-7 proteins to uridylated tails which then recruit the Dcp1-Dcp2 complex. Yeast Cid1-uridylation is hence an alternative pathway to deadenylation-dependent decay [276].

In human, uridylation is associated with short poly(A) tails and detected for most mRNAs. TUT4 and TUT7 enzymes are responsible for addition of oligo-uridine tails to mRNA 3'-ends and their depletion causes stabilization of target mRNAs. TUT4/TUT7 tailing is less efficient on poly(A) tails bound by PABPC1, showing that uridylation is associated with later steps of mRNA decay [277]. Controlled uridylation is utilized by different biological pathways for enhancing RNA decay: induction of apoptosis causes global mRNA degradation, which is mediated by increased uridylation and decay by the cytoplasmic 3'-5'-exonuclease DIS3L2 [278], which preferentially targets uridylated RNA [279]. Uridylation further occurs throughout oocyte development, for selective degradation of mRNAs which is required throughout developmental stages [280]. Histone RNAs, which do not have a poly(A) tail, but are stabilized by a hairpin structure within their 3'-UTRs, are also degraded by selective uridylation at the end of S-phase [281]. The histone mRNA hairpin structure is bound by SLBP which recruits enzymes for uridylation of 3'-ends. Oligo uridine tails are then bound by Lsm1-7 proteins which recruit Eri1, a specific exonuclease which degrades histone mRNAs [282].

Both guanosines and uridines occur at low frequencies in poly(A) tails: Less than 2% of mRNA transcripts had uridinylated tails and less than 1% were shown to contain guanosines in extracts of human and mouse cell lines [227]. Cytosines have also been identified at comparable levels to guanine and uridine [283], although no biological consequence of 'cytosinylation' have been described so far.

1.2.6 mRNA localization and decay

Parts of the cytoplasmic mRNA decay machinery have been shown to be enriched in subcellular, granular structures termed processing bodies (P-bodies). Those membrane-less compartments share liquid-liquid phase separation properties with other subcellular granules, such as stress granules (SG) or Cajal-bodies. P-bodies are composed of translationally repressed mRNAs along with more than 100 proteins for instance components of the mRNA decay machinery as Xrn1, the Ccr4-Not complex, the decapping complex Dcp1/Dcp2 as well as proteins involved in miRNA mediated repression [284]. P-body formation is RNA dependent and phase separation is caused by different biophysical effects such as multivalent protein-protein contacts along intrinsically disordered protein regions (IDRs), or electrostatic interactions between negatively charged RNA and positive amino acid residues [138]. A diverse range of protein-coding RNAs have been found enriched in P-bodies, although typically not more than 30% of the cellular mRNAs are localized to P-bodies [284]. Poly(A) tails of stored mRNAs have a broad ranger range in length when comparing to the cytoplasmic fraction. mRNAs targeted to P-bodies are typically translated less efficiently since RNA binding proteins involved in translational repression such as 4E-T, compete with eIF4G for association with cap-binding protein eIF4E [285]. CPEB1, which mediates repression of mRNA, is also enriched in P-bodies during early development [286] . P-bodies are highly dynamic structures that change during cell growth and in response to extracellular environment, such as nutrition availability or osmotic stress [287]. Despite enrichment of decay components, P bodies are not required for decay. On the contrary, it has been shown that blocking RNAi pathways leads to disappearance of P-bodies [288], which is suggesting that P-bodies could be a consequence of ongoing RNA decay.

1.2.7 Poly(A) tails integrate signals on RNA stability to determine decay

Poly(A) tails are wcentral hubs for deadenylation-dependent decay of messenger RNA in the cytoplasm. Deadenylation rates are most predictive of mRNA half-lives and different decay signals, for instance AU-rich elements, mediate destabilization through recruitment of RNA-binding proteins. Poly(A) tails are important for stimulating translation by mediating interactions between 5'-cap and 3'-poly(A) tail, which greatly enhances initiation. In how far the actual tail length is relevant for translation is a matter of ongoing research since direct correlation of tail length and translational efficiency has only been observed *in vitro* and during early development. Beyond the well-established functions in translation and decay, poly(A) tails and the polyadenylation machinery impact different aspects of nuclear mRNA processing: Nuclear decay through PAXT and hyperadenylation dependent PPD pathways involve

interactions with poly(A) tail binding proteins, which are also involved in connecting mature, polyadenylated mRNAs with export adapters and the nuclear pore. Intermediates of those pathways are transient and as such difficult to investigate, but nuclear RNA decay may have important roles for presetting cytoplasmic mRNA abundance. The impact of poly(A) tails on gene regulation has for the largest part been investigated by bulk analysis of poly(A) tail profiles or for individual genes. Only recently genome-wide methods for poly(A) tail determination were integrated into kinetic measurements to explore poly(A) tail dynamics [265]. The understanding of nuclear RNA processing will greatly profit from genome-wide analysis of poly(A) tails since the many pathways for nuclear decay are not well described with respect to their target genes. This work contributes here first by providing easy-to-use technology for sequencing poly(A) tail and by applying the method for investigating nuclear poly(A) tails. Nuclear poly(A) tail analysis revealed that tails are shortened already in the nucleus after being synthesized at a relatively uniform length of more than 200 nt.

1.3 Technical basis for investigating RNA 3'-ends and poly(A) tails

Different experimental methods have been developed over the last decades which enable measurements of exact polyadenylation sites and poly(A) tail length profiles. Experimental methods can be grouped into low-throughput approaches which measure poly(A) tails or polyadenylation sites of individual isoforms or genes and high-throughput methods which enable quantification of thousands of genes in parallel, usually involving high-throughput sequencing. Measurements of polyadenylation and poly(A) tail length for individual genes can be performed for instance using gel-based radioactive labeling, northern blotting or PCR based amplification of poly(A) tails.

Besides experimental approaches, several databases collect polyadenylation sites and 3'-UTR isoforms of different species, tissues and experimental conditions [289,290]. Those databases are curated from different studies and experimental techniques, which are typically based on high-throughput RNA sequencing approaches. A database collection of poly(A) tail length is not available yet.

This chapter describes the technical foundations for mapping polyadenylation sites, mRNA isofoms as well as measuring poly(A) tail length and sequence on a genome-wide scale using high-throughput sequencing technologies. Both experimental ('wet-lab') and computational concepts are being discussed.

1.3.1 Mapping of polyadenylation sites and poly(A) tail length for individual genes

Alternative polyadenylation (APA) describes a mechanism to generate transcripts of the same gene with differences in 3'-UTR length and sequence. Changes in 3'-UTR length alter the *cis*-regulatory repertoire of a transcript, for instance by inclusion of miRNA binding sites [291]. Different experimental methods have thereby been developed to distinguish 3'-UTR isoforms of the same gene.

3'-UTR isoforms can be analyzed by Northern Blotting using probes for extended (distal) versus proximal 3'-UTR isoforms. Northern blotting relies on hybridization of complementary DNA probes to total RNA extracted from a sample of interest, which has been size separated by electrophoresis and transferred to a membrane[292]. DNA probes are radioactively labeled or can be detected by conjugated antibodies (digoxigenin labeling) to visualize and quantify individual transcripts [292]. Northern blotting is an amplification free method which enables specific detection of RNA but is labor sensitive and highly dependent on probe designs and experimental parameters.

As an amplification-based alternative method, reverse transcription–PCR (RT-PCR) with specific primer pairs which are binding to the extended part of the longest 3'-UTR isoform can be utilized to validate isoform expression. Isoform-specific expression levels can be quantified by quantitative real time PCR (qRT-PCR) also using primers which exclusively hybridize to the 3'-UTR sequences present only in the longest isoform. 3'-Rapid Amplification of cDNA ends (3'-RACE) is another method for targeted amplification of amplicons which include the 3'-UTR ends of different isoforms. For this a reverse transcription primer is used which includes an oligo-dT stretch annealing to the poly(A) tail and a specific adapter which is used for later PCR amplification. With this approach all 3'-UTR isoforms for a given gene can be amplified [293].

Fluorescent *in situ* hybridization (FISH) is an imaging-based approach in which isoform expression can be visualized by annealing of complementary probes which are labeled with fluorophores and within intact tissue structures [294].

Similar methods can be used for quantification of poly(A) tail length: Northern blotting can also be adapted to infer poly(A) tail length for specific genes: Gene-specific oligos are annealed to extracted RNA with or without addition of additional oligo-dT oligos. Hybridase (RNAseH) is then added which cleaves RNA:DNA duplexes, which are formed by annealing of gene-specific and in one condition oligo-dT oligos. After digestion RNA is separated, blotted, and detected using gene-specific primers. Selective annealing with oligo-dT primers digests the poly(A) tails, such that the difference between sizes of +/- oligo-dT conditions on Northern Blots correspond to the poly(A) tail length profile [157].

Poly(A) tail length profiles can also be analyzed by PCR-based methods: The poly(A) test (PAT) assay measures poly(A) tail length for individual genes by first hybridizing short oligo-dT primers to poly(A) tails. The annealed primers are then ligated and a 3'-terminal oligo is added, which contains a PCR handle. The number of annealed oligos is proportional to the poly(A) tail length since longer tails can be coated by more oligo-dT primers. The PCR handle is then used as a primer binding site for priming reverse transcription. cDNA is then PCR amplified using gene specific primers along with a primer binding the terminal PCR handle, which amplifies the poly(A) tail and parts of the transcripts. The amplicon length profiles are then resolved by capillary or gel electrophoresis to determine the amplicon length profiles. The actual poly(A) tail length (distribution) can be obtained by subtracting the number of bases that result from amplification of the transcript body to the 3'-UTR end [295]. Alternative PAT assay methods append tails of guanosines or inosines to the existing poly(A) tail which then serves

as a primer binding site for reverse transcription and PCR amplification. This procedure also includes the full poly(A) tail into the amplified fragments and enabled poly(A) quantification for individual genes [296,297].

Bulk poly(A) tail length can be interrogated by incubating total RNA extracted from a sample of interest with RNAse T1, an enzyme which specifically cleaves after guanosines. RNAse T1 digestion leaves poly(A) tails effectively intact, which can be purified, labeled, and visualized by gel electrophoresis. The advantage of this method is its simplicity for amplification free analysis of global tail length profiles. On the other hand the method is not specific and the compositions of RNA species from which poly(A) sequences are derived cannot be resolved [149,175], which makes comparisons between experimental conditions difficult.

Transcripts can be separated based on their poly(A) tail length using differential poly(U) chromatography approaches [298,299]. Polyadenylated RNA is therefore bound to poly-uridine sephadex-resins, and subsequently eluted using increasing temperatures. Longer poly(A) tails increase the number of A-T Watson-Crick pairings to the resin and hence effective melting temperature of the hybridized transcripts, although other A-rich transcript regions could contribute as well. Poly(A) tail length of different genes can be resolved by temperature dependent elution, since shorter poly(A) tails melt and elute at lower temperatures from resins [298]. RNA composition can then be analyzed for each eluted fraction, either for individual genes or using microarrays in order to 'bin' protein-coding genes by poly(A) tail length [299]. Microarray analysis of those RNA poly(A) fractions was the first approach for investigating differences in poly(A) tail length between genes on a genome-wide scale.

1.3.2 Sequencing-based methods for analysis of poly(A) tails and polyadenylation sites

Genome-wide analysis of gene expression regulation was revolutionized by high-throughput sequencing of DNA and RNA and different sequencing platforms and protocols exists, of which Illumina, Nanopore and PacBio sequencing are discussed in more detail. Illumina sequencing is currently the most frequently used sequencing platform and based on sequencing-by-synthesis chemistry (Figure 3 A).

For Illumina sequencing, in a first step cDNA sequencing libraries are produced. Sequencing libraries contain diverse DNA fragments, for instance generated from extracted RNA which is reverse transcribed into cDNA, along with 5'- and 3'-adapters. Sequencing libraries are bound to oligonucleotides on a flow cell which hybridize to the cDNA adapters. Bound cDNA fragments are then amplified in a PCR reaction which generates dense clusters of several

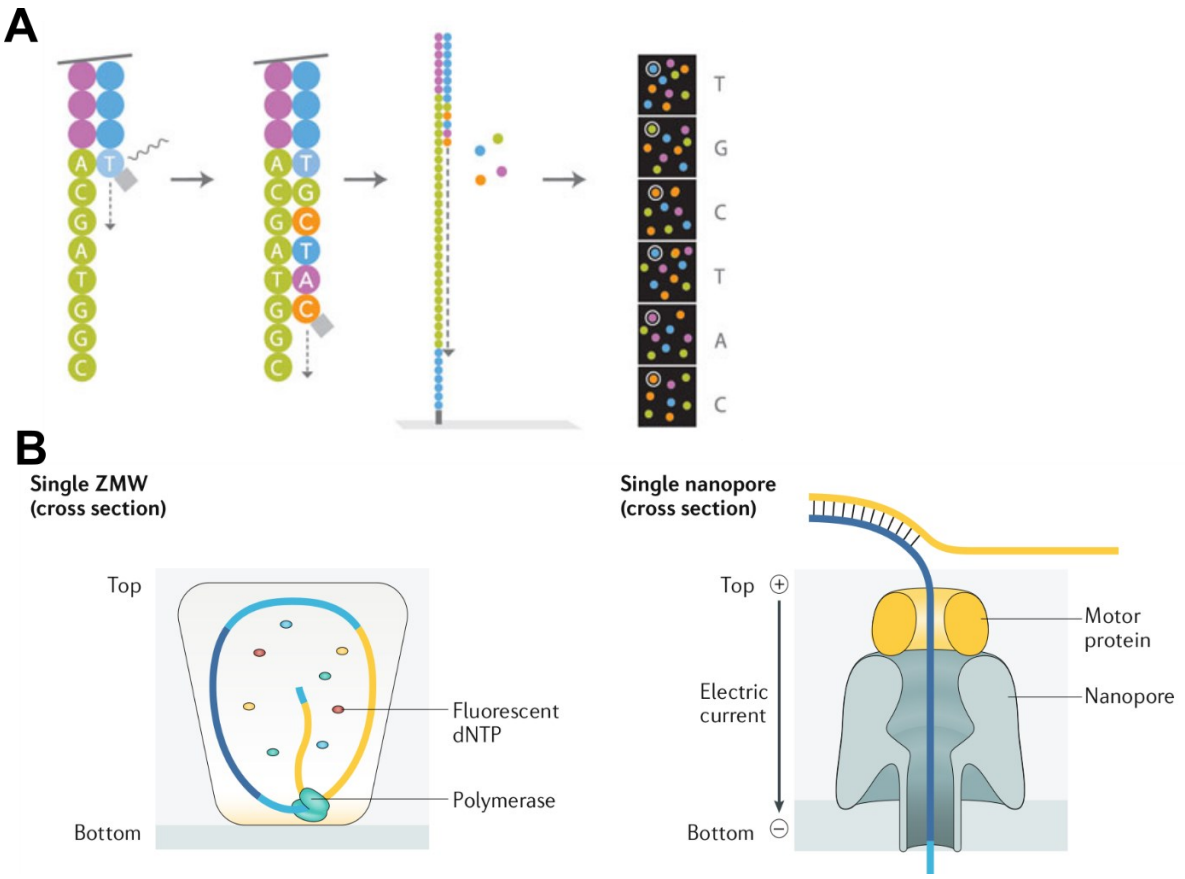

Figure 3 **Technical schematic of sequencing technologies**
A) Illumina Sequencing-by-Synthesis (adopted from Illumina Inc. 2013). **B)** left**:** PacBio Zero Mode Waveguide sequencing. Right: Nanopore sequencing (adopted from Longsdon et al. 2020)

thousand copies of each DNA fragment (cluster generation step). A sequencing primer is then annealed to the adapters contained in each amplicon. This primer is enzymatically extended by incorporating a single, fluorescently labeled base which is complementary to the nucleotide adjacent to the primer in each fragment of the original sequencing library. The extended base contains a blocking group which prevents incorporation of additional nucleotides. The fluorescence signals for each amplicon cluster are recorded by a camera for different excitation and emission channels for optimal optical separation of the four fluorophores encoding four DNA bases. The fluorescent moiety is cleaved along with the blocking group, which enables subsequent rounds of extension, imaging, and cleavage, which is summarized as a sequencing cycle [300]. Up to 500 sequencing cycles can be performed in one run, depending on the Illumina sequencing system. DNA amplicons can be sequenced from both ends, which is referred to as 'paired-end' sequencing [301].

Images of fluorescence intensity for each sequencing cycle and cluster need to be converted to 'reads', which refers to the derived nucleotide sequence for each cluster. This process is termed 'base calling'. First, individual amplicon clusters are computationally extracted from images and intensity profiles across the different channels are extracted for each cycle and each cluster. Base calling algorithms, such as Bustard, infer the most likely nucleotide sequence given the observed intensity profiles for each cluster while taking into account effects such as phasing, which refers to skipping of nucleotide incorporation for some amplicons [302]. Illumina sequencing can produce millions to billions of reads and is currently the sequencing platform with the highest output of sequenced DNA molecules [301].

Illumina sequencing has certain requirements to library designs to assure correct base calling and generation of high-quality reads corresponding to the original DNA sequence. Sequence complexity of the library fragments must be sufficiently diverse such that neighboring DNA clusters differ in their sequence over different cycles. Sequenced amplicons should also not be composed of homopolymer sequences since the base caller can have problems in distinguishing individual sequencing cycles. Both these restrictions are problematic for the analysis of poly(A) tails since those are (mostly) identical in sequence for different genes.

RNA-Seq is by now a standard method for investigating RNA expression patterns using short-read high-throughput sequencing [303]. RNA-Seq is typically performed after poly(A) selection from total RNA extracts to enrich for polyadenylated mRNAs or selective depletion of ribosomal RNA sequences, which would otherwise dominate the sequencing libraries and resulting datasets [304]. RNA is next fragmented, and an adapter is ligated which serves as a primer for reverse transcription. After reverse transcription, cDNA libraries are amplified using PCR primers which contain adapter sequences specific to the sequencing protocol and platform used.

RNA-Seq enables different types of analysis, for instance quantification of mRNAs, differential gene expression analysis between experimental conditions or quantification of alternative splicing events. Converting reads obtained from high-throughput sequencing into interpretable statistics, such as counts for each gene, requires different levels of computational processing: in a first step, the genomic origin of each read must be determined in a process termed alignment. Algorithms such as Needleman-Wunsch [305] and Smith-Waterman [306] were developed as dynamic programming solutions which optimize a similarity score between two sequences. Despite being exact, their asymptotic computational complexity is multiplicative (O(n*m)), which leads to poor performance when aligning for instance a read with a length of

75 nt to 3 billion bases of the human genome. Heuristic approaches such as BLAST [307] first identify possible matches of query substrings to a database of genomes. This is facilitated by 'k-mer indexing' of the database which stores genomic locations of each k-mer in the database and speeds up searches. Seed matches of a query sequence are then extended which enables calculation of scores for potential alignments. Seed extension algorithms reduces the runtime and enable mapping of DNA sequences against large databases such as the human genome. Mapping short reads from high-throughput sequencing requires even faster approaches that enable mapping of millions of reads within hours. Index structures such as Burrows-Wheeler index, which is used by the short read mapper Bowtie [308] or suffix arrays as used by the mapper STAR [309] enable fast searching for potential genomic locations of seeds which are then extended to find optimal matches.

Mapping RNA reads to the genome is particularly challenging since pre-mRNA splicing generates discontinuous reads which do not align as continuous sequences but contain large gaps across introns. Additional challenges include errors in sequencing reads, which occur at a rate of 0.1 – 0.5% [310] for Illumina sequencing, genetic variation of the sample's genome compared to the commonly used reference genome [311,312] or mapping of reads from short RNAs such as miRNAs [313]. Reads can also be mapped to the annotated transcriptome for a given species which narrows the search space and speeds up alignments. This is sufficient for many applications since transcriptomes of commonly used model systems are well annotated [314].

After mapping, reads aligned to each gene can be quantified using existing genome annotations (e.g. Gencode https://www.gencodegenes.org/) and software packages as featureCounts [315], which handle the assignment of alignment coordinates for each read and genomic positions of genes or individual exons. Gene expression between different experimental conditions can then be compared using software packages as EdgeR [316] or DESeq2 [317]. Counts for each gene in each experimental conditions are typically modeled as negative binomial distributions and variance stabilization is performed to account for uncertainty in measurements across typically few replicates. Fold-changes in expression are then tested for significance using parametric testing procedures, such as the Wald-test [316,317], to identify differentially expressed genes in combination with multiple hypothesis testing correction by setting an appropriate False Discovery Rate (FDR).

Analysis of RNA isoform expression is challenging when using short reads since an individual read (or pair of reads for paired-end sequencing) in few cases spans a full transcript. As a consequence, isoform expression in a sample can be model by quantifying reads which map to

alternative exons or splice sites which are characteristic of a given isoform and appropriate statistical treatments of those events [318]. Isoform expression can be quantified both based on known isoform annotations or *de novo* transcriptome assemblies for RNA-Seq. Despite being useful in investigating genome-wide splicing patterns, short-read sequencing is intrinsically limited in analyzing RNA processing events for individual molecules: investigating for instance the splicing status of several adjacent exons is complicated since individual reads cannot be assigned to the mRNA molecule of origin.

Standard RNA-Seq protocols typically involve fragmentation of RNA and obtained reads do not span full transcripts. Analyzing transcript 3'-ends and alternative polyadenylation using standard RNA-Seq is yet less efficient since comparably few reads cover the exact 3'-end of a read. Different experimental protocols for RNA-Seq library preparations were developed which ensure that cDNA fragments in a library originate from the RNA 3'-ends. This can be achieved for instance by fill-in reactions which later enable priming the sequencing reaction directly at the 3'-UTR end [95]. Another method relies on first performing fragmentation and poly(A) selection and then digest of the poly(A) tail and ligation of a sequencing adapter which primes the sequencing reaction directly starting at the end of the 3'-UTR [319].

Quantifying poly(A) tail length on a genome-wide scale using sequence-by-synthesis requires adaptation of the sequencing platform or its base calling software, both of which is challenging to implement. Two methods have been developed for quantifying poly(A) tail length and sequence based on Illumina sequencing. Directly sequencing through poly(A) tails is not possible using standard Illumina protocols since sequencing through homopolymer sequences such as poly(A) tails will produce erroneous reads.

Poly(A) tail length profiling by sequencing (PAL-Seq) [157] measures a fluorescent signal proportional to the poly(A) tail length on an Illumina sequencer. A biotinylated adapter is first ligated to the RNA poly(A) tails by splint ligation. RNA is then partially digested by RNAse T1 which cleaves selectively after G nucleotides, leaving the tail intact. A 5'-RNA adapter is ligated, and RNA is reverse transcribed and loaded on a cBot Illumina cluster generator. A custom sequencing primer is then annealed. This primer is enzymatically extended with dTTP and biotin-dUTP nucleotides where the poly(A) tail is serving as a template. The actual sequence-by-synthesis reaction is then performed from the other end of the cDNA fragment upstream of the polyadenylation site to determine the transcript sequence. Fluorescent streptavidin is added, which binds to the incorporated biotin-dUTP proportional to poly(A) tail length, which is then quantified based on a calibration curve using standards of known length.

PAL-Seq is not directly sequencing poly(A) tails at single base resolution and hence not able to interrogate non-A nucleotides. On the other hand, the method is in principle not limited in maximum poly(A) tail length which can be quantified. PAL-Seq enabled detection of poly(A) tails of up to 10,000 genes with 2,800 genes having more than 100 poly(A) tags (i.e. valid reads).

Another approach, termed TAIL-Seq, directly sequences through poly(A) sequences using Illumina sequencing. The obtained cluster images are then loaded from the sequencer and a hidden Markov model is applied to distinguish poly(A) tails from 3'-UTR sequences and to alleviate the inherent problems with directly base calling homopolymer poly(A) tail sequences [227]. Extracted total RNA is first depleted of highly expressed non-coding RNAs (rRNAs) and a biotinylated 3'-adaptor is ligated to the RNA 3'-ends. RNase T1 is next used to fragment RNA by cleavage after guanosines. Adaptor-ligated fragments are next purified by streptavidin purification and size selection. A 5'-adaptor is ligated, and RNA is reverse transcribed, PCR amplified and sequenced on an Illumina HighSeq sequencer. One sequencing read used for mapping to the genome to identify the transcript of origin for a read. The second read, which is 251 nt in length, is used for identification of the poly(A) tail length. The poly(A) tail sequence is identified by training a Gaussian mixture hidden Markov model on fluorescence intensity measures for each nucleotide and cluster of TAIL-Seq data from cDNA standards with known poly(A) tail length. Since the method is based on directly sequencing poly(A) tails, poly(A) tail modifications such as terminal uridines can be assayed as well. The maximum detectable poly(A) tail length is limited by the read length of 230 nt. TAIL-seq identified poly(A) profiles of ca. 4000 genes with more than 30 valid poly(A) reads per gene. A more sensitive improvement of the protocol termed mTAIL-Seq [320] utilizes an oligo-dT hairpin splint oligo to increase ligation efficiency which greatly increases the fraction of detected mRNAs in mTAIL-Seq libraries.

An improved version of PAL-Seq [265] also uses a direct poly(A) sequencing approach in combination with a modified experimental protocol in which the ligated splint oligo is partly modified with a terminal A to enable efficient ligation to terminal uridylated poly(A) tails. A updated version of PAL-Seq further enables sample multiplexing for sequencing using barcoded reverse transcription primers [228].

1.3.3 Third generation long read sequencing of RNA and DNA

With PacBio and Oxford Nanopore, two new sequencing platforms were commercialized within the last years which break the limit of short reads length obtained from sequencing-by-synthesis applications and enable sequencing of DNA or RNA which are several kilobase pairs (kbp) in length.

The PacBio platform is based on single molecule real time (SMRT) sequencing of individual circularized DNA templates (Figure 3 B). For SMRT sequencing DNA libraries are ligated with a hairpin adapter which produces circularized amplicons (SMRTbell templates). Templates are loaded onto SMRT cells which are composed of hundred of thousands of zero-mode waveguides (ZMWs). Zero-mode waveguides refer to structures with dimensions smaller than the wavelength of light, which are each loaded with a single DNA polymerase and provide confinement for microscopic measurement of fluorescent nucleotide incorporation. The sequencing reactions are primed by DNA oligos complementary to the bell adaptors. During synthesis of the new DNA strand, fluorescent nucleotides are incorporated and fluorescence signals are measured upon incorporation in 'movies', which refer to signal intensities over time. Movies are then converted into base sequences [321]. Since the DNA template is circular, the polymerase reaction is continuously generating passes of the same sequence. Individual base measurements have relatively high error rates of around 10%, yet generation of circular consensus sequence (CCS) reads from individual passes increase accuracy to >99% [322]. PacBio read length can reach more than 60 kbp, with a median of around 10 kbp [323]. Despite advantages in read length, throughput and costs per base are higher compared to Illumina sequencing [324].

PacBio sequencing has been applied in different areas of genomics research. Long reads are particularly beneficial for improving genome references [325,326] or mapping structural genomic variation, which is found in many cancers and difficult to address by short read sequencing [327]. PacBio sequencing is further a useful tool for investigating RNA isoform expression (IsoSeq) since the connectivity of exons is easily inferred from individual long reads [328]. IsoSeq yet relies on oligo-dT-primed reverse transcription for preparation of cDNA sequencing libraries, which do not include the complete poly(A) tail. IsoSeq involves PCR-based amplification of cDNA libraries, which is typically biased towards amplification of shorter cDNA amplicons. In order to increase coverage of transcripts, cDNA libraries can be separated into different length bins and amplified again [329].

Nanopore sequencing emerged as second platform for long read sequencing of DNA and RNA (Figure 3 B). Nanopore sequencing is based on funneling single stranded DNA or RNA through a pore protein embedded in a membrane. DNA is thereby translocated by a motor protein. Translocation through the pore produces alterations in current through the nanopore, depending on the analyzed DNA sequence. This can be utilized to decode the DNA sequences in real-time and represents an alternative to sequencing-by-synthesis based methods [330]. Sequenced fragments are in principle not limited in length and individual reads longer than 1 million basepairs (1 Mbp) have been reported [331], which were produced using a specific protocol which optimizes extraction of non-fragmented high molecular weight DNA [332]. Oxford Nanopore MinION sequencers are small, portable and have successfully been used outside of dedicated genomics laboratories, for instance for determining mutation rate of the Ebola virus during the outbreak in 2016 western Africa [333]. Nanopore sequencing comes with the downside of relatively high error rates per base (5 – 15%), despite recent improvements, which are mainly driven by the chemical design of the nanopores and advanced machine learning models for converting raw current signals into sequences [334].

Nanopore sequencing enables direct sequencing of RNA without intermediate reverse transcription and PCR amplification steps [335], simply by splint-ligation of a Nanopore sequencing adapter to the poly(A) tail. Despite great potential for unbiased and fast 'direct RNA sequencing', current protocols require high amounts of poly(A) selected RNA as input, which becomes a bottleneck for analyzing clinical samples. Nanopore sequencing enables additionally the identification of post-transcriptional RNA modifications such as N6-methyladenosine (m6A) or 5-methylcytosine (5-mC), which are relevant for instance in shaping translation efficiencies [336].

The ability to detect modified bases has also been used for measuring RNA stability for individual mRNA isoforms: for this approach, cells are labeled with the uridine analogue 5-EU, which is incorporated into newly synthesized RNA. The newly synthesized molecules can be distinguished after Nanopore sequencing based on distinct current profiles of 5-EU moieties [337].

High error rates of Nanopore sequencing yet complicate identification and quantification of RNA isoform expression, for instance regarding the precise detection of splice sites [338]. This can be overcome by parallel short-read RNA sequencing [339]. Isoforms are then typically annotated by grouping reads by transcription-start and -end sites as well as splice junctions. A particular challenge relates to fragmented or low quality RNA since in those cases 'real' transcription start sites have to be distinguished from artefacts caused by fragmented 5'-ends.

Long read third generation sequencing has been applied for profiling of gene and isoform specific poly(A) tail length. Workman et al. developed 'nanopolish polyA', a hidden Markov model which segments the raw signals ('squiggles') from Nanopore direct RNA sequencing to identify the length of each sequenced poly(A) tail [340]. This method does not directly report the sequence of the poly(A) tail and the individual nucleotide composition of tails cannot be assessed by this method. Analysis of RNA isoform expression together with poly(A) tail length identified more than 200 genes with multiple expressed isoforms and differences in associated poly(A) tail length.

The PAIso-Seq method [283] anneals a template primer to poly(A) tails which is extended to introduce an adapter to the mRNA 3'-end which is used to anneal a primer for reverse transcription. This procedure preserves the full poly(A) sequence for PacBio sequencing. This approach enables analysis of poly(A) tail length and sequence at nucleotide resolution. The authors also identify different RNA isoforms of the same gene with significant differences in associated poly(A) tail length profiles. Additionally, a significant enrichment of non-A nucleotides was found within poly(A) tails, with uridines mostly occurring within short poly(A) tails, while guanosines and cytosines are most enriched in long tails.

1.4 Investigating genome-wide polyadenylation for different steps of mRNA metabolism

Sequencing based techniques for investigating polyadenylation sites and poly(A) tail length were important for generalizing many of the hypothesis regarding the regulatory role of poly(A) tails and polyadenylation site choice on gene expression and RNA fate. One example is the genome-wide determination of deadenylation rates using PAL-Seq [265], based on metabolic labeling of RNA, which validates previous findings that deadenylation is the rate-limiting step[264] in RNA decay and therefore dictates RNA half-life. In other cases, observations from investigating individual genes could not be directly confirmed on a genome-wide level. One example is the correlation of poly(A) tail length and translation efficiency, which could not be shown for steady state poly(A) tail length distributions. A key challenge for sequencing-based methods is the balance between read length and information content of individual sequenced molecules versus throughput and costs. Recent studies uncovering for instance the order and kinetics of nascent mRNA splicing [43,341] illustrate the power of long read sequencing to uncover the coordination of RNA processing events.

RNA metabolism involves several steps, from transcription to nuclear maturation, export, translation or storge in the cytoplasm and regulated decay. As described throughout this

introduction, many processing steps are interconnected which poses challenges for experimentally investigating isolated steps and discerning direct and indirect effects of experimental perturbations.

Many textbooks separate gene expression control into transcriptional regulation (for instance controlled by transcription factors) and post-transcriptional regulation operating at the level of miRNAs, RNA-binding proteins and 3'-UTR-mediated regulation. Some studies yet show that RNA binding proteins impacting alternative polyadenylation site choice associate with transcription initiation complexes at promotor sites [107]. Important effectors of post-transcriptional processing, as for instance the CCR4-NOT complex, which removes poly(A) tails, were on the other hand first identified as a transcriptional activator or repressor [342]. Both examples show the plasticity which evolved with the complex networks regulating gene expression and the future will certainly bring about many new unforeseen links underscoring the wiring of RNA processing.

RNA metabolism is dynamic, and many maturation steps are fast and transient. One example is nuclear RNA processing, where different pathways regulate transcript decay and export, but the exact targets and mechanism are mostly unknown since many nuclear RNA species are very short lived and only detectable upon depleting components of the nuclear decay machinery, which leads to accumulation of respective target RNAs.

RNA sequencing provides snapshots of RNA abundance at a given timepoint and special protocols, which involve for instance labeling of RNA produced in a defined time interval, are required to resolve temporal dynamics. Recent advanced in long read sequencing provide an opportunity for understanding the role of individual RNA isoforms and investigation of splicing kinetics.

Poly(A) tails have been mostly investigated with a focus on its cytoplasmic roles, which is where most RNA decay occurs. The nuclear role of poly(A) tails is less understood and the discrepancy between the described poly(A) tail length right after transcription in the nucleus and the steady state length, which is much shorter, requires for deeper investigation of the nuclear contribution on poly(A) tail metabolism and links to other nuclear maturation steps.

2 Aims

mRNA poly(A) tails regulate gene expression on different levels, for instance by influencing RNA stability [136] and translation efficiency, in particular during early development [343]. Poly(A) tails are conserved and universal features of eukaryotic messenger RNAs, which highlights their importance for regulating gene expression and need for sound scientific characterization.

Poly(A) tails are also involved in different nuclear RNA maturation pathways and quality control steps [344]. These are often transient, which makes experimental investigation challenging and. Yeast mutants for nuclear factors of RNA decay and export showcased examples of changing nuclear poly(A) tail length. Since poly(A) tail deadenylation is the rate-limiting step in RNA decay, changes in nuclear poly(A) tail length likely impact turnover in the cytoplasm.

Available methods for genome-wide characterization of poly(A) tails were yet complicated and did not cover full transcripts. The **first aim** of this thesis was hence development of a high-throughput sequencing-based method and for quantifying poly(A) tail length and sequence in context of complete mRNA molecules. With FLAM-Seq, a novel experimental protocol for PacBio sequencing was developed including a computational workflow for data analysis and poly(A) extraction.

The **second aim** was investigation of poly(A) tail regulation for different RNA isoforms by computational reconstruction of 3'-UTR isoforms and analysis of sequence motifs. Previous studies also reported incorporation of non-adenosine nucleotides into tails [271,277], which could be analyzed on a genome-wide scale.

Poly(A) tail synthesis after transcription was shown to produce poly(A) tails of around 250 nt in length [149,164] after transcription *in vitro*, although it was unclear whether long poly(A) tails are synthesized for all genes. A **third aim** of this study was investigation of the early steps of poly(A) tail metabolism and possible crosstalk of polyadenylation with splicing and RNA export. Computational analysis of splicing status was used here along with biochemical fractionation, and metabolic labeling experiments to track poly(A) tail length over time.

The **fourth aim** was identification of enzymes which mediate deadenylation of poly(A) tails in the nucleus by using different experimental strategies for RNA knockdown of PAN2-PAN3, CCR4-NOT and PARN deadenylases and investigating the impact on poly(A) tail length in different subcellular fractions.

3 Materials & Methods

3.1 Materials

3.1.1 Chemicals

Name	Vendor	Cat Number
Pladeinolide B	Biomol	Cay16538-100
Ribolock 40U/uL	Thermo Fisher	EO0381
Proteinase Inhibitor cOmplete mini EDTA free	Roche / Sigma	11836170001
Sucrose	Sigma	S0389
Isopropanol	Chemsolute	50295857
Glycoblue	Invitrogen	AM9516
Ethanol	Chemsolute	2286-1L
Page Ruler Plus Prestained	Thermo	26620
RNA XP Beads	Beckman Coulter	A63987
XP DNA Beads	Beckman Coulter	A63881
Dynabeads MyOne Streptavidin C1	invitrogen	65001
EZ-Link Biotin HPDP	Thermo	21341
4-Thiouridine	Chemgenes	RP-2304
Iodoacetamide	Sigma Aldrich	I6125-5G
Chloroform	Roth	Y015.1
Phenol-Chloroform-Isoamylalcohol	Roth	X985.1
Sodium Chloride (NaCl)	Roth	9265.1
Dithiotreitol (DTT)	Roth	6908.1
Guanosin-5-triphosphate (GTP)	Thermo	R0461
Inosin-5-triphosphate	Sigma	I0879-50MG
Desoxynucleotides (dNTP)	Thermo	R0191
Random Hexamer Primers 100 uM	Thermo Scientific	N8080127
Lipofectamine 2000	Thermo	11668030
Lipofectamin RNAiMAX	Thermo	13778100
Doxycycline	Sigma	D9891
Actinomycin D	Sigma	A1410
Dimethylsulfoxide (DMSO)	Sigma	D8418
Skim Milk Powder	Sigma	1.15363

3.1.2 Buffers and working solutions

Name	Components	Vendor	Cat No
Gibco PBS		Fisher Scientific	10010056
2x Lysis Buffer	NaCl 0.28 mM	Roth	S0389
	$MgCl_2$ 3 mM	Roth	KK36.1
	Tris HCl pH 7.5 20 mM	Roth	4855.1
	NP40 Substitute 1%	Sigma	74385
	Ribolock 1:200	Thermo Fisher	EO0381
	Proteinase Inhibitor 1:100 complete ETDA	Roche / Sigma	11836170001
Nuclear Buffer 1	Tris HCl pH 7.9 20 mM	Roth	4855.1
	NaCl 75 mM	Roth	S0389
	EDTA 0.5 mM	Invitrogen	15575020
	Dithiothreitol (DTT) 0.85 mM	Roth	6908.1
	Glycerol 50%	Roth	3783,1
	Ribolock 1:200	Thermo Fisher	EO0381
	Proteinase Inhibitor 1:100	Roche	11836170001
Nuclear Buffer 2	HEPES pH 7.9 20 mM	Roth	6763.1
	Dithiothreitol (DTT) 1 mM	Roth	6908.1
	$MgCl_2$ 7.5 mM	Roth	KK36.1
	EDTA 0.2 mM	Invitrogen	15575020
	NaCl 0.3 M	Roth	S0389
	Urea 1 M	Roth	7638.1
	NP-40 Substitute 1%	Sigma	74385
	Ribolock 1:200	Thermo Fisher	EO0381
	Proteinase Inhibitor 1:100	Roche	11836170001
2.5x RNA Biotin Labeling Buffer	EDTA 2.5 mM	Invitrogen	15575020
	Tris pH 7.4 25 mM	Roth	4855.1
MPG Buffer	Tris pH 7.6 100 mM	Roth	4855.1
	NaCl 1 M	Roth	9265.1
	EDTA 10 mM	Invitrogen	15575020
Methylene Blue Staining Solution	1% Methylene Blue	Roth	A514.1
	0.5 M Sodium Acetate	Roth	6773.1
10% Blocking Solution	10% SDS	Roth	1057.1
	1 mM EDTA	Invitrogen	15575020
	Gibco 1x PBS	Fisher Scientific	10010056

1% Blocking Solution	1% SDS	Roth	1057.1
	1 mM EDTA	Invitrogen	15575020
	Gibco 1x PBS	Fisher Scientific	10010056
0.1% Blocking Solution	0.1% SDS	Roth	1057.1
	1 mM EDTA	Invitrogen	15575020
	Gibco 1x PBS	Fisher Scientific	10010056
SDS Page Lower Gel 12%	12% Acrylamide/Bis	BioRad	#1610156
	0.375 M Tris pH 8.8	Roth	4855.1
	0.1% SDS	Roth	1057.1
	10% APS 200 uL	Serva	13375.01
	TEMED 7.5 uL	Roth	2367.1
SDS Page Upper Gel 4%	4% Acrylamide/Bis	BioRad	#1610156
	0.125 M Tris pH 6.8	Roth	4855.1
	0.15% SDS	Roth	1057.1
	10% APS 200 uL	Serva	13375.01
	7.5 uL TEMED 7.5 uL	Roth	2367.1
1x SDS PAGE Running Buffer	25 mM Tris	Roth	4855.1
	190 mM Glycine	Roth	3790.1
	0.1% SDS	Roth	1057.1
SDS Loading Buffer 5x	10 % SDS	Roth	1057.1
	500 mM DTT	Roth	6908.1
	50% Glycerol	Roth	3783,1
	250 mM Tris pH 6.8	Roth	4855.1
	0.5% Bromophenol Blue	Applichem	A2331
Trizol	38% Phenol	Roth	A980.1
	0.8 M Guanidine Thiocyanate	Roth	2628,4
	0.4 M Ammonium Thiocyanate	Roth	4477,4
	0.1 M Sodium Acetate	Roth	6773,2
	5 % Glycerol	Roth	3783,1
TBS-T Buffer	Tris 20 mM	Roth	4855.1
	NaCl 150 mM	Roth	9265.1
	0.1% Tween 20	Roth	9127.1
Na_3PO_4 Buffer	0.5 M Na_3PO_4	Roth	T107.1
Gibco DMEM High Glucose GlutaMax		Thermo Fisher	31966047
Optimem		Thermo	31985062
FBS / Tet Free FBS		PAN Biotech	P30-3602

Mild Stripping Buffer	Glycine 15% w/v	Roth	3790.1
	SDS 1% v/v	Roth	1057.1
	Tween-20 1%	Roth	9127.1
	pH 2.2 (Adjust HCl 37%)	Roth	4625.1
TE-TW buffer	10 mM Tris pH 8.0	Thermo	AM9855G
	1 mM EDTA	Thermo	15575020
	Tween-20 0.01%	Roth	9127.1

3.1.3 Kits & Enzymes

Name	**Ventor**	**Cat No**
TruSeq mRNA preparation kit	Illumina	RS-122-2102
USB poly(A) length assay kit	Thermo Fisher	764551KT
SMARTScribe Reverse Transcriptase kit	Clontech	639537
Advantage 2 DNA polymerase mix	Clontech	639201
RNA 6000 Pico Kit	Agilent Technologies	5067-1513
High Sensitivity NGS Fragment Analysis Kit	Advanced Analytical Technologies GmbH	DNF-474
Agilent DNA 12000 Kit	Agilent Technologies	cat 5067-1508
Sequel™ SMRT® Cell 1M v3 Tray	Pacific Biosciences	101-531-000
SMRTbell™ Template Prep Kit 1.0-SPv3	Pacific Biosciences	100-991-900
Sequel Sequencing Kit 3.0	Pacific Biosciences	101-597-800
Sequel Binding and Internal Ctrl Kit 3.0	Pacific Biosciences	101-626-600
Direct-zol RNA miniprep kit	Zymo	R2070
Western Blotting Kits	Biorad Trans Blot Turbo	1704157
ECL Select	GE Healthcare / Thermo Fisher	12644055
TurboDNA free Kit	Thermo	AM1907
Poly(A) Polymerase Yeast	Thermo	74225Z25KU
Dynabeads™ mRNA DIRECT™ Purification Kit	Thermo / Invitrogen	61012
neoLab Rotator mit Vortexer	neoLab	7-0045
Nanostring Probe Set A / B	Nanostring	
Nanostring Hybridization Mix	Nanostring	
Nanostring 72-plex Core Set	Nanostring	
SuperScript II Reverse Transcriptase	ThermoFisher Scientific	18064022
Blue S Green qPCR Kit	Biozym	331416S
Qubit RNA HS Assay Kit	Thermo Fisher	Q32852

3.1.4 Antibodies

Name	Ventor	Cat No
Pierce High Sensitivity Strep-HRP	Thermo Fisher	21130
GAPDH Monoclonal Anti GAPDH clone 71.1	Sigma	G8795-200UL
TDP-43 Polyclonal	Proteintech	10782-2-AP
BCAP31 (BAP31) Polyclonal	Proteintech	11200-1-AP
anti-mouse IgG HRP goat pAb to msIgG	abcam	ab97023
anti-rabbit IgG HRP goat anti rabbit IgG-h+I	Bethyl	A120-112P

3.1.5 Oligonucleotides

Name	Sequence	Comment
dC 3T UMI RT primer RT primer 1	GGTAATACGACTCACTATAGCGAGANNNNNNNNNCCCCCCCCCTTT	
dC 3T UMI RT primer RT primer 2	TGAGTCGGCAGAGAACTGGCGAANNNNNNNNNCCCCCCCCCTTT	
FLAM Seq Template switch oligo (isoTSO)	iCiGiCAAGCAGTGGTATCAACGCAGAGTACATrGrGrG	
PCR Primer 1	GGTAATACGACTCACTATAGCGAG	
PCR Primer 2	TGAGTCGGCAGAGAACTGGCGAA	
Universal RV primer (USB Poly(A) tail length kit)	proprietary	
PAT GAPDH	AAAAAGCCTAGGGAGCCGCACCTTG	
PAT RPL37A	CCAGTTCATCTTAAGAATGTCAACG	
PAT SCD	ACGATGACTGGGCAAGAAGA	
PAT BTF3	GAAGAAGCCTGGGAATCAAGTTTG	
TM4SF1_Fwd_PAT	ACGATGACTGGGCAAGAAGA	
ID3_Fwd_PAT	TGACTTTCTGTAACAATGCGATG	
qPCR primer PAN2 fwd	TGTGGATCCTGCCTTCTTGC	
qPCR primer PAN2 rev	CAGCCCTCAGAATCCCCAAA	
qPCR primer PAN3 fwd	CAGGATGCGAAGTGTAAATGAC	
qPCR primer PAN3 rev	ATCCTTCTGAAACTCCGGCC	
qPCR primer CNOT7 fwd	CCCATTGGAGAATTCAGGAGC	

qPCR primer CNOT7 rev	GGCATACATGTCCTCCGTCA	
qPCR primer CNOT8 fwd	CCAGCTGGGCCTTACATTCA	
qPCR primer CNOT8 rev	AGTGCAGTGTGTCAATCCCT	
qPCR primer PARN fwd	GAAGGAGGCTGACAGCAAACGG	
qPCR primer PARN rev	GCCAGCTTCCTCTTGACTAGGAC	
qPCR primer SCD fwd	TGCCCACCACAAGTTTTCAG	
qPCR primer SCD rev	CATCAGCAAGCCAGGTTTGT	
qPCR primer ID3 fwd	CTTGCTGGACGACATGAACC	
qPCR primer ID3 rev	GACAAGTTCCGGAGTGAGCT	
BLOCK-iT™ Alexa Fluor™ Red Fluorescent Control	Proprietary	Thermo Scientific 14750100
PAN2 siRNA	GACCUUGUUUGCUGGAUUA	Dharmacon
PAN3 siRNA	AAAACAAGGUUGCGAGUAA	Dharmacon
CNOT7 siRNA	CAGCUAGGACUGACAUUUA	Dharmacon
CNOT7 siRNA	UUUCGUAGUUCCAUAGAUU	Dharmacon
GAPDH Probe A	GCTCCTGGAAGATGGTGATGGGATTTCCATTGATGACAAGCTTCCCGTTCCCTCAAGA CCTAAGCGACAGCGTGACCTTGTTTCA	NanoString Panel
CNOT7 Probe A	AACTTCACAAATTCTTTGGCTATGATCTACAGTTGCCGCTGGCATAGTGACATCCTCT TCTTTTCTTGGTGTTGAGAAGATGCTC	NanoString Panel
TUBB Probe A	ACTGCTGACACCTCCCTTGAAGCTGAGATGGGAAATGGACATACTTAGAACACAATT CTGCGGGTTAGCAGGAAGGTTAGGGAAC	NanoString Panel
TUT7 Probe A	CCAGGACCTTGGGCCAAAACTGAGTTTTTGAGCTTATCATCTTTTGTTTGCTGTTGAG ATTATTGAGCTTCATCATGACCAGAAG	NanoString Panel
TUT1 Probe A	ATGGAAGAGCCAAAAGGGTGGACCACACAGCCAGGGAAGAACTCTGTGAACAAAGA CGCCTATCTTCCAGTTTGATCGGGAAACT	NanoString Panel
HPRT1 Probe A	TGAGCACACAGAGGGCTACAATGTGATGGCCTCCCATCTCCTTCATCACACGAACCT AACTCCTCGCTACATTCCTATTGTTTTC	NanoString Panel
TBP Probe A	GCACGAAGTGCAATGGTCTTTAGGTCAAGTTTACAACCAAGATTCACTGTCCAATTTG GTTTTACTCCCCTCGATTATGCGGAGT	NanoString Panel
TUT4 Probe A	GTGTCGGTGCACACACTCTGATGTGAACTCTCTGTTTCTAAATTTGACCTCTTTCGGGT TATATCTATCATTTACTTGACACCCT	NanoString Panel
FIP1L1 Probe A	CTCGAGACTCCAATCCCAGGAAGGCGACGGCAGCGGCGGCAAAGATGAAGCAACAG CCACTTTTTTTCCAAATTTTGCAAGAGCC	NanoString Panel
POLR1B Probe A	AGCCAGCAGCCTTTAACATCTCACCAAAGTATTCTAAGGCCGAGTTCTCCCACCGTGT GGACGGCAACTCAGAGATAACGCATAT	NanoString Panel
PABPN1 Probe A	CTCTTTGTCTGAGAACTCTATATACGCAAACCCTTTGGGATGGCCACTAACCTGGAGT TTATGTATTGCCAACGAGTTTGTCTTT	NanoString Panel
PAN3 Probe A	GAGGTGCAGTTGGAGGATAAATATGATAGTTTGGAAACACCATTCCAGTCCAGATAA GGTTGTTATTGTGGAGGATGTTACTACA	NanoString Panel
CNOT8 Probe A	AACTCTTTCATCCTAAAGAAAGCCATTCCTGTCAGCAGTGAGTCTGAGCCCTTCCTTC CTGTGTTCCAGCTACAAACTTAGAAAC	NanoString Panel
TENT4B Probe A	GCCTGAGAGGACTCCAAGGATACATCTTGCGACCCTACTCGCATAAAATTGGTTTTGC CTTTCAGCAATTCAACTT	NanoString Panel
RPL19 Probe A	AATCCTCATTCTCCTCATCCATGTGACCTTCTCTGGCATTCGGGCATTGGCTGGTCAA GACTTGCATGAGGACCCGCAAATTCCT	NanoString Panel

PAN2 Probe A	GGCCCCTCCAGGTATCCCTGAAGCAGTTGGTGACTTGTGCTAAGTCTTGACTTTCGTTGGGACGCTTGAAGCGCAAGTAGAAAAC	NanoString Panel
TENT4A Probe A	AATATTGCTAATGCGTAGAGATTAGTACAGGCCTACAGATCAGTTTTTATCCAGCAGACCTGCAATATCAAAGTTATAAGCGCGT	NanoString Panel
PABPC1 Probe A	CATCTCATCCACAGCTTTCTGTGCATCTTCATGCCTTTCAAAGCTTACAACCTGCCAATGCACTCGATCTTGTCATTTTTTTGCG	NanoString Panel
TENT2 Probe A	GGAACATTACATGGAGCTTGATGTACAAGGTGCAGCTGTATAGCAGGACTCAAACTGGAGAGAGAAGTGAAGACGATTTAACCCA	NanoString Panel
GAPDH Probe B	CGAAAGCCATGACCTCCGATCACTCCGCCAGCATCGCCCCACTTGATTTTGGAGGGATCTC	NanoString Panel
CNOT7 Probe B	CGAAAGCCATGACCTCCGATCACTCCGGATAACTTGACGAATTTTCTTCATCTCTTCATCCAAGTTGCAAGCCCA	NanoString Panel
TUBB Probe B	CGAAAGCCATGACCTCCGATCACTCACAGACTCCTCCAGAGTAGAGCTTGGAGGGAGATTGAAAGTGGAGATAAT	NanoString Panel
TUT7 Probe B	CGAAAGCCATGACCTCCGATCACTCTAAGAGTAAGTGGCTGTACCTTACAGGTATTTGCAGCTGAACTGGTAGCA	NanoString Panel
TUT1 Probe B	CGAAAGCCATGACCTCCGATCACTCCCAGATCCAAGAAGAGGTCAAGATCACAGCCATGGACATCGAAGCTATTT	NanoString Panel
HPRT1 Probe B	CGAAAGCCATGACCTCCGATCACTCCAGTGCTTTGATGTAATCCAGCAGGTCAGCAAAGAATTTATAGCCCCCCT	NanoString Panel
TBP Probe B	CGAAAGCCATGACCTCCGATCACTCTCCTCATGATTACCGCAGCAAACCGCTTGGGATTATATTCGGCGTTTCGG	NanoString Panel
TUT4 Probe B	CGAAAGCCATGACCTCCGATCACTCTCAGAAGCATCTTCTGTAGCTTTGCAGTTGCAAGAGGTAGCAGAT	NanoString Panel
FIP1L1 Probe B	CGAAAGCCATGACCTCCGATCACTCCCCCGAGGCGCGAGAAGGGCGCGAACCCGCCGACGAACGAACGAAGAAAG	NanoString Panel
POLR1B Probe B	CGAAAGCCATGACCTCCGATCACTCTTCTAGCCCACTGATGCCACTATATAACCTCTCGGTGCCATAGAAATTGT	NanoString Panel
PABPN1 Probe B	CGAAAGCCATGACCTCCGATCACTCTGCCTTCCTCTAAATAGGGACTCATCTAAGGCCAAGGAAGTCCTCACTGA	NanoString Panel
PAN3 Probe B	CGAAAGCCATGACCTCCGATCACTCAGCCATGAAGAAGGAAGGTGCGTTTGCTTTCGGTTGCATATAAGCAACGT	NanoString Panel
CNOT8 Probe B	CGAAAGCCATGACCTCCGATCACTCCATAGAGCCGCCCACAGTACTTGGCATCATCAATGCTGTCCTCAAAAAAC	NanoString Panel
TENT4B Probe B	CGAAAGCCATGACCTCCGATCACTCATGTGTTAGTGGTTTGGGTGCTTTGCATTTTCCCAACT	NanoString Panel
RPL19 Probe B	CGAAAGCCATGACCTCCGATCACTCTGGCGATCGATCTTCTTAGATTCACGGTATCTTCTGAGCAGCCGGCGCAA	NanoString Panel
PAN2 Probe B	CGAAAGCCATGACCTCCGATCACTCAATGGTATTGGCAACACTGGTTTGCTTGGGCCACCAACACTCTACTTGCT	NanoString Panel
TENT4A Probe B	CGAAAGCCATGACCTCCGATCACTCGACACGCACTGCCAGACCCGTAGTGGTTTCCTTAATGTGGGTTTACAGTC	NanoString Panel
PABPC1 Probe B	CGAAAGCCATGACCTCCGATCACTCTTCTGAGCTCGACCAACATAAATTTGTTTTCCATTGAGCTCCTTTCCGTT	NanoString Panel
TENT2 Probe B	CGAAAGCCATGACCTCCGATCACTCCCAGTAAGAGGTCCCCAAGGTTTGATTCATTCTTTGAGAGGTAAGGA	NanoString Panel

3.1.6 Plasmids

Name	Description	Comment
Cas13b	Psp-Cas13b (addgene #103862, ref. Cox et al., Science 2017)	
Luciferase	Transfection control	
NT	as13b crRNA direct repeat under a U6 promoter + Non targeting guide RNA	
PAN2_PS1	Cas13b crRNA direct repeat under a U6 promoter + PAN2 guide fwd CACCGCTTTAAGTAGGTAGACTTGAGAGTTGTTA PAN2 guide rev CAACTAACAACTCTCAAGTCTACCTACTTAAAGC	Designed by Ivano Legnini
PAN2_PS2	Cas13b crRNA direct repeat under a U6 promoter + PAN2 guide fwd CACCGCCCTAGCTTGGAATATTTGATGGTCACCT PAN2 guide rev CAACAGGTGACCATCAAATATTCCAAGCTAGGGC	Designed by Ivano Legnini
PAN3_PS1	Cas13b crRNA direct repeat under a U6 promoter + PAN3 guide fwd CACCGTATCCGGTTGGGAGGTGGCAGTGGTTCTA PAN3 guide rev CAACTAGAACCACTGCCACCTCCCAACCGGATAC	Designed by Ivano Legnini
PAN3_PS2	Cas13b crRNA direct repeat under a U6 promoter + PAN3 guide fwd CACCGCTCAGCAAATGCTTTAGTGGTAAATACTT PAN3 guide rev CAACAAGTATTTACCACTAAAGCATTTGCTGAGC	Designed by Ivano Legnini
CNOT7_PS1	Cas13b crRNA direct repeat under a U6 promoter + CNOT7 guide fwd CACCGAATAAAGAATGTACAAGGGAGACAAACCA CNOT7 guide rev CAACTGGTTTGTCTCCCTTGTACATTCTTTATTC	Designed by Ivano Legnini
CNOT7_PS2	Cas13b crRNA direct repeat under a U6 promoter + CNOT7 guide fwd CACCGATCTGAGATAGGAACGGTCATACTTAGTA CNOT7 guide rev CAACTACTAAGTATGACCGTTCCTATCTCAGATC	Designed by Ivano Legnini
CNOT8_PS1	Cas13b crRNA direct repeat under a U6 promoter + CNOT8 guide fwd CACCGAGGAATGGGGAAGACTTATTTACAAATTC CNOT8 guide rev CAACGAATTTGTAAATAAGTCTTCCCCATTCCTC	Designed by Ivano Legnini
CNOT8_PS2	Cas13b crRNA direct repeat under a U6 promoter + CNOT8 guide fwd CACCGGCAAACGAGAATCTGTAAGCAACTTTACC CNOT8 guide rev CAACGGTAAAGTTGCTTACAGATTCTCGTTTGCC	Designed by Ivano Legnini
CasRx plasmids	addgene #109049	
CasRx gRNAs PAN2_1	CasRx crRNA under a U6 promoter + PAN2 guide fwd CACCGCTTTAAGTAGGTAGACTTGAGAGTTGTTA PAN2 guide rev CAACTAACAACTCTCAAGTCTACCTACTTAAAGC	Designed by Ivano Legnini
CasRx gRNAs PAN2_2	CasRx crRNA under a U6 promoter + PAN2 guide fwd CACCGCCCTAGCTTGGAATATTTGATGGTCACCT PAN2 guide rev CAACAGGTGACCATCAAATATTCCAAGCTAGGGC	Designed by Ivano Legnini

3.1.7 Cell lines

Name	Description	Vendor
HeLa S3		Giuseppe Macino Lab; Sapienza University
HEK Flp-In 293 T-REx		Invitrogen R78007
HeLa S3 PARN shRNA 1	HeLa S3 cell lines transduced with Lenti Virus produced from EZ Plko TetON plasmid with shRNA sequences: (shRNA_PARN_1_f) ctagcCCGCACTGTATTTAACTTAATtactagtATTAAGTTAAATACAGTGCGGttttttg (shRNA_PARN_1_r) aattcaaaaaaCCGCACTGTATTTAACTTAATactagtaATTAAGTTAAATACAGTGCGGg	Generated by Ivano Legnini
HeLa S3 PARN shRNA 2	HeLa S3 cell lines transduced with Lenti Virus produced from EZ Plko TetON plasmid with shRNA sequences: (shRNA_PARN_2_f) ctagcCCTATGTATCTCCTAACACTTtactagtAAGTGTTAGGAGATACATAGGttttttg (shRNA_PARN_2_r) aattcaaaaaaCCTATGTATCTCCTAACACTTactagtaAAGTGTTAGGAGATACATAGGg	Generated by Ivano Legnini
HeLa S3 PARN shRNA 1+2	combined Lenti Virus s. above	Generated by Ivano Legnini
HeLa S3 PAN3 shRNA 2	HeLa S3 cell lines transduced with Lenti Virus produced from EZ Plko TetON plasmid with shRNA sequences: (PAN3_shRNA2_f) CTAGCCCCAAGATTACTCCACATATACTAGTTATGTGGAGTAATCTTGGGTTTTTTG (PAN3_shRNA2_r) AATTCAAAAAACCCAAGATTACTCCACATAACTAGTATATGTGGAGTAATCTTGGGG	Generated by Ivano Legnini
HeLa S3 CNOT7 shRNA 1	HeLa S3 cell lines transduced with Lenti Virus produced from EZ Plko TetON plasmid with shRNA sequences: (CNOT7_shRNA1_f) CTAGCCAGCTAGGACTGACATTTATACTAGTTAAATGTCAGTCCTAGCTGTTTTTTG (CNOT7_shRNA1_r) AATTCAAAAAACAGCTAGGACTGACATTTAACTAGTATAAATGTCAGTCCTAGCTGG	Generated by Ivano Legnini
HeLa S3 CNOT7 shRNA 1+2	Lenti Virus from construct above + EZ Plko TetON plasmid with shRNA sequences: (CNOT7_shRNA2_f) CTAGCGACTCTATAGAGCTACTAATACTAGTTTAGTAGCTCTATAGAGTCTTTTTTG (CNOT7_shRNA2_r) AATTCAAAAAAGACTCTATAGAGCTACTAAACTAGTATTAGTAGCTCTATAGAGTCG	Generated by Ivano Legnini

3.1.8 Datasets

Overview of FLAM-Seq datasets

Sample	Experiment	Description
NR_JA_Pb_72_4SU_PD_8_20_min_SN	4SU - Pulldown	4sU Pulldown 20 min label - supernatant
NR_JA_Pb_71_4SU_PD_8_15_min_SN	4SU - Pulldown	4sU Pulldown 15 min label - supernatant
NR_JA_Pb_70_4SU_PD_8_10_min_SN	4SU - Pulldown	4sU Pulldown 10 min label - supernatant
NR_JA_Pb_69_4SU_PD_8_0_min_SN	4SU - Pulldown	4sU Pulldown 0 min label (ctrl)- supernatant
NR_JA_Pb_68_4SU_PD_8_20_min_PD	4SU - Pulldown	4sU Pulldown 20 min label - pulldown
NR_JA_Pb_67_4SU_PD_8_15_min_PD	4SU - Pulldown	4sU Pulldown 15 min label - pulldown
NR_JA_Pb_66_4SU_PD_8_10_min_PD	4SU - Pulldown	4sU Pulldown 10 min label - pulldown
NR_JA_Pb_65_4SU_PD_8_0_min_PD	4SU - Pulldown	4sU Pulldown 0 min label (ctrl)- pulldown
NR_JA_Pb_78_4SU_PD_5_90_min_PD	4SU - Pulldown	4sU Pulldown 90 min label - pulldown
NR_JA_Pb_77_4SU_PD_5_45_min_PD	4SU - Pulldown	4sU Pulldown 45 min label - pulldown
GI_KD_7_Ctrl1_Cyto	PARN KD Ctrl	PARN shRNA 1+2 Cytoplasm Rep 1 Ctrl 5d
GI_KD_7_Ctrl2_Cyto	PARN KD Ctrl	PARN shRNA 1+2 Cytoplasm Rep 2 Ctrl 5d
GI_KD_7_Ctrl1_Nuc	PARN KD Ctrl	PARN shRNA 1+2 Nucleoplasm Rep 1 Ctrl 5d
GI_KD_7_Ctrl2_Nuc	PARN KD Ctrl	PARN shRNA 1+2 Nucleoplasm Rep 2 Ctrl 5d
GI_KD_7_Ctrl1_Chr	PARN KD Ctrl	PARN shRNA 1+2 Chromatin Rep 1 Ctrl 5d
GI_KD_7_Ctrl2_Chr	PARN KD Ctrl	PARN shRNA 1+2 Chromatin Rep 2 Ctrl 5d
GI_KD_7_Dox1_Cyto	PARN KD Dox	PARN shRNA 1+2 Cytoplasm Rep 1 Dox 5d
GI_KD_7_Dox2_Cyto	PARN KD Dox	PARN shRNA 1+2 Cytoplasm Rep 2 Dox 5d
GI_KD_7_Dox1_Nuc	PARN KD Dox	PARN shRNA 1+2 Nucleoplasm Rep 1 Dox 5d
GI_KD_7_Dox2_Nuc	PARN KD Dox	PARN shRNA 1+2 Nucleoplasm Rep 2 Dox 5d
GI_KD_7_Dox1_Chr	PARN KD Dox	PARN shRNA 1+2 Chromatin Rep 1 Dox 5d
GI_KD_7_Dox2_Chr	PARN KD Dox	PARN shRNA 1+2 Chromatin Rep 2 Dox 5d
gi_kd_9_CNOT7_Ctrl1_Chr	CNOT7 Ctrl	CNOT7 shRNA 1+2 Chromatin Rep 1 Ctrl 5d
gi_kd_9_CNOT7_Ctrl1_Cyto	CNOT7 Ctrl	CNOT7 shRNA 1+2 Cytoplasm Rep 1 Ctrl 5d
gi_kd_9_CNOT7_Ctrl1_Nuc	CNOT7 Ctrl	CNOT7 shRNA 1+2 Nucleoplasm Rep 1 Ctrl 5d
gi_kd_9_CNOT7_Ctrl2_Chr	CNOT7 Ctrl	CNOT7 shRNA 1+2 Chromatin Rep 2 Ctrl 5d
gi_kd_9_CNOT7_Ctrl2_Cyto	CNOT7 Ctrl	CNOT7 shRNA 1+2 Cytoplasm Rep 2 Ctrl 5d
gi_kd_9_CNOT7_Ctrl2_Nuc	CNOT7 Ctrl	CNOT7 shRNA 1+2 Nucleoplasm Rep 2 Ctrl 5d
gi_kd_9_CNOT7_Dox1_Chr	CNOT7 Dox	CNOT7 shRNA 1+2 Chromatin Rep 1 Dox 5d
gi_kd_9_CNOT7_Dox1_Cyto	CNOT7 Dox	CNOT7 shRNA 1+2 Cytoplasm Rep 1 Dox 5d
gi_kd_9_CNOT7_Dox1_Nuc	CNOT7 Dox	CNOT7 shRNA 1+2 Nucleoplasm Rep 1 Dox 5d

gi_kd_9_CNOT7_Dox2_Chr	CNOT7 Dox	CNOT7 shRNA 1+2 Chromatin Rep 2 Dox 5d
gi_kd_9_CNOT7_Dox2_Cyto	CNOT7 Dox	CNOT7 shRNA 1+2 Cytoplasm Rep 2 Dox 5d
gi_kd_9_CNOT7_Dox2_Nuc	CNOT7 Dox	CNOT7 shRNA 1+2 Nucleoplasm Rep 2 Dox 5d
kd_9_PAN3_Dox2_Nuc	PAN3 Dox	PAN3 shRNA 2 Nucleoplasm Rep 2 Dox 5d
kd_9_PAN3_Dox2_Cyto	PAN3 Dox	PAN3 shRNA 2 Cytoplasm Rep 2 Dox 5d
kd_9_PAN3_Dox1_Nuc	PAN3 Dox	PAN3 shRNA 2 Nucleoplasm Rep 1 Dox 5d
kd_9_PAN3_Dox1_Cyto	PAN3 Dox	PAN3 shRNA 2 Cytoplasm Rep 1 Dox 5d
kd_9_PAN3_Dox1_Chr	PAN3 Dox	PAN3 shRNA 2 Chromatin Rep 1 Dox 5d
kd_9_PAN3_Ctrl2_Nuc	PAN3 Ctrl	PAN3 shRNA 2 Nucleoplasm Rep 2 Ctrl 5d
kd_9_PAN3_Ctrl2_Cyto	PAN3 Ctrl	PAN3 shRNA 2 Cytoplasm Rep 2 Ctrl 5d
kd_9_PAN3_Ctrl2_Chr	PAN3 Ctrl	PAN3 shRNA 2 Chromatin Rep 2 Ctrl 5d
kd_9_PAN3_Ctrl1_Nuc	PAN3 Ctrl	PAN3 shRNA 2 Nucleoplasm Rep 1 Ctrl 5d
kd_9_PAN3_Ctrl1_Cyto	PAN3 Ctrl	PAN3 shRNA 2 Cytoplasm Rep 1 Ctrl 5d
kd_9_PAN3_Ctrl1_Chr	PAN3 Ctrl	PAN3 shRNA 2 Chromatin Rep 1 Ctrl 5d
kd_15_nuc_dox_2	PAN3 Dox	PAN3 shRNA 2 Nucleoplasm Rep 2 Dox 3d
kd_15_nuc_dox_1	PAN3 Dox	PAN3 shRNA 2 Nucleoplasm Rep 1 Dox 3d
kd_15_nuc_ctrl_2	PAN3 Ctrl	PAN3 shRNA 2 Nucleoplasm Rep 2 Ctrl 3d
kd_15_nuc_ctrl_1	PAN3 Ctrl	PAN3 shRNA 2 Nucleoplasm Rep 1 Ctrl 3d
kd_15_cyto_dox_2	PAN3 Dox	PAN3 shRNA 2 Cytoplasm Rep 2 Dox 3d
kd_15_cyto_dox_1	PAN3 Dox	PAN3 shRNA 2 Cytoplasm Rep 1 Dox 3d
kd_15_cyto_ctrl_2	PAN3 Ctrl	PAN3 shRNA 2 Cytoplasm Rep 2 Ctrl 3d
kd_15_cyto_ctrl_1	PAN3 Ctrl	PAN3 shRNA 2 Cytoplasm Rep 1 Ctrl 3d
kd_15_chr_dox_2	PAN3 Dox	PAN3 shRNA 2 Chromatin Rep 2 Dox 3d
kd_15_chr_dox_1	PAN3 Dox	PAN3 shRNA 2 Chromatin Rep 1 Dox 3d
kd_15_chr_ctrl_2	PAN3 Ctrl	PAN3 shRNA 2 Chromatin Rep 2 Ctrl 3d
kd_15_chr_ctrl_1	PAN3 Ctrl	PAN3 shRNA 2 Chromatin Rep 1 Ctrl 3d
NR_JA_Pb_017_HeLa_F_chr_2	Hela S3 Fractions 1	HeLa S3 Chromatin Rep 2
NR_JA_Pb_016_HeLa_F_cyto_2	Hela S3 Fractions 1	HeLa S3 Cytoplasm Rep 2
NR_JA_Pb_015_HeLa_F_cyto_1	Hela S3 Fractions 1	HeLa S3 Cytoplasm Rep 1
NR_JA_Pb_014_HeLa_F_chr_1	Hela S3 Fractions 1	HeLa S3 Chromatin Rep 1
NR_JA_Pb_013_HeLa_F_nuc_2	Hela S3 Fractions 1	HeLa S3 Nucleoplasm Rep 2
NR_JA_Pb_011_HeLa_F_nuc_1	Hela S3 Fractions 1	HeLa S3 Nucleoplasm Rep 1
MB_nucleiA	Mouse Brain Fractions	Mouse Brain Nuclei Rep 1
MB_nucleiB	Mouse Brain Fractions	Mouse Brain Nuclei Rep 2
MB_cytoA	Mouse Brain Fractions	Mouse Brain Cytoplasm Rep 1
MB_cytoB	Mouse Brain Fractions	Mouse Brain Cytoplasm Rep 2

F2526_Nuc_1	Hela S3 Fractions 2	HeLa S3 Nucleoplasm Rep 1
F2526_Cyto_1	Hela S3 Fractions 2	HeLa S3 Cytoplasm Rep 1
F2526_Cyto_2	Hela S3 Fractions 2	HeLa S3 Cytoplasm Rep 2
F2526_Nuc_2	Hela S3 Fractions 2	HeLa S3 Nucleoplasm Rep 2
F2526_Chr_2	Hela S3 Fractions 2	HeLa S3 Chromatin Rep 2
F2526_Chr_1	Hela S3 Fractions 2	HeLa S3 Chromatin Rep 1
NR_JA_079_TRANS_2_Ctrl1	Transcription Inhibition	Transcription Inhibition Ctrl Rep 1
NR_JA_080_TRANS_2_Ctrl2	Transcription Inhibition	Transcription Inhibition Ctrl Rep 2
NR_JA_081_TRANS_2_2h_1	Transcription Inhibition	Transcription Inhibition 2h Rep 1
NR_JA_082_TRANS_2_2h_2	Transcription Inhibition	Transcription Inhibition 2h Rep 2
NR_JA_083_TRANS_2_6h_1	Transcription Inhibition	Transcription Inhibition 6h Rep 1
NR_JA_084_TRANS_2_6h_2	Transcription Inhibition	Transcription Inhibition 6h Rep 2
NR_JA_085_TRANS_2_12h_1	Transcription Inhibition	Transcription Inhibition 12h Rep 1
NR_JA_086_TRANS_2_12h_2	Transcription Inhibition	Transcription Inhibition 12h Rep 2
GI_FRAC_23_Ctrl1_merge	Splicing Inhibition	Splicing Inhibiton Nuclei Ctrl Rep 1
GI_FRAC_23_Ctrl2_merge	Splicing Inhibition	Splicing Inhibiton Nuclei Ctrl Rep 2
FRAC_23_PlaB1_merge	Splicing Inhibition	Splicing Inhibiton Nuclei PlaB Rep 1
FRAC_23_PlaB2_merge	Splicing Inhibition	Splicing Inhibiton Nuclei PlaB Rep 2
NR_IL_005_0min_R1	SLAM-Seq	SLAM-Seq / FLAM-Seq 4 sU label 0 min Rep 1
NR_IL_006_90min_R1	SLAM-Seq	SLAM-Seq / FLAM-Seq 4 sU label 90 min Rep 1
NR_IL_007_0min_R2	SLAM-Seq	SLAM-Seq / FLAM-Seq 4 sU label 0 min Rep 2
NR_IL_008_90min_R2	SLAM-Seq	SLAM-Seq / FLAM-Seq 4 sU label 90 min Rep 2
NR_IL_009_180min_R1	SLAM-Seq	SLAM-Seq / FLAM-Seq 4 sU label 180 min Rep 1
NR_IL_010_180min_R2	SLAM-Seq	SLAM-Seq / FLAM-Seq 4 sU label 180 min Rep 2
NR_IL_Pb-017_HeLa_Rep_2_merge_clean	HeLa S3	FLAM-Seq bulk HeLa S3 Rep 1
NR_IL_Pb-014_HeLa_Rep_1_merge_clean	HeLa S3	FLAM-Seq bulk HeLa S3 Rep 2
NR_IL_Pb_021_celegans_L4_rep2_merge_clean	*C. elegans* L4	FLAM-Seq bulk *C. elegans* L4 Rep 1
NR_IL_Pb_019_celegans_L4_rep1_merge_clean	*C. elegans* L4	FLAM-Seq bulk *C. elegans* L4 Rep 2

NR_IL_Pb_020_celegans_eggplaying_rep2_merge_clean	*C. elegans* Adult	FLAM-Seq bulk *C. elegans* adult Rep 1
NR_IL_Pb_018_celegans_eggplaying_rep1_merge_clean	*C. elegans* Adult	FLAM-Seq bulk *C. elegans* adult Rep 2
NR_IL_Pb-025-organoid_d30_rep2_merge_clean	Organoids	FLAM-Seq bulk Organoids Rep 1
NR_IL_Pb-024-organoid_d30_rep1_merge_clean	Organoids	FLAM-Seq bulk Organoids Rep 2
NR_IL_Pb-023-iPS_rep2_merge_clean	iPSC	FLAM-Seq bulk iPSC Rep 1
NR_IL_Pb-022-iPS_rep1_merge_clean	iPSC	FLAM-Seq bulk iPSC Rep 2
Dataset ID	**GEO Accession / Source**	**Sample**
SRR8268943	GSM3498219	4-thiouridine (4sU), 500 uM, 8 minutes K562
SRR8268944	GSM3498220	4-thiouridine (4sU), 500 uM, 8 minutes K562
SRR8268945	GSM3498221	4-thiouridine (4sU), 500 uM, 8 minutes K562
SRR10097604	GSM4073917	4-thiouridine (4sU), 500 uM, 8 minutes K562
SRR10097605	GSM4073918	4-thiouridine (4sU), 500 uM, 8 minutes K562
SRR10097603	GSM4073916	4-thiouridine (4sU), 500 uM, 8 minutes K562
Nanopolish poly(A) length estimate from Nanopore data	Dr. Karine Choquet / Stirling Churchman Lab	
Human gene annotation GTF	Gencode	http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_38/gencode.v38.annotation.gtf.gz
Human genome hg38	Gencode	http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_38/GRCh38.primary_assembly.genome.fa.gz
C. elegans genome WB235	NCBI	https://www.ncbi.nlm.nih.gov/assembly/GCF_000002985.6/
C. elegans genome WB235 annotation GTF	ensembl	http://ftp.ensembl.org/pub/release-82/gtf/caenorhabditis_elegans/
Mouse gene annotation GTF	ensembl	http://ftp.ensembl.org/pub/release-101/gtf/mus_musculus/
Mouse genome Grcm38	ensembl	http://ftp.ensembl.org/pub/release-101/fasta/mus_musculus/dna/
PacBio UHRR subreads	PacBio	https://downloads.pacbcloud.com/public/dataset/RC0_1cell_2017/m54086_170204_081430.subreads.bam)
Hs CAGE	FANTOM5	https://fantom.gsc.riken.jp/5/datafiles/latest/extra/CAGE_peaks/hg19.cage_peak_phase1and2combined_ann.txt.gz
C. elegans SAGE Saito et al. 2013	Saito et al	https://wormtss.utgenome.org/browser/download.jsp
PAL-Seq HeLa processed poly(A) tag statistics	Subtelny et al. GSE52809	https://ftp.ncbi.nlm.nih.gov/geo/series/GSE52nnn/GSE52809/suppl/GSE52809_HeLa_total.txt.gz;
TAIL-Seq HeLa cell line poly(A) statistics	Chang et al. Supplement	https://ars.els-cdn.com/content/image/1-s2.0-S109727651400121X-mmc2.xlsx
HeLa Half-life measurements	Tani et al	https://genome.cshlp.org/content/suppl/2012/02/14/gr.130559.111.DC1/Tani_Supp_Tables_revised2.xls
HeLa half-translation rates measurements	Subtelny et al. GSE52809	https://ftp.ncbi.nlm.nih.gov/geo/series/GSE52nnn/GSE52809/suppl/GSE52809_HeLa_total.txt.gz
HEK Half-life rates measurements	Schueler et al 2014 GSE49831	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE49831

List of ribosomal protein genes		RPSA,RPS2, RPS3, RPS3A, RPS4X, RPS4Y, RPS5c RPS6, RPS7, RPS8, RPS9, RPS10, RPS11, RPS12, RPS13, RPS14, RPS15, RPS15A, RPS16, RPS17, RPS18, RPS19, RPS20, RPS21, RPS23, RPS24, RPS25, RPS26,RPS27, RPS27A, RPS28, RPS29 RPS30, RPL3, RPL4, RPL5, RPL6, RPL7c, RPL7A RPL8, RPL9, RPL10, RPL10A, RPL12, RPL13A, RPL14, RPL15, RPL17, RPL18, RPL18A, RPL19, RPL21, RPL22, RPL23, RPL23A, RPL24, RPL26, RPL27, RPL27A, RPL30, RPL31, RPL32, RPL34, RPL35, RPL36, RPL36A, RPL37, RPL39, RPL40, RPL41, RPP0, RPP1, RPP2
List of immediate early genes		CTGF, NR4A2, CYR61, DUSP1, FOSB, FOS, NR4A1, IL6, NR4A3, EGR1, ZFP36, EGR3, JUNB, ATF3, DSCR1, AJ420542, GRO3, BHLHB2, DUSP5, GEM, SLC2A3, NFKBIA, PLAU, IER3, SGK, AL117595, COPEB, GADD45B, TIEG, MAIL, FLG, MCL1, GBP1 JUN, KIAA0469, TNFAIP3, CEBPD, LOC57018, DTR, C8FW, LDLR, TSC22, F3, SCYA2, DUSP6, SRF, AXUD1, PMAIP1, ZFP36L2

3.1.9 Devices

Name	Vendor	Cat
1.5 mL LoBinding Tubes	Eppendorf	0030108051
1.5 mL / 2 mL tubes	Eppendorf	0030120086 / 0030120094
Cell scraper	Sarstedt	SARS83.3951
6-well cell culture dishes	Sarstedt	83.3920.005
10 cm cell culture dishes	Sarstedt	83,3902
15 cm cell culture dishes	Sarstedt	83,3903
MicroAmp™ Optical 96-Well Reaction Plate	Thermo Fisher	N8010560
Dynamag-2 magnet	Thermo Fisher	12321D
254 nm crosslinker	UVP	
Hyperbond N+ membrane Amersham	GE Healthcare	GERPN203B
Dotblot device	not available	
Fusion FX Imager	Vilber	
TransBlot Turbo Western Blot	Biorad	
Fragment Analyzer	Advanced Analytical Technologies GmbH	
Bioanalyzer 21	Agilent	
Cell culture hood		
SurPhob Pipet Tips 10 µL, 200 µL, 1000 µL	Biozym	VT0270X, VT0210, VT0240
Vortex Genie 2	Scientific Industries	
Nanodrop 1000	Thermo Fisher	
Water machine		
Qubit 3 Fluorometer	Invitrogen / Thermo Fisher	Q33216
Master Cycler X50X Thermocycler	Eppendorf	

PegStar 2X Universal	PeqLab	
Whatman paper		
nCounter analysis system	Nanostring	
Step One Plus	Applied Biosystems	
Centrifuge 5415 R	Eppendorf	
Thermo Mixer Compact	Eppendorf	
Eclipse Ti2 microscope	Nikon	

3.1.10 Software / Packages

Name	**Version**
Rstudio	3.6.
Fragment Analyzer Software ProSize	2.0.
StepOne qPCR software	2.3.
Nanostring nCounter	4.0.
bedtools	2.27
STAR	2.5.4b
featureCounts	v1.6.0
python Anaconda, Inc.	3.6.7.
regex	2018.2.21
pysam 0.14	0.14
pandas 0.23.4	0.23.4
yaml 0.1.7	0.1.7
pybedtools	0.7.10
matplotlib	3.0.0
BioMart	2.42.1
peaktutils	1.3.0.
IGV Genome Browser	2.8.3.
minimap2	2.16-r922
samtools	1.9
Fiji / ImageJ	1.53c
SMRT Link browser software	5.0
topGO	2.38.1

3.2 Experimental Methods

3.2.1 RNA extraction from cells and tissues

RNA was extracted using Trizol reagent according to Chomczynski and Sacci [345].

Trizol was added to cells grown on cell culture dishes. Typical volumes were 2.7 mL Trizol per 10 cm dish or 900 µL per well of a 6-well culture dish. Cells were detached from cell culture dishes using a cell scraper for improved lysis. Lysed cells in Trizol were collected in a 1.5 mL Eppendorf tube, vortexed and incubated for 3 min at RT. 100 µL chloroform was added, vortexed and incubated for 3 min at RT. The lysate was centrifuged for 15 min at 16,000 g and 4°C. The aqueous phase was transferred to a fresh 1.5 mL tube, 0.5 µL Glycoblue and 1 vol isopropanol were added. The sample was typically incubated for at least 30 min at -20°C to enhance precipitation before centrifuging for 20 min at 16,000 g and 4°C to separate precipitated RNA from supernatant. The supernatant was discarded, and the pellet was washed with 80% EtOH and then centrifuged for 5 min at 7,500 g and 4°C. The supernatant was carefully removed completely, and RNA pellets were dried for 2-4 min at RT. RNA pellets were typically resuspended in 12-20 µL DNAse/RNAse-free H_2O. After extraction of RNA from tissues, the TurboDNA-free kit was used to remove genomic DNA leftovers. 1/9 volume 10x DNase buffer was added along with 2 µL DNase and samples were incubated for 20 min at 37°C. Reactions were quenched by adding 2 µL DNase inactivation reagent, incubating for 5 min at RT and collecting the supernatant. RNA concentration was quantified using NanoDrop 1000 or Qubit according with RNA HS Assay kit.

3.2.2 RNA purification by phenol-chloroform-isoamylalcohol (PCI) extraction

1 vol of PCI (phenol-chloroform-isoamylalcohol) was added to RNA in solution in a 1.5 mL tube. The sample tube was vortexed and then incubated for 3 min at RT. The tube was next spun for 10 min at 16,000 g. The supernatant was transferred to a fresh 1.5 mL tube and 1/10 vol 5 M NaCl, 0.5 µL Glycoblue and 1.1 vol isopropanol was added. The RNA was precipitated for 20-30 min at -20°C, then centrifµged for 20 min at 16,000 g. The supernatant was discarded, and the pellet was washed with 80% EtOH and then centrifuged for 5 min at 7,500 g at 4°C. The supernatant was carefully removed completely, and RNA pellets were dried for 2-4 min at RT. RNA pellets were typically resuspended in 12-20 µL DNAse/RNAse-free H_2O. RNA concentration was quantified using NanoDrop 1000 or Qubit according with RNA HS Assay kit.

3.2.3 DNA / RNA purification using Ampure XP / RNAClean XP beads

Ampure XP (DNA cleanup) and RNAClean XP (RNA cleanup) beads provide an alternative method for purification of nucleic acids from solution with a minimum size cutoff of around 100 bp. Bead to sample volume ratio can be adjusted to select nucleic acids by minimum size and ratios are specified for each cleanup step [346]. Typically, a 1.8x ratio of beads to sample volume was added, mixed by pipetting up and down 6 times and incubated for 5 min at RT. The sample was placed on a magnetic rack and incubated for 5 min to separate beads from supernatant. The supernatant was discarded, and the bead pellet was washed with 200 µL 80% EtOH. EtOH was removed and washing was repeated once. Bead pellets were dried at RT until no EtOH leftovers were visible. Pellets were resuspended in H_2O (typical volumes 15-50 µL) to elute DNA/RNA and incubated for 5 min at RT before placing samples back on a magnetic rack for 5 min. Supernatants were collected as purified DNA/RNA fraction.

3.2.4 Quantification of nucleic acids

Nucleic acids were quantified using the NanoDrop 1000 or Qubit 3 assay.

The NanoDrop system quantifies RNA, DNA or proteins by UV spectrophotometry. Typically, 1 µL of RNA, DNA or protein sample was loaded on the pedestal and the absorption spectrum was measured. The 260 nm absorption was converted into mass concentration [ng/µL] using default standard curves depending on sample type (for nucleic acids: ssDNA, RNA, dsDNA). The sample buffer solution was used to correct for background absorption. RNA quality was monitored by characteristic A260:A280 and A260:A230 absorption ratios which specify protein or phenol contamination.

The Qubit system has increased sensitivity compared to UV spectrophotometry through specific intercalation of dye molecules into RNA or DNA, which decreases background signal and enhances the detection limit to 10 pg/µL [347]. Typically, 1 µL of RNA or DNA samples were mixed with Qubit buffers and reagent and processed according to the manufacturer's instruction.

3.2.5 Poly(A) tail length assay (PAT Assay)

Poly(A) tail length (PAT) assays were used as a capillary electrophoresis-based method to measure poly(A) tail length profiles for individual genes. PAT assays were performed as described in Bazzini et al. [296], but omitting FAM-labeled primers for electrophoretic analysis.

For GI-tailing, 500 ng extracted total RNA were diluted in a total volume of 6.75 µL H_2O. For each sample 4.75 µL tailing mix (2.5 µL 5x Yeast Poly(A) Polymerase Buffer, 1 µL Ribolock, 1.25 µL 5 mM ITP/GTP mix) was added and mixed. 1 µL Yeast Poly(A) Polymerase was added last and the reaction was incubated for 60 min at 37°C.

Optionally, GI-tailing was performed using reagents from USB poly(A) tail length assay to test performance differences between GI-tailing conditions. For this approach, 500 ng total RNA was mixed with 4 µL 5x Tail Buffer mix and 2 µL 10x Tail Enzyme mix. Tailing reactions were incubated for 60 min at 37°C. 2 µL Stop buffer were added to terminate the tailing reactions. Samples were directly used for reverse transcription as described below.

GI-tailed RNA was adjusted to 200 µL final volume with H_2O and purified by PCI extraction. For reverse transcription 2 µL 5x RT Buffer and 1 µL 10x RT enzyme (USB length assay kit) were added and incubated for 1 h at 37°C. For PCR amplification 7.5 µL H_2O, 2.5 µL 5x PCR Mix (USB length assay kit), 0.5 µL 10 mM Universal RV Primer (USB length assay kit), 0.5 µL 10 mM gene specific PAT assay primer was mixed with 1 µL GI-tailed cDNA and 0.5 µL Taq Hot Start primer was added. For amplification of SCD and BTF3 genes 2 µL cDNA were used as input for PCR, and H_2O volume was reduced accordingly. Amplification was performed in a thermocycler with program settings: 2 min 94°C; [94°C 10 sec; 58°C 10 sec; 72°C 30 sec] for 32 cycles; 72°C 5 min.

2 µL PCR amplicons were loaded on a Fragment Analyzer using the High Sensitivity NGS Fragment Analysis Kit according to the manufacturer's instructions. Electropherogram data were exported as .csv files from Fragment Analyzer ProSize software and analyzed as described below (s. 3.3.1).

3.2.6 Gene expression quantification by quantitative real-time PCR (RT-qPCR)

RNA expression levels were quantified by quantitative real-time PCR (qPCR) of reverse transcribed total RNA. cDNA from each biological sample was measured in triplicates. Each 500 ng of total RNA per sample, adjusted to 11 µL volume in H_2O, was mixed with 1 µL 10 mM dNTPs and 0.5 µL 500 ng/µL random hexamer primers. Samples were incubated for 5 min at 65°C. 4 µL 5x First Strand Buffer (Superscipt II Kit), 2 µL 0.1 M DTT and 1 µL Ribolock were added, and samples were incubated for 2 min at 25°C before adding 0.5 µL Superscript II reverse transcriptase. Reverse transcription reactions were incubated for 10 min at 25°C, 50 min at 42°C and inactivated 15 min at 70°C.

For quantitative real-time PCR, cDNA was diluted 1:10 in H_2O. 3.75 µL diluted cDNA were mixed with 7.5 µL SYBR master mix. 3.75 µL 1 µM primer mix containing qPCR forward and reverse primers were added. 96 well plates containing reaction qPCR mixes were placed in a StepOne thermocycler and incubated using the following program: 20 sec 95°C; [95°C 10 sec; 60°C 20 sec] for 40 cycles. A melt curve analysis was added for testing specificity of amplification.

Raw data (Ct values) were exported as .xlsx files and analyzed as described below (s. 3.3.2).

3.2.7 Gene expression quantification by Nanostring assay

Nanostring enables multiplexed, amplification free quantification of RNA molecules by hybridization of label DNA probes. Samples were prepared according to the manufacturer's instructions using Nanostring reagents. Nanostring Probe Set A stock (5 nM) and Probe Set B stock (25 nM) were diluted 1:30 in 0.1% TE-TW buffer. 130 µL hybridization mix was added to 65 µL 72-plex Core Set, mixed and briefly centrifuged. 15 µL was pipetted into each tube. 12 µL 30x Probe Mix A and 30x Probe Mix B were added, as wells as 144 µL H_2O. 29 µL per tube were dispensed and each 1 µL 100 ng/µL RNA from each sample was added. Probes were hybridized for 18 h at 67°C in a thermocycler and then cooled. Probes were then loaded to a Nanostring cartridge and quantified. Data was analyzed using Nanostring nCounter 4.0 software (s. 3.3.3).

3.2.8 Full-length mRNA and poly(A) tail sequencing (FLAM-Seq)

A detailed protocol for Full-Length mRNA and Poly(A) sequencing (FLAM-Seq) can also be found at protocol exchange: https://protocolexchange.researchsquare.com/article/pex-398/v1 DOI: 10.21203/rs.2.10045/v1

2-10 µg total RNA were used as input for poly(A) selection using reagents from Illumina TruSeq mRNA preparation kit. Total RNA volume was adjusted to 50 µL using RNA purification beads. 50 µL poly(A) selection beads were added and mixed by pipetting up and down 6 times. The samples were incubated for 5 min at 65°C on a thermo block, 5 min on ice and 5 min at RT. Samples were placed in a magnetic rack to separate beads from solution for 5 min. The supernatant was discarded, and samples were removed from the magnetic rack. Beads were resuspended and washed in 200 µL beads washing buffer by pipetting up and down 6 times. Samples were placed back in a magnetic rack for 5 min and supernatant was removed. The sample was removed form the magnetic rack and 50 µL elution buffer was added. Samples

were incubated for 2 min at 80°C in a thermo block and incubated for 5 min at RT. 50 µL bead binding buffer was added and mixed by pipetting up and down 6 times and then incubated for another 5 min at RT. Samples were placed back in a magnetic rack and incubated for 5 min. The supernatant was discarded, samples were removed from the rack and beads were resuspended in 200 µL bead washing buffer by pipetting up and down 6 times. Samples were placed on a magnetic rack for 5 min and the supernatant was removed. Samples were removed from the rack and beads were resuspended in 16 µL H_2O for elution. Samples were incubated for 2 min at 70°C, 1 min on ice and eluted, poly(A) selected RNA was separated from beads by incubation for 5 min on a magnetic rack. 16 µL of the supernatants were collected in fresh PCR tubes.

Poly(A) selection was optionally performed using the Dynabeads mRNA Direct kit. Beads were here resuspended before and 50 µL beads were transferred to a fresh DNA LoBinding 1.5 mL tube and incubated for 30 sec on a magnetic rack. Supernatant was removed and beads were resuspended in 100 µL lysis/binding buffer. Samples were placed back in magnetic rack, incubated for 30 sec before supernatants were removed. Beads were again resuspended in 100 µL lysis/binding buffer and total RNA in 50 µL H_2O was added to beads. Samples were incubated for 5 min on a rotator and 2 min on a magnetic rack. The supernatant was then removed and beads were washed with 150 µL washing buffer A, placed on a magnetic rack for 30 sec. The supernatant was again discarded, and washing was repeated once with 150 µL washing buffer A and once with 150 µL washing buffer B. After removal of the last wash buffer, beads were resuspended in 20 µL elution buffer and incubated for 2 min at 70°C in a thermo block, then 30 sec on a magnetic rack. The eluate was collected in a fresh LoBinding tube and 80 µL lysis buffer was added. Beads were washed twice with 150 µL washing buffer B and 100 µL RNA in lysis buffer was added back to beads, mixed and incubated for 5 min with rotation. Washing steps were performed as above and poly(A) selected RNA was eluted in 16 µL H_2O and transferred to PCR tubes.

For GI-tailing of polyadenylated RNA, the reagents from the USB length assay kit were used. 14 µL poly(A) selected RNA were mixed with 4 µL 5x tail buffer mix. After this 2 µL 10x tail enzyme mix were added. Samples were incubated for 1 h at 37°C in a thermocycler. 1.5 µL stop solution was added immediately after completion and the tailing reaction and samples were incubated for 2 min on ice. GI-tailed RNA was purified using a 1.8x ratio of RNAClean XP beads as described above (s. 3.2.3) and RNA was eluted in 17 µL H_2O.

Reverse transcription reactions were performed using reagents from the SMARTScribe Reverse Transcriptase kit. First, the reaction mix was prepared consisting of 8 µL 5x First Strand buffer, 1.5 µL 20 mM DTT, 4 µL 10 mM dNTPs, 2 µL Ribolock RNase inhibitor, 2 µL 12 µM isoTSO primer, 2.5 µL H_2O and 2 µL SMARTScribe Reverse Transcriptase per sample. 16 µL purified GI-tailed RNA was mixed in a PCR tube with 2 µL 10 µM dC 3T UMI RT Primer 1 or 2. RNA-primer samples were placed in a thermocycler and the following program was started: 72°C for 3 min, 42°C for 1 h, 70°C for 10 min then hold at 4°C. After 3 min incubation at 72°C, 22 µL of the reaction mix was added. After completion of cDNA synthesis, samples were purified with a 0.6x Ampure XP bead ratio (s. 3.2.3) and eluted in 42 µL H_2O.

For PCR amplification of cDNA libraries, the reagents from the Advantage 2 PCR enzyme system were used. 40 µL cDNA library were mixed with 42 µL H_2O, 10 µL 10x Advantage 2SA PCR buffer, 2 µL 10 mM dNTP mix, 2 µL 5'PCR primer II A, 2 µL Universal RV primer and 2 µL 50X Advantage 2 Polymerase Mix. A thermocycler was started using the following program: 98°C for 1 min, [98°C for 10 sec, 63°C for 15 sec, 68°C for 3 min] x 22 cycles, 68°C for 7 min. Samples were placed in thermocycler upon reaching a temperature of 98°C. cDNA libraries were purified twice using 0.6x Ampure XP bead ratios (s. 3.2.3) and eluted in 40 µL H2O. In case of too low (or too high) cDNA library yields, the PCR cycle number was adapted between 20 and 24 cycles.

cDNA library profiles were analyzed by Fragment Analyzer High Sensitivity NGS Fragment Analysis Kit to validate high-quality library profiles.

PacBio sequencing was performed by the Genomics Core Facility of the Max Delbruck Center for Molecular Medicine. Before sequencing, PacBio adapters were added to cDNA libraries, which also enabled multiplexing. This step was also performed by staff from the Genomics Core Facility who also performed processing of PacBio Sequel 'movies' into consensus .fastq reads using PacBio SMRTLink software.

3.2.9 Metabolic labeling and streptavidin pulldown of biotinylated RNA

HEK Flp-In 293 T-REx cells were cultured in 4x 15 cm dishes in DMEM medium. 4-Thiouridine (4sU, dissolved in DMSO) was added at a final concentration of 1 mM and incubated for respective experimental timepoints up to 90 min. For 0 min control, DMSO without 4sU was added. Cells were washed once with 10 mL cold PBS, then 5 mL Trizol was added per dish and RNA was extracted as described above (s. 3.2.1). RNA was resuspended in a total volume of 50 µL.

For each sample, two biotinylation reactions were prepared: 100 µg total RNA input in 200 µL H_2O were each mixed with 200 µL 2.5x RNA Biotin Labeling Buffer and 100 µL 1 mg/mL biotin-EZ-link and incubated for 2 h on a rotator.

Biotinylated RNA was purified by PCI extraction (s. 3.2.2) in a final volume of 40 µL. Samples of biotinylated RNA were taken for dot blot analysis (s. 3.2.11). RNA was then denatured for 3 min at 70°C and placed on ice before performing pulldowns.

Pulldown experiments were performed using MyOne Streptavidin C1 beads. 120 µL bead suspension were washed three times with 150 µL MPG buffer on a magnetic rack. Biotinylated RNA in 40 µL H_2O was added to 120 µL MPG buffer on Streptavidin beads and incubated for 15 min at RT with rotation. Supernatants were separated from biotinylated RNA bound to beads by incubating on a magnetic rack for 1 min and collecting the supernatant as unbound fraction. Beads were washed three times with 150 µL MPG buffer pre-warmed to 37°C. 150 µL 100 mM DTT was then added and incubated for 5 min for elution of biotinylated RNAs as bound fractions. RNA from bound and unbound fractions was purified by PIC extraction (s. 3.2.2). FLAM-Seq libraries were prepared from bound and unbound fractions and sequenced as described above. FLAM-Seq libraries were prepared from bound fractions after 0 min, 10 min, 15 min, 20 min, 45 min, 90 min metabolic labeling and corresponding supernatant fraction of 0 min, 10 min, 15 min and 20 min timepoints. Since FLAM-Seq required comparably high input, RNA designated for technical replicates had to be pooled to guarantee more than 2 µg RNA input for poly(A) selection.

3.2.10 Metabolic labeling of RNA and SLAM-Seq in combination with poly(A) profiling

The SLAM-Seq (SH-linked alkylation for the metabolic sequencing of RNA) protocol [120] was used in conjunction with FLAM-Seq library preparation and PacBio sequencing for investigation of poly(A) tail dynamics over time by labeling of newly synthesized RNA using 4sU. Incorporated 4sU moieties were then derivatized by iodoacetamide (IAA), which causes mismatched cytosines to be build in cDNA during reverse transcription at positions of derivatized 4sU.

HeLa S3 cells were cultured in DMEM medium. Cells were seeded on 6-well plates until reaching 70% confluency. For metabolic labeling, the medium was supplemented with 500 µM 4sU in DMSO or DMSO control and incubated for 0 min, 90 min, 180 min. Cells were harvested, and RNA was extracted using Trizol as described above (s. 3.2.1).

Polyadenylated RNA was extracted from 10 µg total RNA per sample using TruSeq RNA purification beads (Illumina) and eluted in 15 µL H2O as described above for FLAM-Seq library preparation (s. 3.2.8).

GI-tailing was performed using the USB poly(A) length kit (Thermo Fisher). 4 µL 5x tail buffer mix was added to 2 µL 10x tail enzyme mix and incubated for 1 h at 37°C. 1.5 µL Stop solution was added to quench the reaction. GI-tailed RNA was cleaned up using a 1.8x ratio of RNAClean XP beads. For alkylation reactions, leading to T-C conversions, 15 µL GI-tailed RNA was incubated with 5 µL 100 mM iodoacetamide (IAA), 25 µL DMSO and 5 µL 0.5 M $NaPO_4$ pH 8.0 buffer for 15 min at 50°C. The reaction was quenched by addition of 1 µL 1 M DTT. RNA was purified using a 1.8x ratio of RNAClean XP beads. RNA reverse transcription was performed as in the FLAM-Seq protocol (s. 3.2.8) with reagents from the SMARTScribe Reverse Transcriptase kit: 16 µL GI-tailed RNA was incubated with 2 µL dC 3T UMI RT primer for 3 min at 72°C and placed on ice. 22 µL RT Mix (8 µL 5x RT buffer, 1.5 µL 100 mM DTT, 4 µL 10 mM dNTPs, 2 µL Ribolock, 2 µL IsoTSO 12 µM, 2 µL SMARTScribe RTase, 2.5 µL H_2O) were added and incubated for 1 h at 42°C, 10 min 70°C then 4°C hold. cDNA was purified using 0.6x ratio of Ampure XP beads. cDNA was eluted in 42 µL H_2O.

PCR amplification was performed with reagents from the Advantage 2 PCR enzyme system. For PCR amplification, 10 µL 10x Advantage 2SA PCR Buffer, 2 µL 10 mM dNTPs, 2 µL PCR Primer II A (12 µM), 2 µL Universal RV primer 10 µM and 2 µL 50x Advantage 2 Polymerase and 42 µL H_2O were added and incubated using the following program 98°C 1 min [98°C 10 sec, 63°C 15 sec, 68°C 3 min] 68°C 3 min. cDNA libraries were cleaned up 2x 0.6x Ampure XP beads and sequenced.

3.2.11 Dot blot analysis of biotinylated RNA

Dot blots were prepared by spotting 5 µg extracted RNA on an Amersham Hyperbond N+ membrane, which was positioned in a dot blot filtration unit on top of a layer of Whatman paper soaked in water. A vacuum was applied to the filtration unit. The membrane was dried and crosslinked for with 2x 1200 µJ at 254 nm. The membrane was incubated with methylene blue for 10 min and washed several times with H_2O. Stained RNA spots were imaged using standard illumination.

The membrane was blocked 20 min in 10% blocking solution and probed for 10 min with a 1:10.000 dilution of Strep-HRP antibody in 10% blocking solution. Membranes were washed each 2x for 5 min with 10%, 1% and 0.1% blocking solution. ECL select reagent was added to

membranes and membranes were imaged using chemiluminescence detection mode on an imager using auto-exposure settings.

3.2.12 Biochemical fractionation of chromatin, nucleoplasm and cytoplasm

For investigation of subcellular poly(A) tail length distributions, chromatin, nucleoplasmic and cytoplasmic compartments were isolated from HeLa S3 cell lines.

Cells were each seeded in two 10 cm dishes per replicate in DMEM medium. Cells were washed once with cold PBS. 2 ml cold PBS, supplemented with 1:200 Ribolock and 1:100 Protease Inhibitor, was added and cells were scraped from dishes and collected in 2 mL tubes. 50 µL cell suspension was each collected as input fraction for qPCR and Western Blot analysis.
The cell pellet was carefully resuspended in 500 µL 1x lysis buffer and incubated for 5 min on ice. A cushion of 500 µL 1x lysis buffer / 50% sucrose solution was pipetted under the cell lysate. Lysates were centrifuged into the sucrose cushion for 10 min at 16,000 g. Supernatants were collected as cytoplasmic fraction, and 50 µL sample were each taken for Western Blot analysis. Nuclei pellets were again resuspended in 500 µL 1x lysis buffer and 500 µL 1x lysis buffer / 50% sucrose solution was pipetted below the resuspended lysate. The procedure was repeated once. Pellets were carefully resuspended in 100 µL Nuclear Buffer I. 1 mL Nuclear Buffer II was added, then tubes were inverted 5 times and incubated for 15 min on ice. Suspensions were centrifuged for 10 min at 16,000 g. Supernatants were collected as nucleoplasmic fraction and 50 µL were collected for Western Blot analysis. Chromatin pellets were resuspended in 500 µL H_2O and 50 µL were collected for Western Blot analysis. 5 vol Trizol were added to resuspended fractions and RNA was extracted by Phenol-Chloroform extraction as described above (s. 3.2.1). Collected lysates from fractions were analyzed by Western Blot for cytoplasmic contamination using GAPDH as cytoplasmic marker, TBP43 as cytoplasmic / nuclear marker and BCAP31 as ER marker. Western Blots were performed as described below (s. 3.2.12).
FLAM-Seq libraries from subcellular fractions were prepared as described above using 5-10 µg total RNA is input for poly(A) selection (s. 3.2.8) and sequenced.

3.2.13 Western Blot analysis of contamination in biochemical fractions

For Western Blot analysis, 15 µL suspension from input, cytoplasmic and chromatin fraction, as well as 30 µL nucleoplasmic fraction were mixed with 3 µL or 6 µL 5x SDS PAGE loading buffer and denatured for 5 min at 90°C. Samples were loaded on 12% SDS-PAGE gels and run according to standard protocols by Laemmli [348]. Blotting was performed using the BioRad TransBlot Turbo pre-made kits and standard settings for 2 mini gels. Membranes were blocked

using 5% skim milk in TBS-T buffer for 1 h. Membranes were then probed with BCAP31 (1:2000 dilution), GAPDH (1:5000 dilution) and TBP-43 (1:5000 dilution) antibodies in 5% skim milk / TBS-T overnight. Membranes were washed three times for 5 min with TBS-T and incubated with 1:10.000 anti-mouse- or anti-rabbit-HRP antibodies in 5% skim milk in 1x TBS-T buffer. Membranes were washed three times for 5 min with TBS-T buffer. Membranes were then probed with ECL Select solution and imaged.

For certain experiments, membranes were stripped and re-probed with antibodies: For stripping, membranes were covered with mild stripping buffer and incubated for 10 min while shaking. Buffer was discarded and fresh stripping buffer was added and incubated for another 10 min. Membranes were washed twice in 1xPBS with 10 min incubation. Next, membranes were washed twice with TBS-T for 5 min. Before addition of a new antibody, membranes were blocked again for 1 h in 5% skim milk.

3.2.14 Splicing inhibition in HeLa S3 nuclei

Splicing inhibition was performed using SF3b inhibitor Pladienolide B (PlaB) [28]. HeLa S3 cells were treated with PlaB and nuclei were extracted to enrich for unspliced pre-mRNAs.

For each replicate 2x 10 cm dishes HeLa S3 cells were grown to 70-80% confluence. For splicing inhibition, medium was changed to DMEM + 10% Tet-free FBS supplemented with 10 µL 100 nM PlaB in DMSO or 10 µL DMSO as control. Cells were incubated for 3 h with PlaB. Two replicates for PlaB and control samples were prepared. Isolation of nuclei was performed using the biochemical fractionation protocol as described above (s. 3.2.12) but stopping after complete separation of cytoplasmic fraction and 3x centrifugation through a sucrose cushion. For extraction of nuclear RNA, nuclei were lysed in Trizol after three rounds of centrifugation through a sucrose cushion. Nuclear RNA was extracted as described above (s. 3.2.1). 10 µg total RNA input was used for poly(A) selection and FLAM-Seq library preparation as described above (3.2.8).

3.2.15 Transcription inhibition in HeLa S3 cell lines

Transcription inhibition using Actinomycin D (ActD) was performed in HEK Flp-In 293 T-Rex cells. Cells were seeded at $0.5x10^6$ cells per well in 6 well plates. DMEM medium supplemented with 5 µg/mL ActD was added to each well. Cells were incubated for up to 12 h. Replicate samples for each time point were harvested every 2 h by scraping cells in cold 1x PBS buffer and centrifuging cells for 5 min at 300 g. PBS supernatant was removed and RNA was extracted

using Trizol reagent as described above (s. 3.2.1). To assess the efficiency of transcription inhibition, expression levels for less stable SCD and ID3 genes were quantified by qPCR as described above (s. 3.2.6), in technical triplicates and biological duplicates. For the 12 h ActD timepoint, only one biological replicate was measured.

FLAM-Seq libraries were prepared for ActD time course experiments as described above (s. 3.2.8).

3.2.16 Cas13b RNA knockdown of PAN2, PAN3, CNOT7 & CNOT8

The Cas13b system [349] was used for targeted mRNA knockdowns of PAN2, PAN3, CNOT7 and CNOT8 deadenylases. Knockdown experiments were performed in HEK Flp-In 293 T-REx cells. $0.5x10^6$ cells were seeded in for triplicates in 6 well cell culture dishes for 24 h in DMEM medium. Before transfections, medium was changed. For each transfection, 5 µL Lipofectamin 2000 was mixed with 150 µL Optimem and incubated for 5 min. 1 µg Cas13b plasmid was mixed with 75 µL Optimem for each transfection. Two guide RNA expressing plasmids were mixed for PAN2, PAN3, CNOT7 and CNOT8 at a total mass of 3 µg and mixed with 75 µL Optimem. 3 µg non-targeting guide RNA was used as control. Guide RNA and Cas13b plasmids were mixed to a total volume of 150 µL per transfection. 150 µL plasmids and 150 µL lipofectamine were mixed and incubated for 10 min at RT before adding to cell culture dishes. Cells were transfected for 48 h, afterwards medium was removed, cells were washed with 1x PBS and 900 µL Trizol was added to each well for RNA extractions as described above (s. 3.2.1). Expression levels were quantified by qPCR using primer pairs for PAN2, PAN3, CNOT7, CNOT8 and GAPDH as reference gene (s. 3.2.6). Expression levels were further quantified by Nanostring nCounter measurements using gene panels containing target genes, according to the Nanostring standard protocol (s. 3.2.7). qPCR and Nanostring measurements were analyzed as described below (s. 3.3.2 / 3.3.3).

3.2.17 siRNA knockdown PAN2, PAN3, CNOT7 & CNOT8

Knockdowns of PAN2, PAN3, CNOT7 and CNOT8 deadenylase complexes was performed by siRNA interference in HEK Flp-In 293 T-REx cells. Cells were seeded at $0.5x10^6$ / well in triplicates in 6-well cell culture dishes in DMEM medium. Before transfection of siRNAs, medium was carefully replaced. For each sample, 150 µL Optimem medium was mixed with 9 µL RNAiMAX reagent. In a separate tube, siRNAs were mixed with each 150 µL Optimem. For each transfection, a total of 96 pmol siRNA was transfected. siRNAs against PAN2 and PAN3, as wells as CNOT7 and CNOT8 were co-transfected for double knockdowns or

transfected all at once for simultaneous knockdown of all targets. BLOCK-iT™ Alexa Fluor™ Red Fluorescent Control siRNA was used for control transfections. 150 µL each Optimem plus siRNA and RNAiMAX were mixed an incubated for 10 min at RT, before adding to cells. Cells were typically grown for 48 h after transfection.

For analysis of knockdown efficiencies, cells were washed in 1x PBS and RNA was extracted as described above (s. 3.2.1) and gene expression levels were quantified by qPCR as described above (s. 3.2.6) using qPCR primer pairs against PAN2, PAN3, CNOT7, CNOT8 and GAPDH as a reference gene.

3.2.18 CasRx knockdown PAN2

As an alternative to Cas13b, the CasRx system [350] was used for knockdown of the PAN2 target gene. CasRx and guide RNA plasmid transfections were performed by Ivano Legnini (Max Delbruck Center). Cell were washed once in PBS and RNA was extracted as described above (3.2.1). Gene expression levels were quantified by qPCR as described above (s. 3.2.6) using qPCR primer pairs for PAN2 and GAPDH as a reference gene.

3.2.19 shRNA knockdown PAN3, CNOT7 & PARN using stable, doxycycline inducible shRNA expressing cell lines

Stable shRNA expressing cell lines against PAN3, CNOT7 and PARN deadenylases were produced by Ivano Legnini (Max Delbruck Center). shRNA expression could be induced by doxycycline (dox) through a Tet-inducible promoter system. To test RNA knockdown efficiencies of target genes, shRNA expressing cell lines were seeded in 12-well cell culture dishes at $5x10^4$ cells per well. shRNA expression was induced by supplementing DMEM medium with 100 ng/mL or 500 ng/mL doxycycline (dox). As a control, DMEM medium without dox was added. Cell lines were treated with dox up to 6 days and fresh DMEM medium (plus dox) was added every 2 days. Cells were harvested in each 1 mL PBS and scraped from wells. 200 µL cell suspension was each kept for Western Blot analysis. 800 µL cell suspension was each centrifuged for 5 min at 300 g. Supernatants were removed and RNA was extracted from cell pellets as described above. To assess the efficiency of shRNA knockdowns, gene expression levels of PAN3, CNOT7, PARN and GAPDH as a reference gene were quantified by qPCR as described above (s. 3.2.6).

shRNA knockdown experiments were performed in combination with biochemical fractionation. For fractionation experiments, PARN-1+2, PAN3-2, CNOT7-1+2 cell lines were

seeded at $2x10^6$ cells per 10 cm dish in DMEM medium. shRNA expression was induced changing medium to DMEM supplemented with 500 ng/mL dox or DMEM medium without dox as control. shRNA induction was performed for 5 days for each cell line and additionally for 3 days for PAN3-2 shRNA cell line. Medium was changed every 48 h to fresh DMEM supplemented with 500 ng/µL dox.

Biochemical fractionation experiments were performed as described above (s. 3.2.12). RNA was extracted from input cell fractions as described to assess knockdown efficiencies by qPCR (3.2.6) and Western Blot (3.2.13) as described above. Cytoplasmic contamination of nuclear fractions was assessed by Western Blot as described above (s. 3.2.12). RNA from cytoplasmic, nucleoplasmic and chromatin fractions was extracted, and FLAM-Seq libraries were prepared from 4-10 µg total RNA input as described above (s. 3.2.8) and sequenced.

3.2.20 Growth curve measurements of PAN3, CNOT7 & PARN shRNA cell lines

Cell growth was measured upon shRNA knockdowns in PAN3-2, CNOT7-1+2 and PARN-1+2 cell lines to assess phenotypic effects of PAN3, CNOT7 and PARN depletion. Cells were seeded at $0.3x10^6$ cells per well on 6-well plates in 2 replicates. shRNA expression was induced by adding DMEM medium supplemented with 500 ng/mL doxycycline (dox). Respective shRNA cell lines without dox induction were used as a baseline control. Microscopy images were taken every 24 h for 5 consecutive days at 10x magnification. 9 images were taken for each timepoint and replicate. Microscopy images were analyzed as described in computational methods (s. 3.3.21).

3.3 Computational Methods

3.3.1 Analysis of PAT assay electropherograms

Electropherogram `.csv` files exported from Prosize Fragment Analyzer software were loaded which specify UV absorption as a function of a marker with known DNA length for each sample analyzed. For converting the fragment analyzer length profiles to poly(A) tail length estimates, the amplified sequence upstream of the poly(A) site required for gene specific amplification were subtracted from the marker lanes for each primer used (GAPDH: 90, RPL37: 88, SCD: 95, BTF3: 90).

A baseline was subtracted from each measured length as min (UV absorption) and profiles were scaled by dividing each measurement by

$$x_{scaled} = \frac{x}{\max(UV - Absorption) - min\ (UV - Absorption)}$$

with a maximum length cutoff of 350 nt. Scaled poly(A) tail length profiles were then plotted.

3.3.2 Analysis of quantitative read-time PCR (qPCR) data

Ct values obtained from each measured target gene in each sample describe the PCR cycle number upon which the fluorescent dye signal surpasses a threshold value upon amplification. Each sample was measured in triplicates (technical replicates) and average Ct values were calculated. Average gene Ct values were normalized on a quantified reference (housekeeping) gene also measured for the same sample by subtracting average Ct values which was defined as DCt (delta Ct value).

$$DCt = Ct_{average,gene-}\ Ct_{average,house\ keeping}$$

To compare gene expression levels between a control and perturbation, perturbation Ct values were subtracted from control DCt, yielding DDCt (delta delta Ct) values.

$$DDCt = DCt_{perturbation-}\ DCt_{control}$$

DDCt values were converted into fold changes between perturbation and control by calculating

$$FC = 2^{-DDCt}$$

3.3.3 Analysis of Nanostring gene expression quantification data

Nanostring data were analyzed using nCounter4.0 software. Normalization was performed using 'Positive control normalization' and counts for each sample were plotted.

3.3.4 FLAM-Seq computational pipeline (FLAMAnalysis)

FLAM-Seq libraries were prepared as described above (s. 3.2.8) and sequenced on the PacBio Sequel system according to the manufacturer's guidelines. Sequel movies were converted to Circular Consensus Sequence (CSS) reads using the SMRT Link browser 5.0 software and exported as `.bam` files., which were converted to `.fastq` files using `bedtools bamtofastq`.

Poly(A) tails were extracted for each read and clipped reads were mapped to the respective genome to quantify gene expression using the FLAMAnalyis Pipeline (https://github.com/rajewsky-lab/FLAMAnalysis). FLAMAnalysis outputs a table containing read name, poly(A) tail sequence, poly(A) tail length and gene identity for mapped reads. In a first step reads are preprocessed to filter reads containing a characteristic substring consisting of a minimum of 9 C and 10 T nucleotides (nt), with 1 mismatch at maximum, which must be present on *bona fide* polyadenylated transcripts through appending a 3'-terminal GI-tail. Reads were oriented as reverse complement to the original mRNA sequence.

The poly(A) tail sequence was identified as majority vote of two algorithms: Algorithm one ('poly(A) extension' algorithm) identified *n* subsequent T nucleotides at the preprocessed read start and extends the identified poly(A) tail end by iteratively searching for *n+1* T nucleotides with a maximum of n/T_1 allowed mismatches. T_1 is an empirically defined parameter for calibrating allowed errors in poly(A) tails. Once no extended T sequence is identified, the tail is trimmed at its distal position to start in a T nucleotide adjacent to the putative 3'-UTR. Algorithm two ('poly(A) sliding window' algorithm) utilizes a sliding window approach to identify substrings of size L=20 starting from the preprocessed read start until the T nucleotide fraction drops below a defined threshold T_2, upon which the identified tail segment is trimmed from the distal positions until a TT dinucleotide is reached. T_2 is an empirically determined cutoff parameter. Both algorithms were run each with parameters T_1=[25,30,35,40] and T_2=[25,30,35,40]. Poly(A) tail sequences were identified by majority vote between the identified poly(A) tail coordinates for each algorithm-parameter combination. Poly(A) tail sequence and adjacent PCR adapter sequences were removed from the reads. Reads were mapped to human h38 (extended with ERCC spike in sequences), *C. elegans* WB235 or mouse GRcm38 genome using STAR with parameters: `'–outFilterMultimapScoreRange 20', '–`

`outFilterScoreMinOverLread‘, ‘–outFilterMatchNminOverLread 0.66’, ‘–outFilterMismatchNmax 1000’,’– winAnchorMultimapNmax 200’,’–seedSearchStartLmax 12’, ‘–seedPerReadNmax 100000’,’–seedPerWindowNmax 100’,’–alignTranscriptsPerReadNmax 100000’,’– alignTranscriptsPerWindowNmax 10000’,’–outSAMtype BAM SortedByCoordinate’.`

Feature counts was used to assign aligned reads in `.bam` format to gene annotations using FeatureCounts with parameters `'-L', '-g', 'gene_name', '-s 2', '-O','--fracOverlap','0.3', '-R', 'CORE'`. Gencode v28 annotation was used for human gene annotations, Gencode `GRCm38.101.gtf` annotation for mouse annotations and `WBcel235_82` for *C. elegans*. Reads and resulting from internal priming of cDNA synthesis were removed and poly(A) tail starts were clipped from genomic encoded nucleotides. For this, the coordinates 100 nt upstream of the 3’-UTR were extracted and genomic DNA sequence for corresponding coordinates was extracted from genome `.fa` files. The genome sequence was compared to the nucleotide sequence in raw reads to identify positions which are not encoded by the genome. This position was used to define the start of the poly(A) tail and genomic encoded nucleotides in poly(A) tails were removed from defined poly(A) sequences. This step also removed internal priming artefacts. Unique molecular identifier (UMI) sequences were further extracted and reads with identical UMIs were collapsed.

Read name, identified poly(A) sequence, poly(A) tail length, and gene name for each retained read were written to output files. Different processing statistics were collected by the pipeline during the analysis: the fraction of reads which contain a poly(A) tail sequence, fraction of mapped reads and fraction of mapped reads which could be assigned to annotated gene locus.

3.3.5 Visualization of FLAM-Seq genome browser tracks

Read FLAM-Seq alignments were visualized including associated poly(A) tail length using the IGV genome browser. To also visualize the poly(A) tails of the alignments, poly(A) tail length was first extracted for each read. The poly(A) tail was then added to the alignment by modifying the CIGAR string of `.bam` alignments adding mismatches of the same length as the poly(A) tail for each read as `(polyA_length)X`. To visually distinguish poly(A) tails from the templated part of a read, different colors were used which mark mismatches and highlight the poly(A) tail in IGV visualizations of alignments.

3.3.6 Read length, gene quantification, coverage and transcription start site analysis

Sequencing statistics for FLAM-Seq datasets were in all cases compared to a PacBio IsoSeq dataset prepared from Universal Human Reference RNA (UHRR) as gold standard reference

for long read RNA/cDNA sequencing. UHRR subreads were downloaded (https://downloads.pacbcloud.com/public/dataset/RC0_1cell_2017/m54086_170204_081430.subreads.bam) and converted to single molecule consensus reads using PacBio ccs software. Reads were then aligned to the human genome hg38 using STAR as described above. Features for mapped reads were annotated as described above using FeatureCounts and Gencode v28 annotation.

Read lengths distributions were extracted from FLAM-Seq `.fastq` files by opening files using `pysam` and determining length for each read. To obtain gene counts from FLAMAnalysis result `.csv` files, reads were aggregated and counted for each gene. Pairwise Pearson correlation matrices were calculated by joining gene counts obtained for human or *C. elegans* datasets. Expression of genes not detected in either dataset was set to 0.

To test the overlap of read starts with annotated human transcription start sites, CAGE (cap analysis gene expression) peaks were downloaded (http://fantom.gsc.riken.jp/5/data/hg19.cage_peak_phase1and2combined_ann.txt.gz) from the FANTOM5 database [351] and converted to BED coordinates. TSS peaks were sorted by gene. Aligned reads in .bam files were similarly grouped by genes. Reads were next assigned to TSS annotation for each gene by comparing coordinates of read starts with TSS coordinates. Reads which could not be assigned to an annotated TSS were assigned to a 'no_tss' group for each gene. For analysis of TSS in *C. elegans* samples, TSS peaks identified from SAGE (serial analysis of gene expression) datasets [352] were converted to BED format. Next adjacent peaks within less that 10 nt distance were collapsed. TSS were next grouped by genes and read starts were assigned to TSS bins as above. The fraction of reads mapping to annotated TSS for each gene were calculated as sum of all read in annotated TSS bins for a gene by all reads. The median fraction of reads mapping to TSS was then computed across all detected genes in a sample along with the standard deviation. To assess specificity of mapping read starts to TSS, reads were shortened *in silico* by removing *n* nucleotides from the reads start, i.e. modifying the read start coordinate in alignment files depending on strand orientation of the alignment. Reads were then assigned to TSS as described above and median assigned reads per gene as well as standard deviations were reported.

Coverage across genes of different length was analyzed by computing normalized coverage for each gene. Exon coordinates were extracted for each protein coding transcript and sorted by genes as 'meta-exons' after joining exons with overlapping coordinates. Next, read coverage was calculated for each meta-exon using `pybedtools`. For each gene, an array containing coverage counts for 'meta-transcripts' was concatenated from meta-coverage exons. Coverage

profiles of ‘meta-transcripts’ were normalized by scaling by maximum coverage for a rolling average across 25 positions. ‘Meta-transcripts’ were then sorted by length and visualized as heat maps using `matplotlib`. A minimum coverage of 5 counts was required across at least one position for each gene in the analysis.

3.3.7 Poly(A) tail length calibration using poly(A) standards and PAT assays

cDNA and RNA standards with known poly(A) tail length were used to produce FLAM-Seq libraries (Dr. Ivano Legnini, MDC). One sequencing library was produce for 4 mixed cDNA standards, which were only PCR amplified and one library for the RNA standard that was processed through the complete FLAM-Seq protocol. Obtained PacBio datasets were processed using FLAMAnalysis (s. 3.3.4) but performing only the `preprocess` and `quantTail` steps which filtered datasets and estimated the poly(A) tail length for each read. Each cDNA standard contained a unique barcode which was extracted from each read to identify the associated cDNA standard and expected poly(A) tail length. Reads were then grouped by cDNA standards. Median poly(A) tail length and standard deviations were calculated for each cDNA and RNA standard.

Two algorithmic approaches were combined to estimate poly(A) tail length from FLAM-Seq reads (s. 3.3.4). To benchmark the results of different algorithms and parameters to define poly(A) tail length, 50 randomly sampled reads from HeLa S3 datasets were analyzed with each algorithm using different parameters: For the ‘poly(A) extension’ algorithm 1 the threshold parameter T_1 was set to T_1=[15,20,25,30,35,40] and the resulting poly(A) tail length for each parameter was plotted for each read. For the ‘poly(A) sliding window’ algorithm 2, parameter T_2 was set to T_2=[0.7, 0.85, 0.95] keeping window size (WZ) parameter at WZ=30 or modifying WZ=[10,30,50] while keeping T_2=0.8. Poly(A) tail length for each read and parameter combination was plotted as well.
Poly(A) tail length for the 50 random reads was further annotated manually and compared to algorithmic quantification using the majority vote algorithm implemented in the FLAMAnalysis pipeline.

Poly(A) tail length profiles were validated by PAT assay (3.2.5) (Ivano Legnini, Max Delbruck Centrum) for BTF3, GAPDH, MT-CO1, MT-CO3 and RPL37 in HeLa S3 RNA. PAT assay electrophoresis datasets were processed as described above (s. 3.3.1) and compared to poly(A) tail length profiles of the same genes in HeLa S3 FLAM-Seq datasets. For FLAM-Seq poly(A) distributions, histogram frequencies were normalized to the maximum frequency of each replicate.

3.3.8 Poly(A) tail distributions, comparisons between replicates and sequencing technologies

Poly(A) tail bulk distributions per unique sequenced molecule (UMI) were calculated as densities across 10 nt bin intervals up to 2000 nt maximum poly(A) tail length for each replicate. The same procedure was applied for calculating median poly(A) tail length per gene density profiles. Scatter plots and Pearson correlation coefficients between replicates were calculated after first joining median poly(A) tail length matrices between replicates and removing genes not detected in both replicates and second filtering genes for expression values greater than the filter value $f = 1$ to $f = 50$ in each replicate.

For comparisons between poly(A) sequencing technologies, PAL-Seq data for HeLa cell line were downloaded as processed dataset from Gene Expression Omnibus (GEO) accession GSE52809 (https://ftp.ncbi.nlm.nih.gov/geo/series/GSE52nnn/GSE52809/suppl/GSE52809_HeLa_total.txt.gz; access 20.05.2021) containing counts and median poly(A) tail length estimates for each detected gene. TAIL-Seq data for HeLa cell line were downloaded from publication supplement (https://ars.els-cdn.com/content/image/1-s2.0-S109727651400121X-mmc2.xlsx; access 20.05.2021). Median poly(A) tail length estimates per gene were compared after datasets were joined by gene and excluding genes not detected in both datasets and in both datasets expressed with less than f counts.

3.3.9 Poly(A) tail comparison gene expression, half-life, TE and GO term enrichment

HeLa S3, iPS cells, organoid, and *C. elegans* replicate datasets were merged and median poly(A) tail length was each plotted against log2-transformed gene expression counts. Pearson correlation coefficients were calculated between median poly(A) tail length per gene and log2 gene expression counts. Statistical significance was tested using R cor.test function. HeLa mRNA half-life data from Tani et al. [119] were downloaded (https://genome.cshlp.org/content/suppl/2012/02/14/gr.130559.111.DC1/Tani_Supp_Tables_revised2.xls; access 21.05.2021). Gene accessions were converted to HGNC symbols using R biomaRt package. HeLa translation efficiency data from Subtelny et al. [157] were downloaded from Gene Expression Omnibus (GEO) dataset GSE52809 (https://ftp.ncbi.nlm.nih.gov/geo/series/GSE52nnn/GSE52809/suppl/GSE52809_HeLa_total.txt.gz; access 20.05.2021). Median poly(A) tail length per gene was then plotted against half-life and translation rates.

Enrichment of gene ontology (GO) terms for genes sorted into 'short' or 'long' poly(A) tail length bins were calculated using R topGO package [353]. Genes were first binned by median poly(A) tail length into 'short', 'medium' or 'long' poly(A) tail length bins, where 'short' and 'long' were defined as median tail length shorter/long than lower/upper quartile of the median poly(A) tail length per gene distribution. Enrichment of GO terms was then calculated for genes in 'short' or 'long' poly(A) tail length bins against a background of all expressed genes in the merged HeLa dataset. For top enriched GO terms in 'short' and 'long' poly(A) tail bins, all in genes detected in HeLa associated to the GO term were extracted and the fraction of genes for the GO term categorized into 'short', 'medium' or 'long' bin were plotted.

3.3.10 Statistical modeling of differences in poly(A) tail length distributions

Since poly(A) tail length distributions were continuous and had a non-normal distribution for most genes, different statistical methods were required to identify relevant differences between poly(A) tail length distributions. To model the technical error for sequencing longer poly(A) tails, a linear model was fit on standard deviations from cDNA synthetic spike in data as a function of the median poly(A) tail length.

To evaluate power of different statistical approaches for identifying differences in poly(A) tail length distributions, three methods were tested on simulated poly(A) distributions: First, a 'poly(A) tail resampling' method explicitly modeled individual measurement errors for each quantified tail based on the technical error for a given poly(A) tail length inferred from synthetic cDNA standards: For measured poly(A) tail length, the expected standard deviation was inferred from the cDNA linear model fit for the expected variability of a tail with the measured length. Poly(A) tail length was then resampled based on a Gaussian with a mean corresponding to the measured tail length and inferred standard deviation. The medians of the two resampled poly(A) distributions were then compared. Poly(A) distributions were resampled 1000 times and a p-value was calculated based on the frequency of randomly observing the shorter poly(A) tail length distribution with longer tails after reshuffling. Second, a non-parametric Wilcoxon test was applied to compare differences between poly(A) tail length distributions. Third, a 'label swap' test was constructed to compare two poly(A) distributions: Sample assignments were randomly distributed across individual poly(A) tail measurements, then medians in distributions were calculated and compared to the original poly(A) difference. A p-value was calculated based on the frequency of observing the randomized differences in medians larger than original difference. To test sensitivity for detecting differences in medians of poly(A) tail length distributions, each two poly(A) distributions were simulated as Gaussians with standard

deviation of 49 nt and a mean of ‘basal poly(A)’ for the shorter poly(A) distribution and ‘basal poly(A)’ + ‘simulated difference in medians’ for the longer poly(A) distribution, where ‘simulated difference’ was between 0 nt and 50 nt. For each distribution, a different number of counts, i.e. poly(A) tails from the distribution were drawn, with counts between 5 and 100. 100 samples were simulated for each combination of counts, ‘basal poly(A)’ and ‘simulated difference in medians’ parameters and the median of the resulting p-values distribution and standard deviation was displayed. Median poly(A) tail length between merged iPSC/organoids and *C. elegans* samples was compared using the ‘poly(A) tail resampling’ method described above, for all genes with on average more than 5 counts. Resulting p-values for each gene were corrected for multiple hypothesis testing using the Benjamini-Hochberg method.

3.3.11 3’-UTR isoform annotation of FLAM-Seq datasets

Gene 3’-UTR end annotations and 3’-UTR isoform models were *de novo* constructed from FLAM-Seq reads. Coordinates of alignment ends with correspond to 3’-ends polyadenylated RNA, were extracted for each gene for FLAM-Seq datasets. 3’-UTR ends were predicted from coordinates of alignment ends for a given gene by peak detection over the coordinates from individual genes using the Python `peakutils` module (https://bitbucket.org/lucashnegri/peakutils/src/master/; access 27.05.2021). The peakutils `peakutil.index` function was used and with peak height threshold `thres=0.1` and minimum peak distance `min_dist=30`. After identification of 3’-UTR end peaks, reads were classified by the distance of alignment end coordinates to the 3’-UTR end peaks with a maximum distance of 15 nt. This allowed assignment of poly(A) tail length estimates defined by the FLAMAnalysis pipeline before for each read to 3’-UTR isoforms.

For calculation of 3’-UTR length for each defined isoform, annotated 3’-UTR starts / coding sequence (CDS) 3’-ends were extracted from Gencode human GTF annotation v28 (https://www.gencodegenes.org/human/; access 27.05.2021) or *C. elegans* Wormbase GTF WBcel235_82 annotation (https://parasite.wormbase.org/ftp.html; access 27.05.2021). Coordinates of the splice site closest to the 3’-end in FLAM-Seq alignments were compared to last exons starts in GTF annotation files to extract possible 3’-UTR starts for each isoforms. For each 3’-UTR end isoform, the last exon start which occurred most, was selected as 3’-UTR start coordinate. 3’-UTR length was then calculated as absolute difference between 3’-UTR start and end coordinates.

Nucleotide frequencies adjacent to 3’-UTR end / cleavage site were extracted from each preprocessed alignment coming from FLAMAnalysis pipeline (`cleaned.bam` alignments) and

computed across all sequenced reads for positions upstream of cleavage site. Nucleotide frequencies across poly(A) tails were calculated from poly(A) tail 5'-ends. Genomic nucleotide frequencies downstream of the cleavage site were calculated for the first 20 positions downstream of the cleavage site, which were extracted form the genomic DNA sequence as defined in genome `.fa` files for the FLAMAnalysis pipeline by comparing the alignment end coordinates to the genome sequence.

DNA hexamer sequences were extracted from alignment 3'-ends and hexamer occurrences were counted for each 3'-UTR isoform to identify genes where 3'-UTR end cleavage position was less specific. Gini coefficients, which quantity the inequality across factor counts, were calculated for counts of all possible hexamers (4096 total) for 3'-UTR end hexamers. Gini coefficients were then scaled to span the range of 0 to 1.

Polyadenylation signal (PAS) usage counts were extracted by counting occurrences of any previously identified possible PAS variant [61] within a window of 60 nt from each reads 3'-end. 'NA' PAS was assigned when no PAS hexamer could be detected. PAS counts were then normalized to total counts, compared between replicates and visualized, along with distributions of most frequently occurring PAS positions from the 3'-end of a read.

Differences in 3'-UTR lengths between genes were calculated by first computing the average 3'-UTR length for each gene in cases of multiple 3'-UTR isoforms and then calculating the differences in 3'-UTR lengths for each gene between different samples.

3.3.12 Identification of alternative polyadenylation and transcription start site isoform associated differences in poly(A) tail length profiles

Isoform specific differences in poly(A) tail length profiles were analyzed for all FLAM-Seq datasets. Reads were grouped based on annotated 3'-UTRs as described above. Poly(A) tail length distributions associated with each 3'-UTR isoform were then tested for statistically significant differences in median poly(A) tail length against all other isoforms detected for a given gene using the 'poly(A) tail resampling' method described above and p values were adjusted to a false discovery rate of 5% using the Benjamini-Hochberg method.

3'-UTR isoforms were annotated as 'proximal' or 'distal' based on the coordinates of 3'-UTR end annotations and the gene orientation. Median poly(A) tail length distributions per gene were tested for statistically significant differences between 'proximal' and 'distal' isoform distributions across all genes using a two-sided Wilcoxon test.

For analysis of transcription start site (TSS) related differences in poly(A) tail length distributions, reads were grouped by annotated human and *C. elegans* transcription start sites annotated in CAGE / SAGE dataset as described above. Differences in poly(A) tail length profiles between transcription start site isoforms were then tested using 'poly(A) tail resampling' method and p-values were corrected adjusted to FDR of 5% using the Benjamini Hochberg method. To remove spurious association that may result from truncated reads and reduce the analysis, only those TSS isoforms were considered that were covered by at least 20% of all reads for a given gene.

3.3.13 Identification of non-A nucleotide sequences in poly(A) tails

Non-A nucleotides in poly(A) tails were identified by counting T, G and C nucleotides in valid poly(A) tails of each read identified by the FLAMAnalysis pipeline in merged HeLa S3, organoids, iPS cells, and *C. elegans* samples, as well as cDNA and RNA standards. Non-A counts were in one instance normalized to all sequenced nucleotides in a sample or normalized to the nucleotide content of a given tail, and then averaged across all sequenced poly(A) tails in a sample. The fraction of sequenced molecules for each gene which contain non-A nucleotides was extracted by counting reads which contain at least one non-A nucleotide against all reads mapped to a gene.

Poly(A) were aligned at their 5'- or 3'-ends by counting non-A nucleotides occurring for an index running from the beginning, (i.e. 1,2,…) or end (tail_len – 1,2,…,) for all sequenced poly(A) tails and calculation of non-A frequencies for each index position. The index value here ranged from 1 to 200 for human samples and 1 to 120 for *C. elegans* samples, since few tails were here detected with tails longer than 120 nt in *C. elegans* samples. As an orthogonal approach, poly(A) tails were binned in 10 nt length bin intervals and frequencies of non-A nucleotides were calculated for each bin.

3.3.14 Analysis of transcription inhibition experiments using Actinomycin D

FLAM-Seq replicate samples obtained from control (0 h), 2 h, 6 h and 12 h Actinomycin D (Act D) treatment were processed using the FLAMAnalysis pipeline. Gene half-life measurements for HEK293 cells were obtained from [354] and averaged across replicates. Genes were binned by half-life and median tail length was visualized for each bin.

3.3.15 Analysis of unspliced, intronic reads in FLAM-Seq datasets

As a first step in identification of unspliced, intronic reads in the FLAM-Seq datasets, a reference of intronic sequences was curated which contained intron coordinates which

exclusively did not overlap with any exonic sequences of other isoforms of the same gene (curated by Ivano Legnini, Max Delbruck Center). Intron, exon and 3'-UTR coordinates were downloaded from UCSC table browser using human hg38 and mouse mm10 annotations. Genes were filtered for 'protein-coding' biotype. Next all genes were filtered which overlapped with each other using `bedtools intersect`. Finally, all introns which overlapped any exons were filtered using `bedtools intersect` without allowing overlap between intron and exon coordinates. Alignments in `.bam` format were first matched against the curated database of introns and thereby required to have a minimum overlap of 50 nt using `bedtools intersect -a *_Aligned.sortedByCoord.out.bam -b intron_database.bed -wo –split -bed -S | awk '$19>=50' | awk '$10==1' | cut -f1-6,16 > intron_candidates.bed.` Intron coordinates were then matched against annotated 3'-UTRs to filter artefacts, e.g. from non-annotated transcripts or missing exon annotation, using `bedtools intersect -a intronic_reads.bed -b utr_annotation.bed -wa -S > intronic_reads.bed.`

Poly(A) tail length distributions were then binned by length and normalized to total reads in each FLAM-Seq sample.

Read length of intronic and all ('bulk') reads were computed by filtering reads from `.fastq` files based on read names of identified intronic reads. Read length distributions were then calculated for intronic and bulk read length bins.

The representation of genes detected in FLAM-Seq samples in the intron reference was calculated to define an upper bound of the number of genes for which unspliced reads could be detected. Since the read length limits the probability of detecting introns in large distances from annotated transcript 3'-ends, genes with closest introns with more than 3 kb distance from the 3'-end were removed and the fraction of detected genes with annotated introns was computed.

Downsampling of intronic reads was performed to assess whether unspliced reads were sampled from the majority of expressed genes in a FLAM-Seq dataset, limited by the sequencing depth, or restricted to subsets of genes. Intronic reads for FLAM-Seq HeLa S3, iPSC and organoid datasets were merged, then the number of genes with intronic reads was calculated. The merged intronic reads dataset was then randomly sampled to a fraction of the original dataset and the number of genes was calculated for downsampled datasets and divided by total genes detected in all FLAM-Seq samples.

Venn diagrams were calculated based on calculating the union and intersection of all detected intronic genes in FLAM-Seq datasets.

Intron length was compared by extracting first calculating the intron length distributions from the intron coordinates in the intron reference. Length distributions were then compared between all introns in the reference and introns overlapping with unspliced reads. Second, the length of all introns for each gene genes were extracted from Gencode v28 GTF annotations. The length of intronic genes length distributions were then compared between all expressed genes in a FLAM-Seq sample and genes with associated intronic reads.

Intronic poly(A) tail length distributions were plotted for each gene expression bin by first binning all genes by expression into 4 bins. Intronic reads and associated poly(A) tail length were then assigned for each bin by gene name and poly(A) distributions were visualized as box plots. Similarly, the fraction of intronic reads was calculated for each gene expression bin by computing the fraction of the number of intronic reads by total counts in each gene expression bin.

3.3.16 Analysis of splicing inhibition experiments using SF3b inhibitor PlaB

Poly(A) tail length distribution of intronic reads were calculated as described above, as well as Venn diagrams and poly(A) tail length and fraction of intronic reads by gene expression bin.

Differences in median poly(A) tail length per gene were calculated between merged control and PlaB-treated samples and genes were grouped into 30 bins based on median poly(A) tail length difference. The average expression was then calculated for each group of genes and visualized as color scale on median poly(A) difference distributions.

Log2 fold-changes changes were calculated based on gene counts in control and PlaB-treated samples and plotted against poly(A) difference. Genes with striking differences in poly(A) length and expression were annotated. Pearson correlation coefficient were calculated between poly(A) tail length differences and log2 fold-changes for each gene.

Genes with poly(A) tail length differences >50 or <-50 were labeled as ‘shorter’ or ‘longer’ upon PlaB treatment. Half-lives measured for each gene in HeLa cell lines [119] were compared between each poly(A) difference group as well as 2 random control groups which were size matched to the number of genes in ‘shorter’ or ‘longer’ bins. Similarly, 3’-UTR length was compared between bins. 3’-UTR length was calculated for individual 3’-UTR isoforms in FLAM-Seq datasets as described above (s. 3.3.11). 3’-UTR length was then compared for each 3’-UTR which is associated with the genes in each bin. Significance of differences in half-life and 3’-UTR lengths were compared by Wilcoxon test.

3.3.17 Analysis of Nanopore direct RNA sequencing of nascent RNAs

Published Nanopore direct RNA sequencing datasets of chromatin associated 4sU labeled-RNA from human K562 cell line was downloaded from GEO under accession number GSE123191 [43]. Poly(A) tail length estimates for Nanopore reads were kindly provided by Dr. Karine Choquet from Prof. Stirling Churchman's lab from Harvard Medical School, Boston, MA. Groupings of read ends into 'exonic', 'intronic', 'polyA', 'post_polyA', 'RNAPET' and 'splice site' were provided as metadata in the GEO archive. Reads from `.fastq` files were mapped using Minimap2 (version 2.16-r922) [355] with recommended parameter settings for Nanopore direct RNA sequencing data: `minimap2 -ax splice -uf -k14 -t 8 index fastq > sam`. Alignments were annotated by using featureCounts software and human gencode version28 gtf annotation `featureCounts -L -g gene_name -s 0 -t gene -O –fracOverlap 0.3 -R CORE -a gtf -o out sam`. Unspliced reads were extracted as described above, without requiring opposite strandedness (`'-S'`) for `bedtools intersect`. Poly(A) tail length profiles were then visualized for spliced and unspliced reads from 'intronic', 'polyA' and 'post_polyA' bins.

3.3.18 Analysis of RNA metabolic labeling and pulldown experiments

Metabolic labeling was performed in combination with streptavidin pulldowns of biotinylated RNA in HEK Flp-In 293 T-Rex cells and preparation of FLAM-Seq libraries from labeled (pulldown) and unlabeled (supernatant) fractions after 0, 10, 15, 20, 45 and 90 min 4sU labeling. No supernatant fractions were available for 45 and 90 min labeling. Poly(A) tail length profiles were compared between pulldown (PD) and supernatant (SN) fractions after calculating density for poly(A) tail length distributions. Intronic reads were extracted from pulldown and supernatant datasets using the computational pipeline described above and merged (s. 3.2.15).

Median poly(A) tail length per gene was compared between pulldown and supernatant by first calculating the median poly(A) tail length per gene for each labeling timepoint and the merged supernatant datasets and then calculating the difference between labeled and unlabeled. The difference in median poly(A) tail length per gene was then computed as the difference between labeled and control length and differences were visualized as cumulative density distributions for all genes with 3 or more counts. Poly(A) tail length distributions were then compared between different gene sets (immediate early genes (IEGs), lncRNAs, ribosomal proteins). lncRNAs were defined based on all genes with biomaRt 'lncRNA' biotype. Lists of IEGs and ribosomal proteins were curated manually for this analysis (s. datasets).

3.3.19 Analysis of SLAM-Seq / FLAM-Seq combination experiments

SLAM-Seq [120] was used in combination with a modified FLAM-Seq protocol as an orthogonal method to measure kinetics of newly synthesized RNA and poly(A) tails over time. For SLAM-Seq, cells were labeled with 4sU for 0, 90 and 180 min in replicates. 4sU incorporated into newly synthesized RNA was derivatized, which leads to effective conversions of uracil to cytosine (T-C conversion when compared to genomic DNA).

SLAM-Seq / FLAM-Seq raw data were processed as described above using the FLAMAnalysis pipeline. `.bam` alignment files which were annotated with MD tags for each alignment, which describes mismatching nucleotides towards the hg38 reference genome for alignment position. Mutations in alignments and their positions could then be quantified in a computational model. Reads with more than 30 mutations or read quality below 85 were excluded. The first 20 bases for each read were clipped since the first read positions showed an unexpected increase in mismatches which may have resulted from imprecise alignments at read starts. Similar to the GRAND-SLAM model proposed by Jürges et al. [356], the T-C conversion rate per thymidine ('labeling rate' p_{label}) was calculated based on the fraction of observed T-to-C conversions to sequenced Ts in each SLAM-Seq sample. A background T-C conversion rate (p_{error}), which corresponds to sequencing and other technical sources of observed T-C mutations was estimated from the observed non-T-C mutations in SLAM-Seq datasets under the assumption of uniform distributions of mutations. The probability of observing n T-C mutations for the read coming from labeled mRNA was calculated along the probability of observing n T-C mutations under the error model. Both probabilities were calculated as Binomials with $B(n;k;p_{label})$ and $B(n;k;p_{error})$ with n being the number of Ts in a read and k the observed T-C conversions. A read was defined as coming from labeled RNA if the log-likelihood for the reads was >1.15. This threshold was manually optimized to minimize the number of labeled reads in 0 min timepoints, where no labeled reads were expected while maximizing the number of detected labeled reads in 90 and 180 min labeling timepoints. The threshold was applied to all sequenced SLAM-Seq samples.

Intronic reads were extracted from SLAM-Seq datasets as described above (s. 3.3.15). Differences in median poly(A) tail length per gene was calculated between labeled and all detected reads ('steady state') for each gene for genes with 3 or more counts each sample.

3.3.20 Analysis of biochemical fractionation data in HeLa S3 and mouse brain samples

Sequencing data from FLAM-Seq library preparation of HeLa S3 subcellular fractions and mouse brain cytoplasmic and nuclear fractions were processed using the FLAMAnalysis pipeline as described above (s. 3.3.4) using human hg38 and mouse mm10 reference genomes and respective annotations. 3'-UTR isoforms annotations and identification of intronic reads in each dataset were also performed as described above (s. 3.3.11). Intronic reads were extracted from FLAM-Seq datasets as described above (s. 3.3.15).

Correlations of gene expression counts and median poly(A) tail length per gene between technical replicates of a biological sample were calculated between all genes with more than 10 counts per replicate.

Poly(A) tail length density distributions were calculated by computing the density in bins of 10 nt for poly(A) tail length for each replicate and then calculating average density and standard deviation for each bin, both for bulk poly(A) tail length and intronic poly(A) tail length. For intronic poly(A) tail length a smoothing step was added using the R `smooth` function.

A table with experimental parameters was curated for each biological replicate which contained 'median poly(A) tail length', 'median poly(A) tail length per gene', 'fraction mitochondrial reads', 'fraction intronic reads', 'RNA concentration after fractionation', 'number of sequenced reads' and 'subcellular fraction'. The median 'sample poly(A) tail length' was calculated as average median poly(A) tail length in cytoplasm, nucleoplasm and chromatin for each of the 12 fractionation replicates. An average poly(A) length was then calculated across all 'sample poly(A) length' and for each replicate a 'sample scaling factor' was calculated as 'sample poly(A) length' / average sample poly(A) length. This sample specific scaling factor was added as additional parameter. A linear model explaining median poly(A) tail length per sample was fitted as function of all described parameters and different models were evaluated based on adjusted R^2, Cp and BIC values. Poly(A) tail length in each sample was next scaled by calculated 'sample scaling factors'.

Poly(A) tail length between genes for which intronic reads were detected versus those without intronic reads was compared by subsetting HeLa S3 replicate FLAM-Seq datasets based on subcellular fractions. Poly(A) tail length density distributions and standard deviations were calculated between replicates as described above.

To investigate the molecular features of genes and in context of their poly(A) tail length in subcellular fractions, FLAM-Seq datasets were pooled across all replicates. For analysis of gene

expression by median poly(A) tail length per gene, genes were binned by median poly(A) tail length and boxplots were generated for each bin and subcellular fraction. Similarly, genes were binned by median poly(A) tail length and intronic reads associated with genes in each poly(A) bin were visualized. Half-lives for genes in each bin were calculated similarly based on half-life measurements in Tani et al. 2012. 3'-UTR length was plotted for isoforms binned by median poly(A) tail length for subcellular fractions.

Genes annotated as lncRNAs were extracted based on annotations downloaded from bioMart [357] for human and mouse using gene biotype 'lncRNA'.

Poly(A) tail length was compared between fractions for genes with more than 5 counts in either fraction. Individual gene sets for ribosomal protein genes, IEGs, and lncRNAs were visualized based on annotations as described above. Cytoplasmic-to-nuclear ratios calculated from ENCODE datasets were obtained from Yi et al. 2018. Cytoplasmic-to-nuclear ratios were calculated from FLAM-Seq datasets by summing chromatin and nucleoplasm gene counts and calculating the ratio compared to cytoplasmic expression. ENCODE cytoplasmic-to-nuclear ratios were compared to median poly(A) tail length per gene in each fraction.

3.3.21 CNOT7, PAN3 and PARN shRNA cell lines growth curve quantification and analysis

Image series for different knockdown timepoints and control versus shRNA cell lines were processed using a custom ImageJ / Fiji macro and respective ImageJ functions: For each series of images, input images were converted to 8-bit. Background was removed using `Substract background…` with `rolling=20 light`. `Smooth` function was applied twice to each image, before converting image to binary using `Make binary`. `Fill Holes` was then applied to each binary image and the covered area was quantified running `Measure`. Results for each series was then exported as `.csv` files for each timepoint and contained the area per image covered by cells for each one control and two dox induction series (s. 3.2.19). Covered area was plotted for each image per timepoint and for each series. Differences in total covered area per timepoint were compared for 116 h timepoint using two-sided Student's t-test.

To model grow with a simple exponential model, covered area was log-transformed for each image and a linear model was fit on transformed data using R `lm` function predicting covered area from timepoint, optionally considering dox induction and interaction of timepoint and dox induction as additional parameters.

3.3.22 CNOT7, PAN3 and PARN shRNA subcellular fractionation analysis

FLAM-Seq datasets obtained from biochemical fractionations after doxycycline induction or control from CNOT7, PAN3 and PARN shRNA inducible cell lines were analyzed using the FLAMAnalysis pipeline (s. 3.3.4). For each sample, the cumulative density distribution for median poly(A) tail length per gene was computed and plotted. For comparison of median poly(A) tail length between fractions, replicates for control and dox induction were merged and the difference in median poly(A) tail length per gene was calculated for each subcellular fraction and shRNA cell lines between control and dox induction poly(A) tails.

4 Results

The results of this thesis are structured according to the aims of the work: The first part of the results describes development and validation of FLAM-Seq, a novel method for investigating poly(A) tail length and sequence in context of complete mRNAs. FLAM-Seq was applied for investigating poly(A) tails of mRNA from HeLa S3 and iPS cell lines, organoids, and *C. elegans* samples from larval stage 4 (L4) and adult animals. Poly(A) tail length was investigated in context of mRNA features such as expression level and 3'-UTR isoforms and poly(A) tail nucleotide content and poly(A) dynamics were measured upon inhibiting transcription.

The second part of the results chapter explores poly(A) tail metabolism right after synthesis of poly(A) tails. Poly(A) tails were first investigated in context of unspliced reads, which were computationally identified from FLAM-Seq datasets. The analysis was validated by inhibiting splicing and investigating consequences on poly(A) tail length in Hela S3 cell lines. Metabolic labeling was used to track poly(A) tail length for the first minutes and for up to 3 hours after synthesis. Biochemical fractionations were performed to address poly(A) dynamics across subcellular fractions. The final part explores different options for perturbing RNA expression of deadenylase enzymes by targeted knockdown of CCR4-NOT subunits CNOT7 and CNOT8, the PAN2-PAN3 complex and the deadenylase PARN, using RNA targeting Cas systems, siRNAs and inducible shRNA expression.

4.1 Full-length mRNA and poly(A) tail sequencing (FLAM-Seq)

4.1.1 FLAM-Seq enables quantitative analysis of full-length RNA molecules

Genome-wide determination of poly(A) tail length has so far been performed by using protocols based on short read Illumina sequencing, including TAIL-Seq [227,320] and PAL-Seq [157]. Both methods require interference with Next Generation Sequencing machines and software, since direct analysis of homopolymer sequences, such as poly(A) tails, is technically not possible. PAL-Seq and TAIL-Seq methods are both difficult, time-consuming, and expensive to set up for most potential users. Short read sequencing further limits the possibility to investigate poly(A) tails in context of other gene regulatory elements, such as (alternative) splicing and different 3'-UTR isoforms or transcription start sites. Direct mRNA sequencing using the Nanopore platform has been used to investigate poly(A) tails in context of splicing, yet Nanopore sequencing suffers from higher error rates and requires large quantities of input mRNA (> 100 µg total RNA input) [340]. Both limitations require for a method which combines

a simple library preparation protocol with the ability to PCR-enrich smaller quantities of mRNA which is critical for instance for the potential analysis of clinical samples.

Full-length mRNA and Poly(A) tail sequencing (FLAM-Seq)[358] was devised as a method which enables genome-wide analysis of poly(A) tails and full-length mRNA sequences. The protocol (Figure 4 A, detailed in add methods) combines enzymatic addition of a guanosine / inosine tail (GI-tailing) to the 3'-end of poly(A) selected RNA. The GI-tail is used as a priming site for reverse transcription and applied in combination with a template-switch reverse transcription reaction (TSO-RT [359]) to enrich for full-length cDNA sequences.

Reverse-transcription reactions for cDNA synthesis are typically performed using oligo-dT primers, which prime the reaction at the poly(A) tail for unbiased amplification of polyadenylated transcripts. For inclusion of poly(A) tails into sequencing cDNA libraries, the tail needs to be reverse transcribed along with the mRNA body. Addition of a GI-tail is hence required to introduce a universal priming site independent of the poly(A) tail, which is then used for oligo-dC primed reverse transcription. This concept has also been applied for electrophoretic investigation of poly(A) tail length by gene-specific amplification of GI-tailed samples [297]. To increase specificity polyadenylated transcripts and reduce contamination by rRNAs or other highly expressed, non-polyadenylated RNAs, 3 Ts were added at the 5'-end of the oligo-dC reverse transcription primer (RT primer), which selects for terminally polyadenylated RNAs. The RT primer further contains a 10 nucleotide (nt) unique molecular identifier (UMI) sequence by which PCR duplicates originating from identical cDNAs can be distinguished, along with a PCR handle for amplification.

For enrichment of cDNAs which correspond to full-length mRNAs, a template switch reaction is utilized [359]. Template-switching utilizes the terminal nucleotidy transferase properties of certain reverse transcriptase enzymes such as Moloney Murine Leukaemia Virus (MMLV) RTase to append 3 untemplated cytosines to the cDNA 3'-end upon completion of cDNA synthesis. A short primer containing 3 ribo-guanosines (rG) and a PCR handle (TSO-primer) can hybridize at the overhang, which is then again copied by the RTase (template switch), thereby introducing a PCR handle for PCR amplification. The template-switch reaction can in principle occur multiple times, in particular when little substrate for reverse transcription is available. This can lead to formation of concatemers originating from the TSO-primer, which dominate sequencing libraries (data not shown). To prevent concatemer formation, two isomeric nucleotides were added at the 5'-end of the TSO-primer, which has been shown to

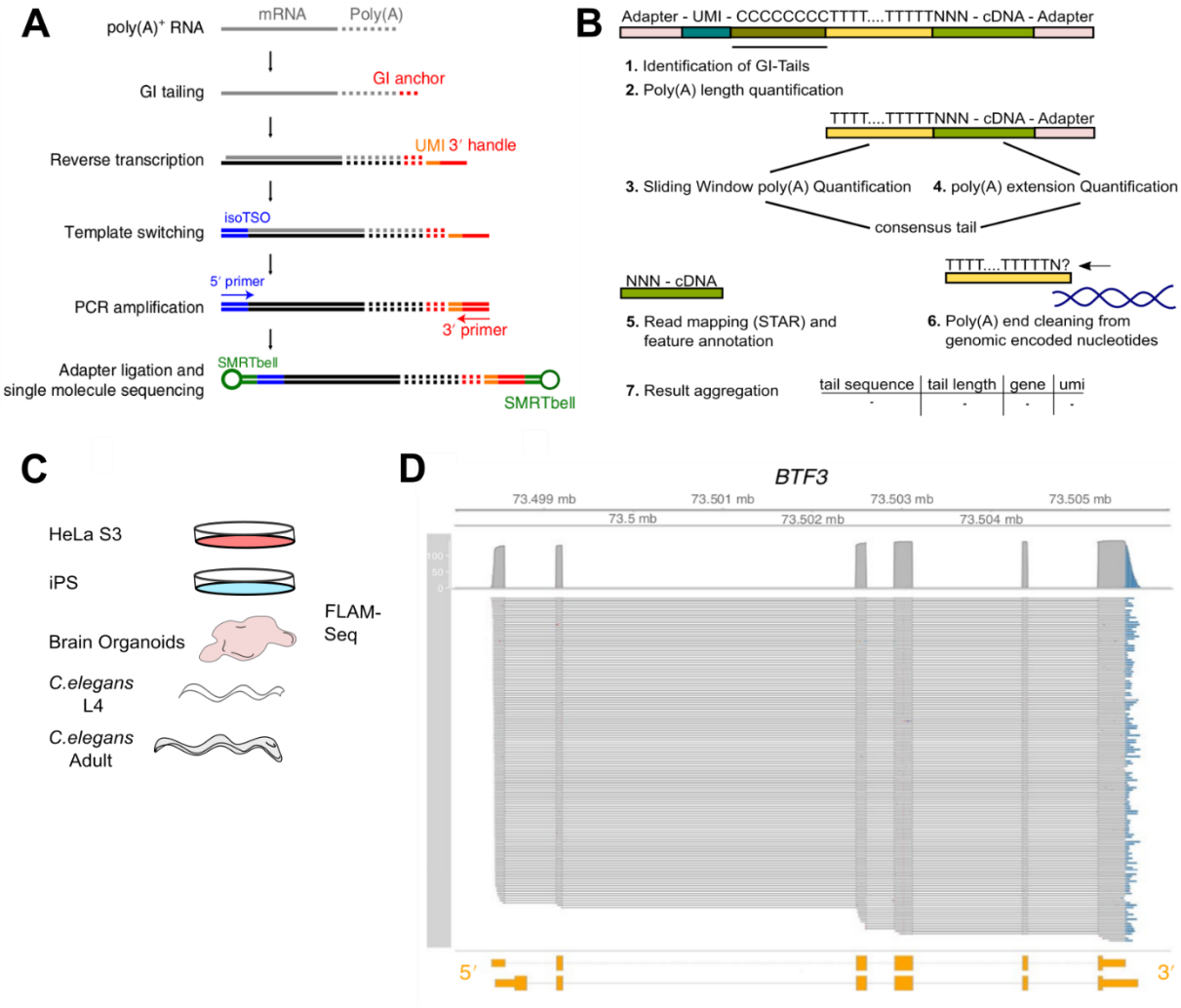

Figure 4 **FLAM-Seq long read sequencing of polyadenylated RNAs from different biological samples**
A) Outline of experimental FLAM-Seq protocol beginning with poly(A) selected RNA which is converted into full-length cDNA libraries. (Adopted from Legnini et al. 2019) **B)** Outline of computational FLAMAnalysis pipeline for processing raw PacBio reads and extraction of gene and poly(A) tail information. **C)** FLAM-Seq libraries were produced from RNA of HeLa S3, iPS, brain organoids, *C. elegans* L4 and adult. **D)** Example browser shot for HeLa FLAM-Seq dataset of BTF3 locus with individual rows corresponding to aligned PacBio sequencing reads. Identified poly(A) tail length for each read are highlighted as blue horizontal bars (Adopted from Legnini et al. 2019).

prevent concatemer formation [360]. This modification drastically reduced the fraction of concatemers in sequencing libraries (data not shown). Full-length cDNA libraries were PCR amplified and sequenced on the PacBio Sequel platform.

FLAM-Seq reads were processed using the FLAMAnalysis pipeline (Figure 4 B, detailed in section 3.3.4). In brief, reads were first filtered for occurrence of GI- and poly(A)-tails. In a next step, poly(A) tails were quantified and clipped from reads along with PCR and sequencing adapters, before mapping reads to the genome using the STAR aligner for long reads [361]. Mapped reads were assigned to genomic features using FeatureCounts [315]. Reverse transcription

can lead to “internal priming” artefacts in cases where the RT primer binds to A-rich regions (or G-rich regions when using an oligo-dC primer) within the transcript bodies [362] or as a consequence of template switching from partly transcribed cDNAs [363]. To eliminate reads that resulted from internal priming, and to remove leftover sequences in poly(A) tails, which were templated and encoded in the genome, reads were compared to genomic sequence on a per-nucleotide basis to precisely distinguish genomic mRNA regions and non-templated poly(A) tail sequences. Poly(A) tail sequence and length was refined in case of remnant nucleotides which were likely encoded by the genome. By the same logic, reads from internal priming were removed since identified poly(A) or A-rich sequences were in those cases templated by the genome. Finally, PCR duplicates were removed based on UMIs, and poly(A) tail length and sequences were identified for each read and corresponding gene.

FLAM-Seq libraries from human HeLa S3 cells, induced pluripotent stem cells (iPSCs), iPSC-derived brain organoids and *C. elegans* adult and L4 stage were initially prepared and sequenced (Figure 4 C, Library preparation by Ivano Legnini, Max Delbruck Center). Alignments from FLAM-Seq datasets can be visualized in a genome browser, including poly(A) tail length, which was appended to alignments using a custom script (Figure 4 D).

Sequencing statistics for the FLAM-Seq protocol were benchmarked against a publicly available PacBio IsoSeq dataset (PacBio IsoSeq UHRR) prepared from human universal reference RNA, which is a standardized mixture of high-quality RNA from 10 human cell lines [364]. IsoSeq is the gold-standard method for long read sequencing of cDNA libraries, but does not enable analysis of poly(A) tails [328]. Between 205,000 and 1,210,000 reads were obtained for HeLa S3, iPSC, organoids and *C. elegans* FLAM-Seq replicates, compared to 63,000 for the IsoSeq UHRR dataset. The average read length for all sequenced samples was 1346 nt compared to 2280 nt for the PacBio IsoSeq UHRR sample (Figure 5 A). Median read length for *C. elegans* samples was slightly shorter than for human samples (1285 nt vs. 1494 nt) which could be caused by on average shorter protein-coding transcripts for *C. elegans* with a median length of 1574 nt [365] compared to 2938 nt for average human mRNAs [140]. Peaks in read length distributions hinted at highly abundant amplicons of identical length in some sequencing libraries, which could be caused by few highly expressed or overamplified amplicons. Peaks disappeared after processing reads through the FLAMAnalysis pipeline which filters for poly(A) tails and removes PCR duplicates. A poly(A) tail sequence could be detected for on averge 38% of input reads. Of those poly(A) reads, around 90% could be uniquely mapped to the human or *C. elegans* genome and 83% of mapped reads could be assigned to a uniquely annotated gene locus. In total, around 28% of raw reads were retained for downstream analysis

after extracting poly(A) tails, mapping and removal of duplicates (Figure 5 B), with little variation between samples. One exception was organoid replicate 1 sample, which contained a large fraction of short amplicons which may hint at a low library quality.

Between 8344 and 13168 genes were detected for each human dataset and between 6067 and 6874 in *C. elegans*. As a comparison, 12626 genes were detected in the IsoSeq UHRR dataset. A median of 2–6 unique molecules (UMIs) per gene were detected for FLAM-Seq samples (Figure 5 C), and a median of 3 reads per gene for IsoSeq, although this comparison is biased since IsoSeq does not incorporate unique molecular identifiers (UMIs) which allow for removal of PCR duplicates. In general, PCR overamplification appears as a less important factor given that the fraction of mapped and assigned reads (28.3% of total reads) is very similar to the fraction of usable reads after PCR collapse (28.0%).

Pairwise correlations of gene expression counts between replicates indicated good agreement between biological replicates with a Pearson correlation coefficient between r = 0.76 and r = 0.82 for human samples and r = 0.79 for *C. elegans* replicates (Figure 5 D).

Read 5'-ends were mapped to annotated transcription start sites (TSS) to estimate the fraction of FLAM-Seq reads which span full-length transcripts. Human samples (Figure 5 E) were compared to TSS annotated in the FANTOM5 database [351], which are based on CAGE peaks, a method for sequencing of capped RNA 5'-ends [366]. Since the FANTOM5 project is limited to human and mouse annotations, *C. elegans* read starts were compared to a 5'-SAGE based annotation (Figure 5 F) which is a similar method for analysis RNA 5'-ends by next-generation sequencing [352]. The analysis of transcription start sites in *C. elegans* is yet more challenging as most mature mRNAs undergo trans-splicing, by which the first exon is spliced to a splice leader sequence [367,368]. For human samples we detected a median of 62% of reads per gene mapping to annotated transcription start sites for HeLa S3 and iPSC datasets and a median of 40% for organoid datasets, which may be an effect of generally longer 3'-UTR isoforms and mRNA transcripts in neuronal systems [101]. For longer IsoSeq reads, around 72% of all reads mapped to annotated TSS, which shows that read length impacts the fraction of reads reaching the 5'-end of a transcripts but also shows that a number of 5'-ends may be missing in the FANTOM annotation. To validate specificity of reads starting at TSS, nucleotides were *in silico* clipped from the 5'-read start and mapped to FANTOM TSS annotations. As expected, the fraction of assignable reads dropped proportional to the number of clipped nucleotides and

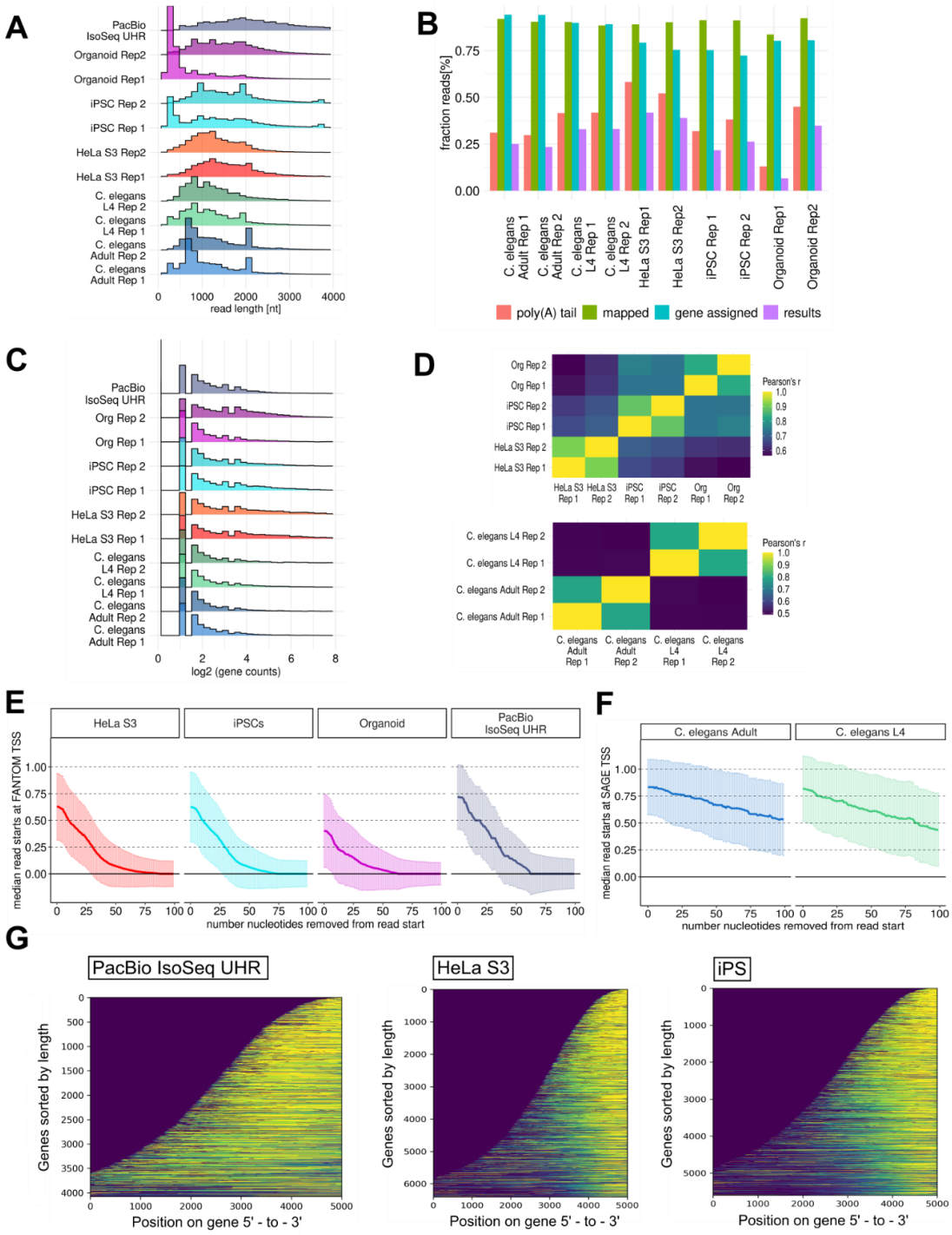

Figure 5 **FLAM-Seq quantification of gene expression and transcript coverage**
A) Read length distribution for sequenced FLAM-Seq samples and PacBio IsoSeq UHR reference dataset
B) FLAMAnalysis pipeline processing statistics. Indicated are the fraction of input reads with identified poly(A) tail, fraction mapped reads ('mapped'), fraction of reads assignable to annotated genes ('gene assigned') and the fraction of reads retained after complete annotation and filtering ('results') **C)** Histograms of expression counts per gene for each FLAM-Seq sample. **D)** Pairwise correlations in gene expression for human (left) and *C. elegans* (right) FLAM-Seq samples. **E)** Median fraction of human FLAM-Seq reads aligning to annotated FANTOM transcription start site (TSS) per gene after clipping indicated number of nucleotides from read start. Error bars indicate one standard deviation across genes. **F)** Median number of *C. elegans* FLAM-Seq read starts aligning to annotated SAGE transcription start site (TSS) per gene after clipping indicated number of nucleotides from read start error bars indicate one standard deviation across genes. **G)** Normalized sequencing coverage for across exons of detected genes. Genes were sorted by length. Bright yellow indicates high coverage at respective positions in genes.

converged to 0%. Both for *C. elegans* adult and L4 stages, around 82% of reads could be mapped to annotated TSS. The increased fraction of assignable reads for *C. elegans* could be caused by on average shorter transcripts for *C. elegans*, although clipping of reads showed that mapping to annotated TSS was not fully abolished even when 100 nt were clipped from the read start, which might indicate artefacts in SAGE TSS annotations.

For each sample the relative coverage over each gene was calculated to assess biases in coverage for longer transcripts (Figure 5 G, illustrated for IsoSeq, HeLa S3 and iPSC samples). We noticed a drop in coverage in transcript positions more than ca. 1500 bp away from the 3'-end, which is expected given the average read length of ca. 1350 nt. Longer IsoSeq UHR reads accordingly produced higher coverage towards transcript 5'-ends.

In summary, FLAM-Seq is quantitative method for genome-wide sequencing of full-length cDNAs and biased in coverage only for very long transcripts.

4.1.2 FLAM-Seq accurately quantifies genome-wide poly(A) tail length profiles

FLAM-Seq enables quantification of full-length mRNAs including the length and sequence of associated poly(A) tails. To validate the accurate quantification of poly(A) tails, FLAM-Seq sequencing libraries were prepared from synthetic cDNA standards with known poly(A) tail length (synthetic standards library preparation by Ivano Legnini, Max Delbruck Center). cDNA standards comprised 4 chemically synthesized single stranded DNA standards with poly(A) tail length ranging from 30 to 120 nt. A RNA standard was prepared by ligation of a chemically synthesized 50 nt poly(A) sequence with an *in vitro* transcribed RNA of 200 nt in length by splint ligation (Ivano Legnini, MDC) (Figure 6 A). The RNA standard was added to address potential biases in poly(A) tail length and sequence analysis which may have been caused by reverse transcription or tailing.

Poly(A) tail length for each read were quantified using the FLAMAnalysis pipeline. Median poly(A) tail length for each standard followed a normal distribution with one distinct mode (Figure 6 B). For cDNA standards, median poly(A) tail length per standard was in each case slightly shorter than expected with a median of 28 nt for the 30 nt poly(A) standard, 57 nt for the nt 60 poly(A), 85 nt for 90 nt poly(A) and 114 nt for the 120 nt poly(A) standard. The RNA standard had a median sequenced poly(A) tail length of 51 nt. The standard deviation of poly(A) tail length for each cDNA standard was calculated and a slight trend towards larger standard deviations was observed for standards with longer poly(A) tails (Figure 6 C). Of note, a second minor mode for the 90 nt poly(A) tail standard was observed at around 60 nt, which inflated the

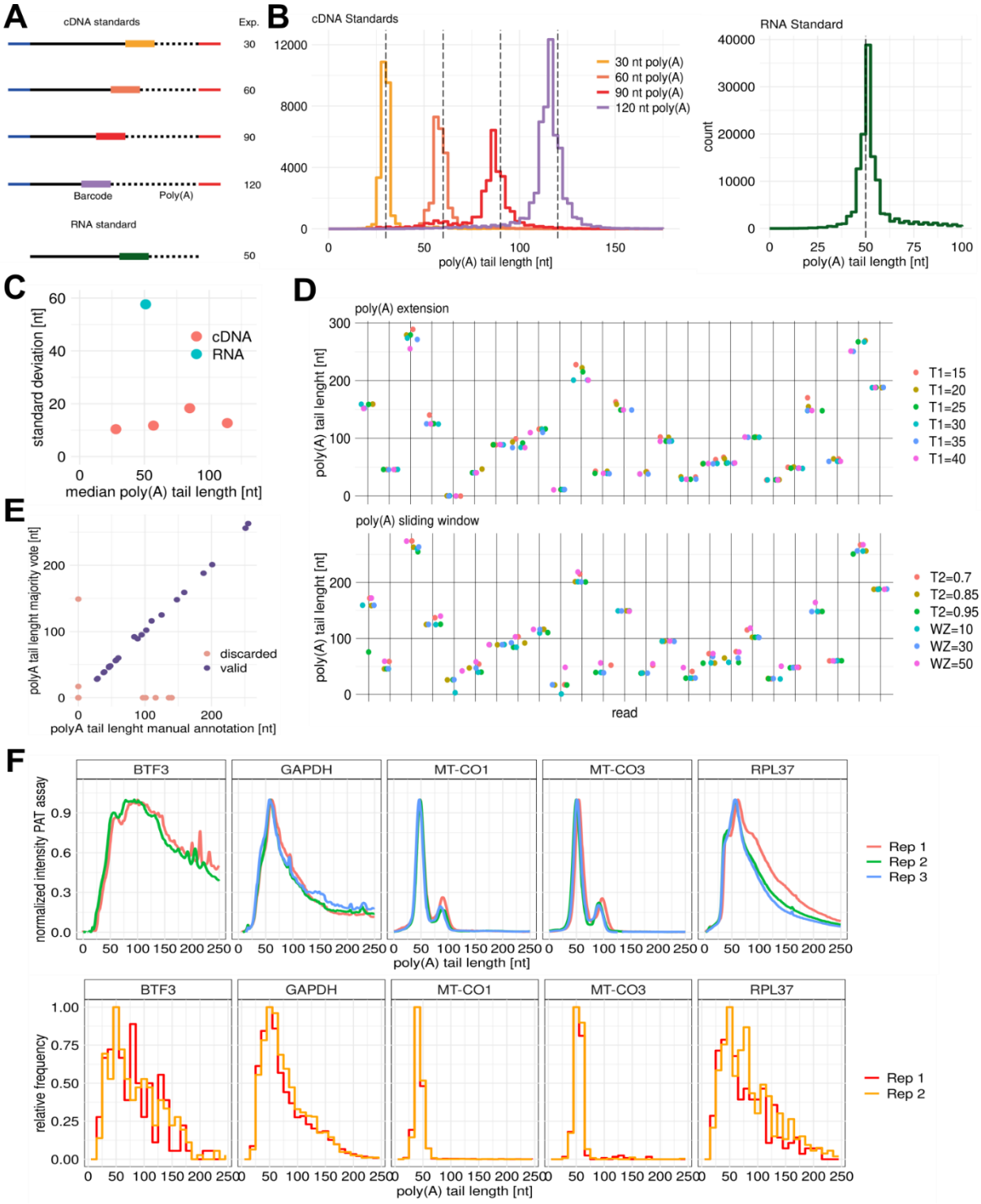

Figure 6 **FLAM-Seq benchmark for quantifying poly(A) tail length from PacBio sequencing**
A) Design of synthetic oligonucleotide standards for validation of accurate poly(A) tail length estimation by FLAM-Seq. (Adopted from Legnini et al. 2019) **B)** Poly(A) tail length estimated from FLAM-Seq sequencing data for four synthetic cDNA standards (left) and RNA standard (right). Dashed lines indicate the expected poly(A) tail length for each standard. **C)** Median poly(A) tail length versus standard deviation for synthetic poly(A) standards. **D)** poly(A) tail length estimates for 25 FLAM-Seq reads using (top) 'poly(A) extension' algorithm with threshold parameter $T_1 = 15$ to $T_1 = 40$, (bottom) 'poly(A) sliding window' with threshold parameter $T_2 = 0.7$ to $T_2 = 0.95$ and window size $WZ = 10$ to $WZ = 50$. **E)** Poly(A) tail length identified by manual annotation versus poly(A) tail length from majority vote algorithm. Reads discarded by FLAMAnalysis pipeline are highlighted. **F)** Poly(A) tail length profiles measured for individual genes by PAT assay (top) and by FLAM-Seq (bottom) in HeLa S3 cell lines.

standard deviation measured for this standard. The standard deviation for the RNA standard was 5 to 6 times higher than for the cDNA standards, which was mostly caused by long poly(A) tails which could have resulted from ligation of multiple poly(A) 50-mer tails to one RNA adapter when producing the RNA standard.

Two algorithmic approaches were used in combination to extract and quantify poly(A) tail length by the FLAMAnalysis pipeline. To understand the behavior of each algorithm in quantifying poly(A) tail length, 50 reads were randomly sampled from HeLa S3 sequencing datasets and processed with both algorithms and different parameter combinations (Figure 6 D). The first 'poly(A) extension' algortihm quantifies the poly(A) tail by matching a short poly(A) substring and itertatively extending the poly(A) tail while increasing allowed mismatches to account for sequencing errors and possible non-A nucleotides in poly(A) tails. A parameter T_1 specifies the allowed mismatches (Figure 6 D Top). The second algorithm uses a sliding window which identified the poly(A) tail by relative adenosine nucleotide content within a specified window of size WZ and a minimum adenosine fraction of T_2 (Figure 6 D Bottom). The 'poly(A) tail extension' algorithm was more conservative in estimating poly(A) tail length and discarded many reads where no poly(A) seed could be identified. A minimum seed length of 10 nt also defined the shortest possible poly(A) tail length. The 'sliding window' approach identified similar poly(A) tail length as the 'extension' algorithm for almost all reads. Modulating threshold and window size parameters had small effects on poly(A) tail length estimates, which was reassuring that both algorithms produce robust poly(A) tail length estimates that were stable for different parameter combinations. For optimal performance, both algorithms were combined in the FLAMAnalysis and reported poly(A) tail for each read was determined by majority vote.

Poly(A) tail length for the sample of HeLa S3 FLAM-Seq reads was also manually annotated and plotted against the FLAMAnalysis estimates (Figure 6 E). All reads which were considered valid by the FLAMAnalysis pipeline were found on the diagonal showing excellent agreement between manual length assignment and the FLAMAnalysis pipeline. Few reads were found where no poly(A) tail was detected by the FLAMAnalysis pipeline. Manual inspection showed that these reads had valid poly(A) tails but errors in the adapter sequences and were hence removed from downstream analysis. In other cases, reads appeared concatenated, such that multiple putative poly(A) tail sequences were present. Those poly(A) tails could be quantified by the FLAMAnalysis pipeline, but the concatenated reads were later discarded in the mapping steps.

Poly(A) tail length profiles measured by FLAM-Seq for HeLa S3 cell were compared to PAT assay measurements, which quantified poly(A) tail length profiles by electrophoresis for individual genes (Figure 6 F). PAT assays were performed by adding a GI-tail to total RNA and using the GI-tail as priming site for reverse transcription, such that the complete poly(A) tail sequence is preserved. Gene-specific primers were then used for PCR amplification of poly(A) tail distributions of individual genes. PAT assays were performed for BTF3, GAPDH, MT-CO1, MT-CO3 and RPL37 genes in triplicates in RNA extracted from HeLa S3 cell lines. The obtained poly(A) tail length distributions were highly reproducible between replicates and in good agreement with poly(A) tail length profiles measured by FLAM-Seq. BTF3 profiles determined by PAT assay were slightly longer than profiles measured by FLAM-Seq and for mitochondrial genes MT-CO1 and MT-CO3 an additional peak at ca. 100 nt was observed. This second mode may result from unspecific RNA amplification. mRNA poly(A) tails of genes encoded by the mitochondrial genome had a relatively uniform length of around 50 nt [151], which was validated by FLAM-Seq and PAT assay for two mitochondrial genes.

In summary, FLAM Seq enabled accurate quantification of poly(A) tails from long-read sequencing data and accurate recovery of poly(A) tail length profiles for different genes.

Poly(A) tail length profiles were quantified for HeLa S3 and iPS cells, organoids, and *C. elegans* adult worms and L4 larval stage and between 59,000 and 290,000 RNA individual molecules could be analyzed for each sample. Poly(A) tail length profiles over all sequenced polyadenylated RNAs showed major differences between model systems (Figure 7 A). HeLa S3 cells had a median bulk poly(A) tail length of 82 nt, which was shorter than the medians for iPS cells (102 nt) and organoids (110 nt). *C. elegans* L4 and adult samples had markedly shorter tails than human model systems with a median length of 54 nt for L4 stage and 47 nt for adult (Figure 7 A). Each poly(A) distribution had its mode at around 50 nt independent of the biological sample. Poly(A) tail length of mRNA encoded by the mitochondrial genome had a median poly(A) tail length of 42 nt – 47 nt for human samples and 35 nt-36 nt for *C. elegans* samples, which is in line with PAT assay poly(A) quantification and previously reported tail length for human cell lines [151]. The longest detected poly(A) tail had a length of 1131 nt and was assigned to the ACOT8 gene. The shortest detected tails were 10 nt in length, which is also the lower limit defined by the FLAMAnalysis processing steps for identification of poly(A) tails. Replicate distributions were in good agreement for all analyzed samples. Organoid replicate 2 had an increased fraction of long tails compared to replicate 1, which is likely a consequence of the three-fold difference in sequencing depth or accumulation of short reads for replicate 1.

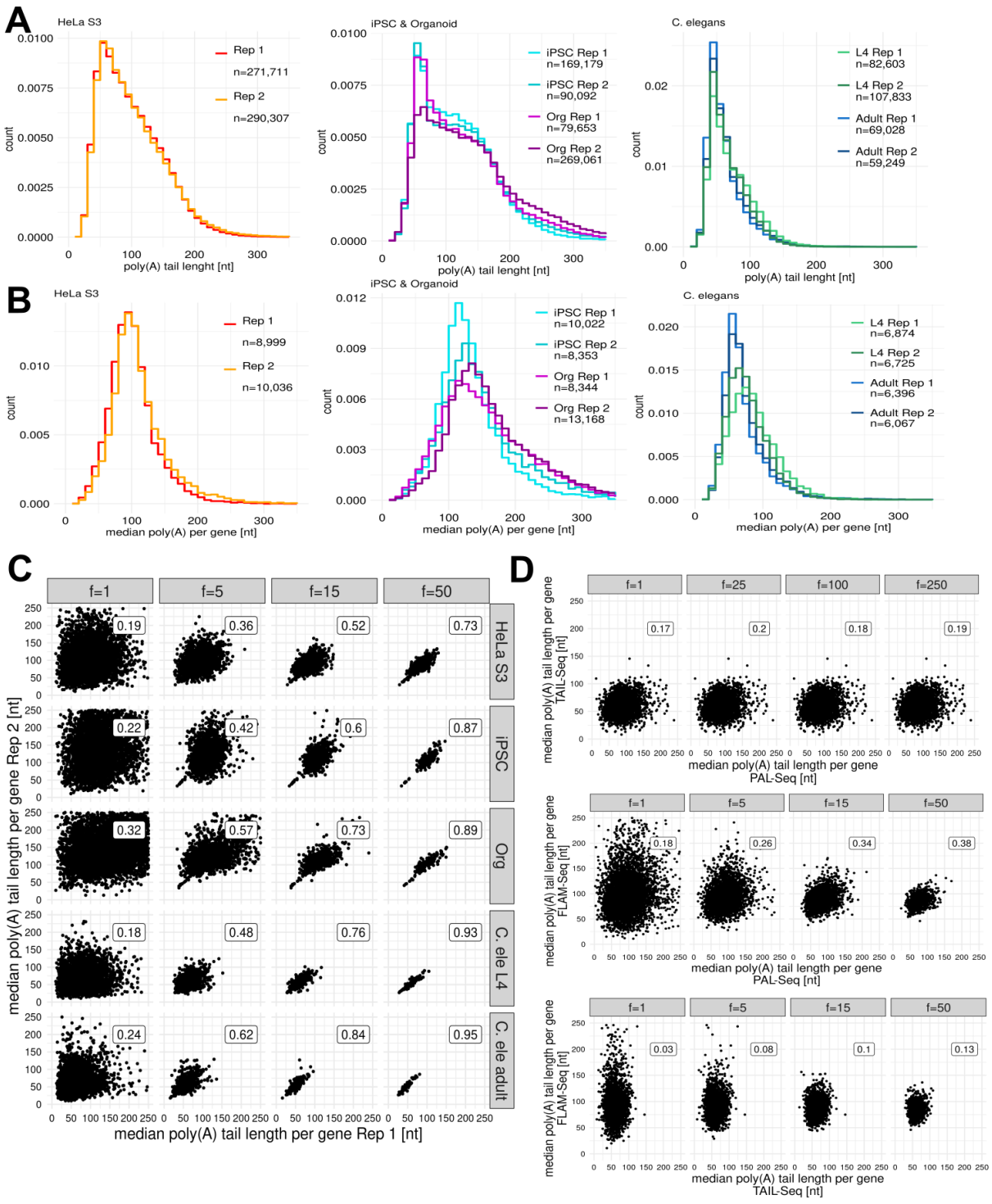

Figure 7 **Poly(A) tail length profiles of HeLa S3, iPSC and organoids show model system specific differences** **A)** Poly(A) tail length per sequenced RNA molecule for HeLa S3 cells, iPS cells, organoids, and *C. elegans* L4 and adult stages. Numbers in legend indicate number of reads per replicate dataset. **B)** Median poly(A) tail length per gene for HeLa S3 cell, iPS cells, organoids, and *C. elegans* L4 and adult stages. Numbers in legend indicate detected genes per dataset. **C)** Scatterplots for median poly(A) tail length measurements per gene between replicates for HeLa S3, iPSC, organoid, *C. elegans* L4 and adult stage. Columns indicate the minimum required counts for each gene to be retained in analysis (cutoff *f*=1, *f*=5, *f*=15, *f*=50). Number in boxes denote Pearson correlation coefficients. **D)** Scatterplots for median poly(A) tail length per gene for HeLa cell line comparing results from PAL-Seq versus TAIL-Seq (top), PAL-Seq versus FLAM-Seq (middle) and TAIL-Seq versus FLAM-Seq (bottom). Columns indicate minimum counts for each gene to be retained in analysis (*f*=1-*f*=50).

Aggregating bulk poly(A) profiles into median poly(A) tail length profiles per gene indicated that median profiles are longer than bulk distributions for all analyzed samples (Figure 7 B), which in turn required that genes with lower expression levels have on average longer poly(A) tails.

Between 6,000 and 13,200 genes were detected per replicate, highlighting the ability of FLAM-Seq to investigate polyadenylation on a genome-wide scale. Median poly(A) tail length per gene was 92 nt for HeLa S3 cells, 122 nt for iPS cells, 140 nt for organoids, 68 nt for *C. elegans* L4 and 55 nt for adult stage. Comparing iPS and organoids and *C. elegans* L4 and adult poly(A) profiles further highlighted the developmental dynamics of poly(A) tail length regulation. Comparing poly(A) tail length per gene between replicates showed good agreement between poly(A) tail length measurements upon filtering for genes with sufficient counts (Figure 7 C). Pearson correlations were modestly positive (r = 0.18–0.32) taking account all genes expressed in both replicates (cutoff f = 1) and independent of the sample. Correlation coefficients yet increased when setting higher expression thresholds up to r = 0.73–0.95 for genes with more than 50 counts (f = 50). Higher gene expression more accurately estimated medians from poly(A) tail distributions for each gene and the analysis showed that uncertainty in median tail length between replicates was mostly related to sampling and sequencing depth.

PAL-Seq [157] and TAIL-seq [227] protocols were developed as alternative methods for quantifying poly(A) tail length using the Illumina sequencing platform and both methods were also applied to analyze HeLa cell line poly(A) tail profiles. Subtelny and colleagues report an average median poly(A) tail length per gene of 83 nt using PAL-Seq, while Chang et al. measure an average of 60 nt per tail, which is in both cases shorter that the average tail profiles measured by FLAM-Seq which was 92 nt. Comparing median poly(A) tail length estimates per gene between FLAM-Seq, PAL-Seq and TAIL-Seq showed modest positive correlations which slightly increased when excluding lowly expressed genes (Figure 7 D). TAIL-Seq and PAL-Seq poly(A) tail length per gene had a correlation coefficient of r = 0.17-0.21 also when only including highly expressed genes. FLAM-Seq and TAIL-Seq had correlation coefficients between r = 0.03 and r = 0.13, while comparing FLAM-Seq and PAL-Seq showed better agreements in poly(A) estimates for higher expressed genes (r = 0.38). The larger overall poly(A) tail length per gene (83 nt PAL-Seq, 92 nt FLAM-Seq) went along with a more reproducible gene-wise poly(A) quantification between PAL-Seq and FLAM-Seq. In summary FLAM-Seq revealed species and developmental stage dependent poly(A) tail length profiles which were reproducible between replicates, while increasing sequencing depth improved the reproducibility in estimating median poly(A) tail length between replicates.

4.1.3 Highly expressed, stable housekeeping genes have short poly(A) tails

Bulk poly(A) tail length distributions were on average shorter than median poly(A) tail length per gene, which implied that highly expressed genes, which dominated bulk poly(A) distributions, had shorter poly(A) tails than lower expressed genes. To further explore this relationship, gene expression was compared to median poly(A) tail length per gene for HeLa S3, iPS cells, organoids and *C. elegans* adult and L4 stage (Figure 8 A). In all cases, gene expression was negatively correlated with median poly(A) tail length per gene with comparable Pearson correlation of r = -0.17 to r = -0.27, which was in all cases highly significant. Slope coefficients for linear regressions were larger for organoid (a = -9.8) and iPS samples (a = -7.7) compared to HeLa S3 cell lines (a = -3.5), which may indicate a larger dynamic range covered by iPS/organoid poly(A) tails compared to HeLa that may be associated with gene expression.

A similar weak but significant negative correlation was found between HeLa S3 mRNA half-life and median poly(A) tail length (Figure 8 B) as well as translational efficiency and median poly(A) tail length per gene (Figure 8 C), which is expected given that highly expressed genes are typically more stable.

To find molecular and/or biological processes which are associated with genes having either long or short steady state poly(A) tails, a Gene Ontology (GO) term enrichment analysis was performed for genes grouped in a ‘short’ or ‘long’ poly(A) bin (Figure 8 D). Gene-poly(A) length bins were defined by median poly(A) tail length below the first (‘short’) or above the third quartile (‘long’) of the median poly(A) tail length distribution. GO term associations of genes grouped in each bin were then calculated. Genes with short poly(A) tails tended to be associated with housekeeping functions such as ‘translation initiation’ or ‘protein targeting to ER’. Long poly(A) tails were on the other hand associated with developmental functions (‘cell development’).

In summary, genes with short poly(A) tails tend to be higher expressed, more stable, more efficiently translated and associated with housekeeping functions.

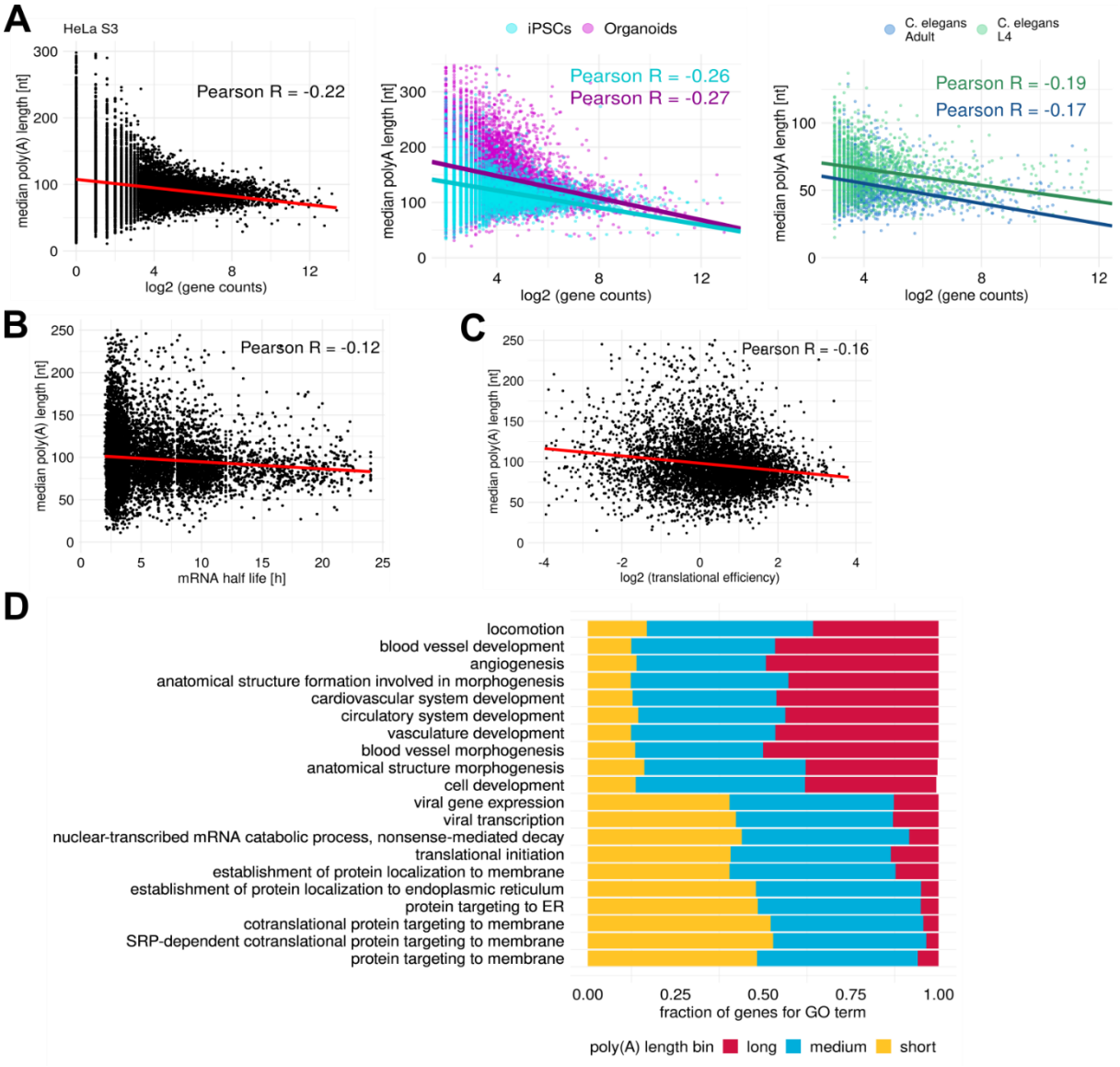

Figure 8 **Correlation of poly(A) tail length with expression, RNA stability and translational efficiency A)** Gene expression versus median poly(A) tail length for HeLa S3 (left), iPS cells and organoids (middle), *C. elegans* L4 and adult (right). **B)** mRNA half life (Tani et al. 2012) versus median poly(A) length per gene for HeLa S3 cell line. **C)** mRNA translation efficiency (Subtelny et al. 2014) versus median poly(A) tail length per gene for HeLa S3 cell line. **D)** GO term enrichment for HeLa S3 sample for genes binned by 'long', 'medium' and 'short' poly(A) tail bins. Shown are top 10 GO terms with the strongest association to short or long median poly(A) tails per gene. For each GO term the proportion of genes which are associated with this GO term grouped into long, medium and short poly(A) bins are shown.

4.1.4 Statistical modeling of differences in poly(A) tail length distributions

Comparison of poly(A) tail length distributions between samples required thorough understanding of the statistical properties since distributions varied in shape and spread, ranging from relatively sharp and centered normal distributions as observed for mitochondrial genes to broad, highly skewed distributions as for GAPDH (Figure 6 F). Comparing gene expression counts to standard deviations of poly(A) tail length in HeLa S3 samples showed that on average the standard deviation for genes encoded by nuclear DNA was slightly decreasing towards

shorter poly(A) tails (Figure 9 A). The residuals of standard deviation for each gene against a linear fit also decreased towards higher expressed genes, which is expected given that a larger number of sampled poly(A) tails per gene should improve robustness of the median estimators. Genes encoded by mitochondrial DNA on the contrary had much smaller standard deviations, in line with the observation that mitochondrial poly(A) tail length profiles were centered around 50 nt.

Comparing poly(A) tail length standard deviations with median poly(A) tail length per gene revealed a trend towards increased standard deviations for longer median poly(A) tail length (Figure 9 B). This trend was already observed when analyzing cDNA standards poly(A) tail length along with their respective standard deviations (Figure 6 C). The slopes of a linear regression model fitted to standard deviations as a function of poly(A) tail length for all genes or for cDNA standards only revealed a stronger increase in variability for longer tails than expected by cDNA standards (a = 0.15 all genes, a = 0.05 for cDNA standards). This also hinted at increased biological variability for longer poly(A) profiles, beyond the increased technical uncertainty in quantifying longer tails observed for cDNA standards. The average standard deviation was 49 nt across all genes.

Assessing the statistical significance of differences between poly(A) tail length distributions was challenging through non-uniform and broad spreads of poly(A) profiles for individual genes. To identify under what conditions statistical tests had sufficient power, different simulations were performed which model each two poly(A) tail length distributions under the following assumptions: Each distribution came from a Gaussian with a standard deviation corresponding to 49 nt, which was the average across all genes. Gaussian distribution means were defined by the ‘basal poly(A) tail length’, which was the center of the shorter poly(A) distribution (Figure 9 C), while the longer poly(A) distribution was defined by addition of 0 to 50 nt (‘simulated difference in medians’) to ‘basal’ poly(A) distributions. Different ‘basal poly(A) tail length’ values were used, taking into account the poly(A) tail length dependent increase in poly(A) quantification error. Comparisons were performed by simulating different sample sizes (‘gene counts’) ranging from 5 to 100 counts per distribution to test sampling (‘sequencing depth’) related effects. For each set of parameters, 100 samples per parameter combination were simulated and the resulting distributions were tested for differences using three statistical models:

First, poly(A) tail length distributions were resampled taking into account the technical error of each poly(A) measurement. Individual measured poly(A) tails were resampled according to a

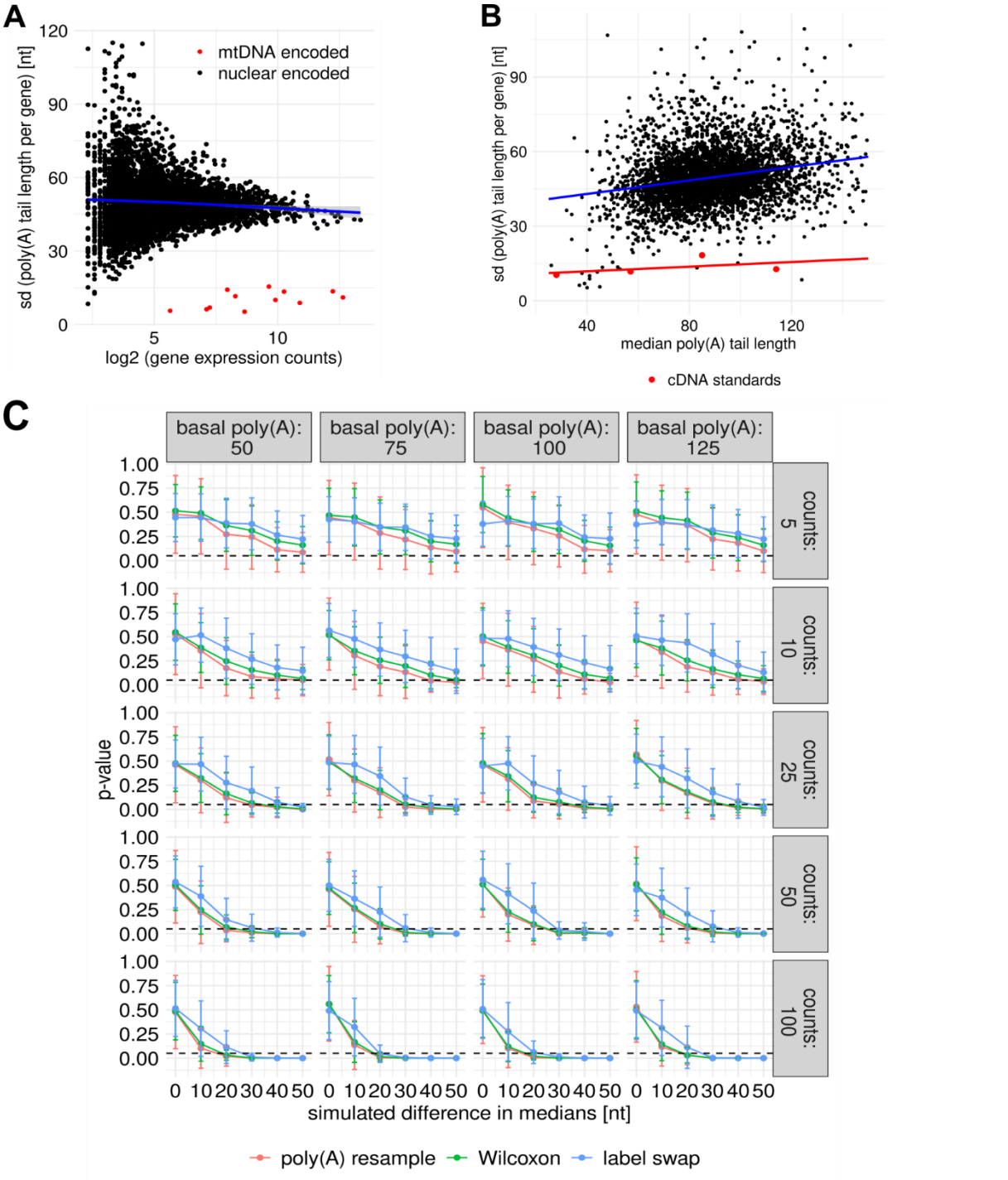

Figure 9 **Statistical modeling of poly(A) tail length differences**

A) HeLa S3 gene counts and standard deviation of associated poly(A) tail length distributions for genes encoded by nuclear (black) and mitochondrial DNA (red). Line denotes LOESS fit. **B)** HeLa S3 median poly(A) tail length per gene and standard deviation for all expressed HeLa genes (black) and synthetic cDNA standards (red). Lines denote linear regression fits to all genes (blue line) and cDNA standards (red line). **C)** Simulation of statistical testing power for differences in poly(A) tail length distributions given the number of counts for each poly(A) distributions (counts, rows), the median poly(A) tail length of the shorter poly(A) distribution ('basal poly(A)', columns). Average simulated p-values are plotted against the difference in median for poly(A) tail length distributions for three tested statistical models. Error bars denote standard deviations in p-values after 100 simulations.

Gaussian centered around the measured poly(A) length with a standard deviation inferred from cDNA standards. A p-value was calculated by comparing the original to resampled distributions after resampling 1000 times. As a second model, a non-parametric Wilcoxon test was used for assessing differences in poly(A) distributions. Third, a 'label swap' test was designed in which labels of simulated distributions were shuffled between the shorter and longer distribution and p-values were calculated by comparing the median of the original to reshuffled distributions. p-value distributions were then compared for different parameters to identify how many counts per gene were required and for which differences in means changes in poly(A) tails distributions could be reliably detected (Figure 9 C). Significant differences were assumed when simulated p-values were on average below 0.05 (significance level). The total number of poly(A) tails for each distribution, which corresponds to the sequencing depth, appeared as one important factor: Below 10 counts per gene, none of the applied statistical models was able to identify distributions as significantly different. Above 10 counts this was possible, although only median poly(A) tail length differences of 20-30 nt and above could be identified as significantly different, which provided an estimate of the required effect sizes. Differences in 'basal poly(A)' parameter did not appear to greatly impact assessment of significance, indicating that differences in variability for quantifying longer tails were less relevant. Comparing the three statistical models showed that the 'poly(A) resampling' approach was most sensitive in detecting differences in poly(A) tail lengths, before the Wilcoxon test and 'label swap' approach. Differences between models became most apparent for low poly(A) counts. All methods reported similar p-values for 0 difference control, while here the standard deviation was highest for the 'poly(A) resampling' method.

Since FLAM-Seq datasets typically had few counts per gene (32% – 68% of genes < 5 counts in merged FLAM-Seq replicates), the 'poly(A) resampling' method was used. Poly(A) tail length distributions were compared between iPS cells and organoids (Figure 10 A), as well as *C. elegans* L4 and adult samples (Figure 10 B) which each corresponded to different developmental stages of the two biological systems. 27 genes were detected with significantly longer poly(A) tails in iPS cells than in organoids after multiple hypothesis testing correction (1% of all genes) while 264 genes had longer poly(A) tails in organoids (9% of all genes), which was expected given the differences in global poly(A) distributions per gene (Figure 6 B). The maximum difference in poly(A) tail length was observed for CTTN gene which was 201 nt longer in organoids. NUFIP2 genes on the contrary was 107 nt longer in iPS cells. Organoid genes not detected in iPS cells has overall longer median poly(A) profiles of 148 nt compared to 130 nt for all organoid genes, which hinted at increased poly(A) tail length of genes induced

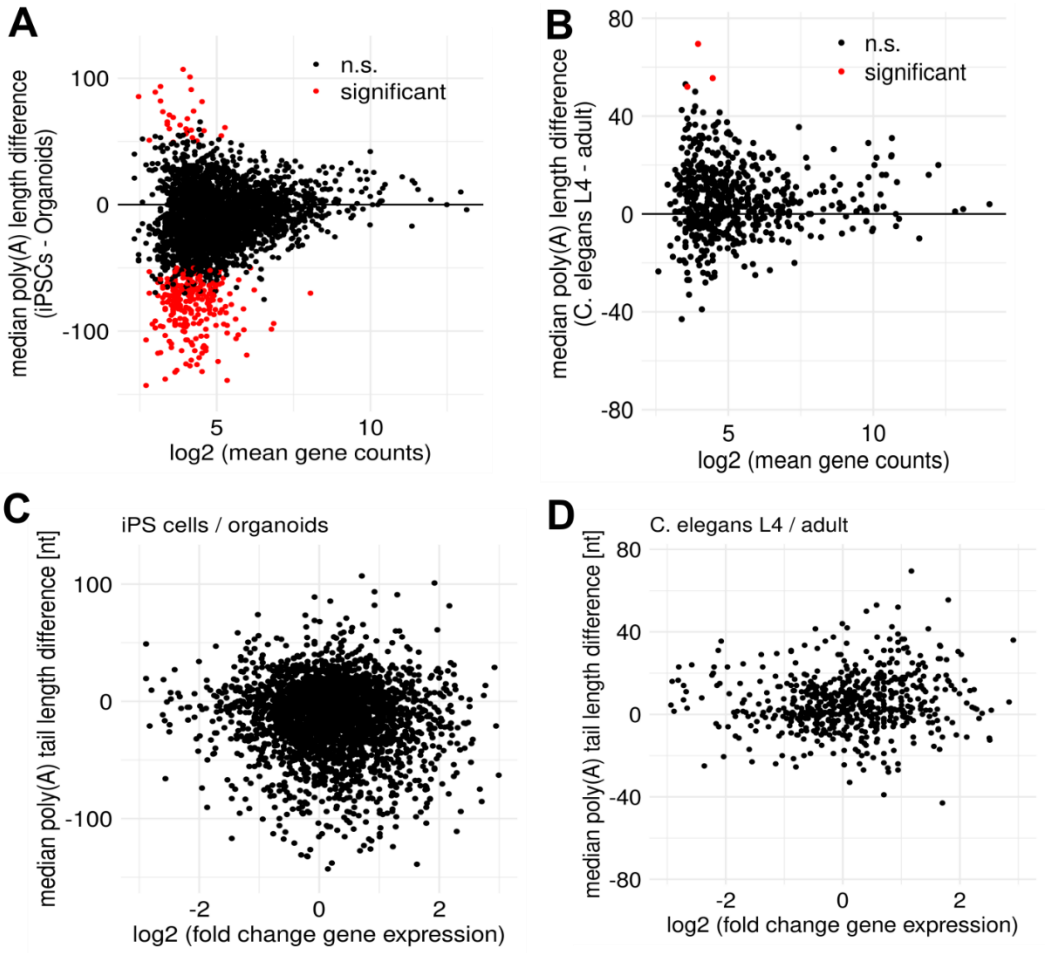

Figure 10 **Poly(A) tail length differences per gene between developmental stages**
A) Median poly(A) tail length difference between organoids and iPS cells (iPSCs) versus mean expression counts between samples (n.s.: non-significant). **B)** Median poly(A) tail length difference per gene between *C. elegans* L4 and adult versus mean gene expression (n.s.: non-significant). **C)** Median poly(A) tail length difference versus gene expression fold change for iPS cells and organoids (poly(A) difference iPSC – organoids, expression fold change iPSC/organoid). **D)** Median poly(A) tail length difference versus gene expression fold change for *C. elegans* L4 and adult samples. (poly(A) difference L4 – adult, expression fold change L4/adult).

during neuronal development. Only 3 genes were detected in *C. elegans* samples with statistically significant differences in median poly(A) tail length per gene (Figure 10 B).

Comparing global median poly(A) tail differences per gene showed overall shorter poly(A) tails in L4 samples (Fig 3 B), which could not be resolved on the basis of individual genes given the sensitivity of the FLAM-Seq analysis. Differences in poly(A) tail length were not correlated to differences in gene expression between iPS and organoids (Figure 10 C) or L4 and adult stages (Figure 10 D) with Pearson correlation coefficients of r = -0.08 and r = 0.06 respectively.

4.1.5 Precise annotation of 3'-UTR isoforms uncovers elements of polyadenylation regulation

FLAM-Seq reads were for each sample on average more than 1000 nt in length, which enabled computational reconstruction of precise RNA isoforms. Different RNA isoforms, which can be the product of alternative polyadenylation, can have drastically different properties with respect to RNA stability, localization and translation [72], which makes their genome-wide investigation in context of poly(A) tails highly relevant.

Manual inspection of FLAM-Seq alignments in a genome browser revealed that in many cases alternative RNA 3'-ends are clearly detectable (examples in Figure 11 A). Yet, the gene model annotations coming from Gencode or RefSeq databases (Figure 11 A bottom) did not provide a precise definition of polyadenylation sites for FLAM-Seq data of human samples. Intron and splice site annotations were on the other hand more accurate (as judged by manual inspection) and agreed better between Gencode gene models and the inspected FLAM-Seq alignments.

Since Gencode or Refseq did not provide a sufficient basis for mapping of 3'-UTR isoforms, 3'-UTR ends for each RNA isoform were annotated *de novo* from FLAM-Seq alignments (s. Comp. methods). For HeLa S3 samples, 4821 3'-UTR isoforms were detected from 3698 expressed genes, for iPS cells 3521 isoforms were detected for 2788 genes, while this number was higher for organoid datasets with 5347 different isoforms for 4261 genes. Alternative polyadenylation in *C. elegans* was less extensive, with 422 and 335 isoforms detected from 395 and 319 genes from L4 and adult samples.

To validate the sequence composition around the cleavage and polyadenylation site, the nucleotide distributions around the end of each read and beginning of the poly(A) tail sequence were calculated (Figure 11 B). As expected, downstream of the cleavage site the sequence content was almost exclusively adenosines within the identified poly(A) tail sequence. Upstream of the cleavage site, the sequence composition was mostly AU-rich, with a U-rich sequence stretch preceded by an A-rich stretch which likely captured the polyadenylation signal AAUAAA occurring around the -20 position from the cleavage site [66]. CPSF73 is reported to preferentially cleave the nascent transcript at a CA dinucleotide [369], yet an increase in cytosine could not be detected around the cleavage site. Towards the end of the annotated genome-templated read sequences, the adenosine frequencies increase even before the actual start of the poly(A) tail. It can however not be excluded that a fraction of adenosines at the end of the aligned reads are actually remnants of the poly(A) tail that the pipeline erroneously

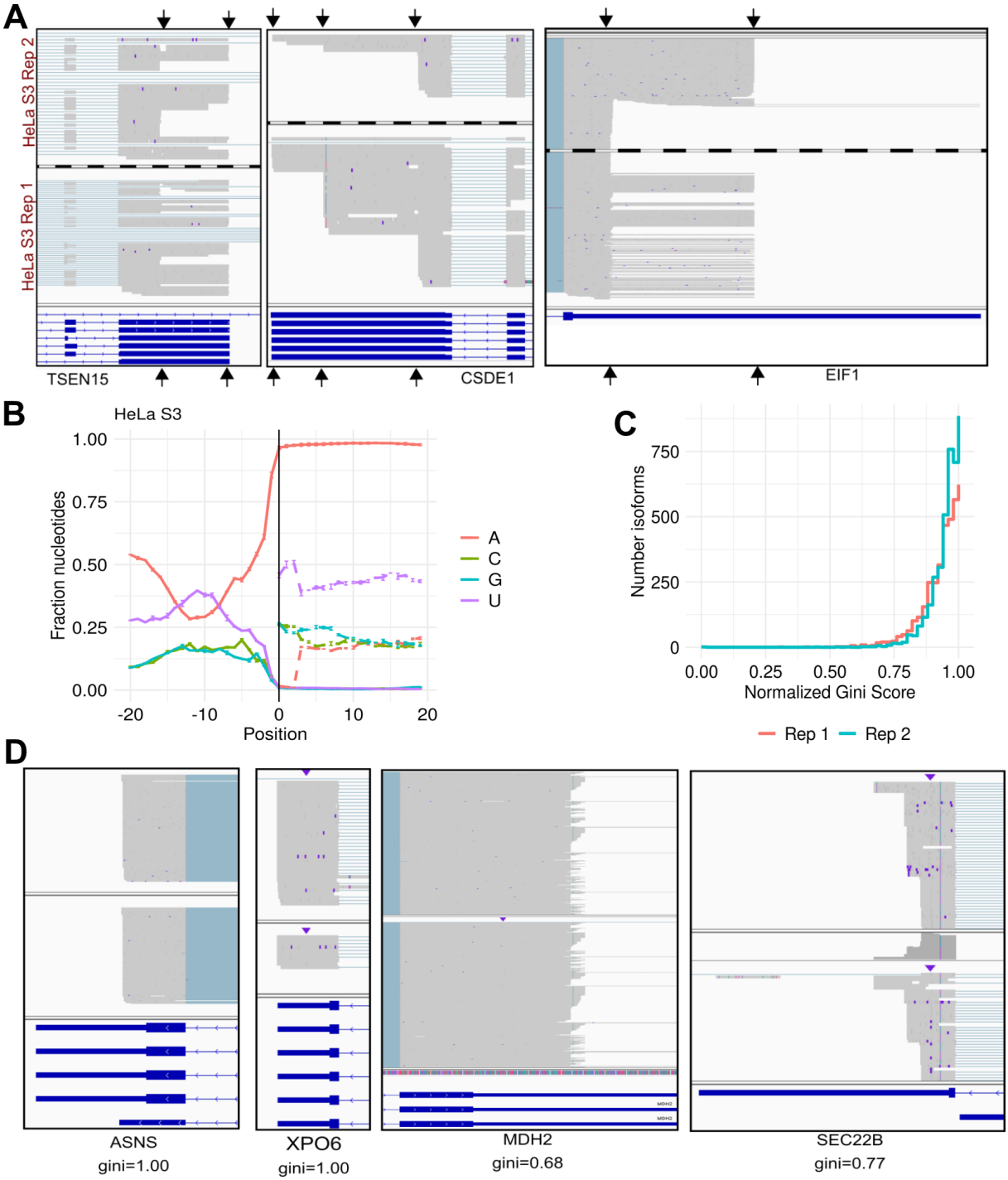

Figure 11 **Annotation of 3'-UTRs from FLAM-Seq data**
A) IGV Genome browser shots for TSEN15, CSDE1 and EIF1 loci visualizing aligned reads for HeLa S3 FLAM-Seq replicate datasets: Arrows indicate 3'-UTR RNA isoforms annotated by computational pipeline and comparison to IGV gene models. **B)** Nucleotide frequencies occurring upstream and downstream (poly(A) tail) of the cleavage site as annotated from FLAM-Seq datasets for HeLa S3 replicates. Dashed lines indicate nucleotide content of genomic sequence downstream of cleavage site. Vertical bar indicates identified beginning of the poly(A) tail. **C)** (Normalized) Gini coefficients of hexamer frequencies occurring at read ends at cleavage site for individual 3'-UTR isoforms. **D)** IGV genome browser shots of example genes with low or high normalized Gini scores highlight putative differences in 3'-end definition.

assigned to the templated 3'-UTR end of the RNA. The genomic sequence downstream of the cleavage site is U-rich which is in line with previous investigations of the sequence composition around polyadenylation sites [370].

Manual inspection of read ends in a genome browser showed that for many genes the putative 3'-UTR end from alignments for a given isoform were not 'sharp' at one coordinate, but rather scattered at different positions around an annotated 3'-UTR end, within a range of around 10 nt. To quantify the extent of this effect, hexamer frequencies at the aligned 3'-end positions of each read were calculated for each isoform. A Gini coefficient was then calculated, which measures the degree of inequality in hexamer usage for all alignments associated with a given 3'-UTR end. For a 'sharp' cleavage site, this would result in a high inequality in hexamer usage i.e. a high Gini score, since only a single hexamer should be occurring at the read end, whereas low Gini scores were associated with less defined cleavage sites (Figure 11 C). A number of genes were observed with Gini scores below 0.8, which indicated less 'precision' in cleavage site usage compared to isoforms with a Gini score of 1 (Figure 11 D).

Different variants of the canonical polyadenylation signal AAUAAA have been described previously [61] to alternatively occur upstream of the cleavage site in the nascent transcript. To quantity the extent of alternative PAS variant usage, frequencies of previously described PAS variants most proximal to the annotated 3'-UTR ends were counted for human and *C. elegans* FLAM-Seq datasets (Figure 12 A). As expected, AAUAAA was the most frequently occurring PAS variant, found at ca. 50% of sequenced transcripts, with a slightly reduced frequency for *C. elegans* samples. Second most frequent was AUUAAA in human samples and AAUAAU in *C. elegans* hinting at species dependent differences. Species related differences were also observed for many minor PAS variants (e.g. AACAAG or AAUGAA) while differences between for instance HeLa S3 and iPS were negligible. For ca. 10% of reads no polyadenylation signal was detected. Comparing the PAS positional frequencies showed that for most reads PAS were detected 20-21 nt upstream of the cleavage site, independently of the sequenced sample, although for human samples PAS positions appeared slightly more shifted towards more distal positions from the cleavage sites (Figure 12 B).

3'-UTR length profiles for individual isoforms were compared between HeLa S3, iPS cells, organoids and *C. elegans* samples (Figure 12 C). iPS cells had overall slightly longer 3'-UTRs compared to HeLa S3 cell lines with a median of 439 nt compared to 501 nt. Organoid median 3'-UTR length was 695 nt and longer than for iPS cells which is expected given the lengthening of 3'-UTRs observed in other neuronal systems [101]. *C. elegans* 3'-UTRs were much shorter

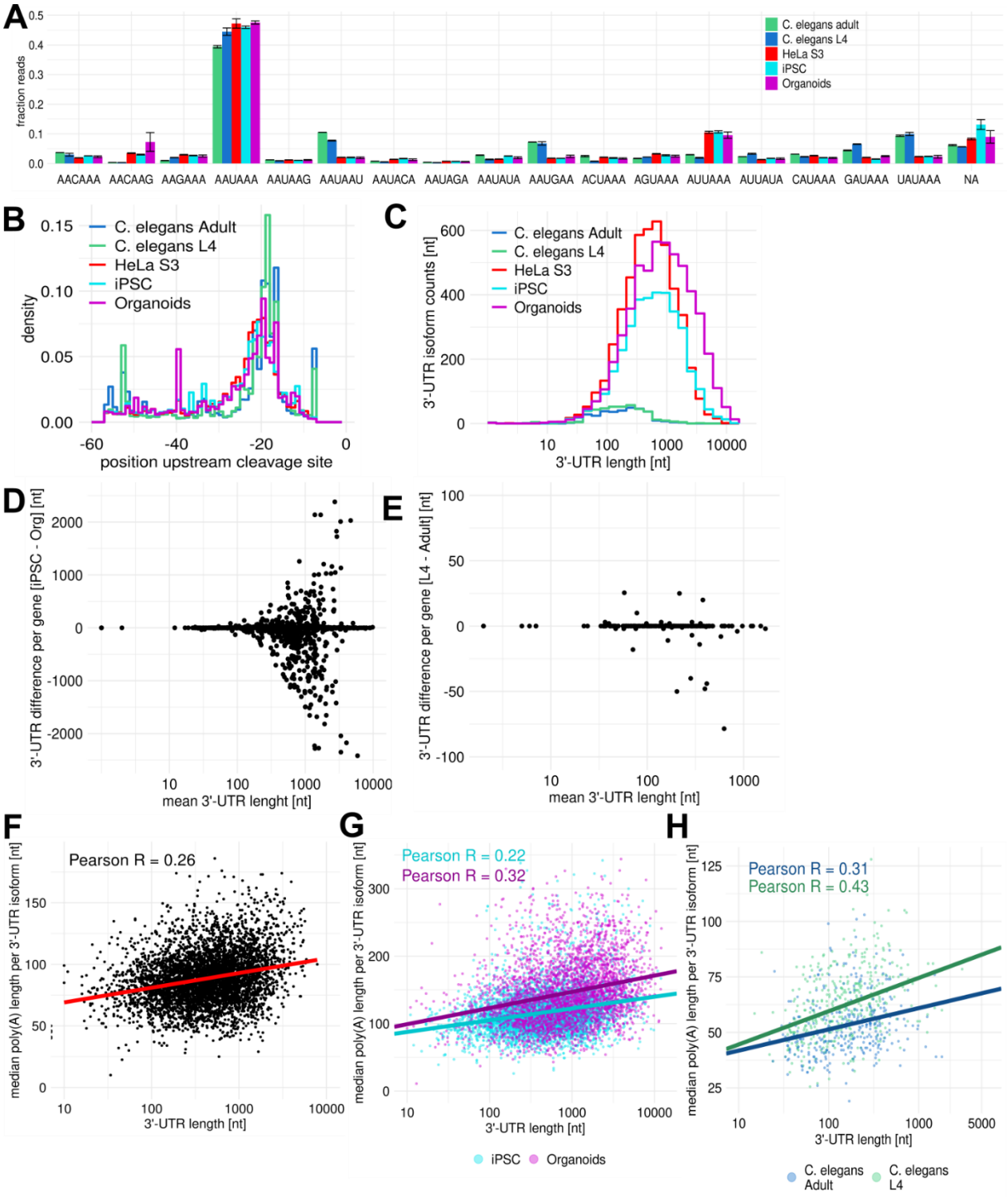

Figure 12 **Dynamic polyadenylation site choice and 3'-UTR length regulation**
A) Relative polyadenylation signal (PAS) usage in reads from FLAM-Seq datasets within a 60 nt window from the 3'-UTR end. Error bars denote standard error of the mean between FLAM-Seq replicates. NA denotes no identified PAS. **B)** Distributions of positions of polyadenylation signals upstream of cleavage site at the 3'-UTR ends of FLAM-Seq reads. **C)** 3'-UTR length distributions of RNA isoforms identified in FLAM-Seq datasets. **D)** Difference in 3'-UTR length per gene between iPSC and Organoids FLAM-Seq datasets versus median 3'-UTR length of each gene. **E)** Difference in 3'-UTR lengths between *C. elegans* L4 and adult FLAM-Seq datasets versus median 3'-UTR length in samples. **F)** Median poly(A) tail length per 3'-UTR isoform versus 3'-UTR length in merged HeLa S3 FLAM-Seq samples. **G)** Median poly(A) tail length per 3'-UTR isoforms versus 3'-UTR length in merged iPSC and Organoids FLAM-Seq samples. **H)** Median poly(A) tail length per 3'-UTR isoform versus 3'-UTR length in merged *C. elegans* L4 and adult FLAM-Seq samples.

with a median of 165 or 139 nt for adult and L4 stage respectively. 3'-UTR length was compared between identical genes for iPS and organoids samples (Figure 12 D). 281 genes with extended 3'-UTRs in organoids were detected with an absolute length difference of more than 30 nt. Only 152 genes had shorter 3'-UTRs upon differentiation, which corresponded to a total of 17% of all expressed genes with changing 3'-UTR isoforms. On the contrary, only 5 genes had longer 3'-UTRs in *C. elegans* adult samples, hinting at less detectable 3′-UTR plasticity throughout *C. elegans* development (Figure 12 E). 3'-UTR length was also compared to median poly(A) tail length per 3'-UTR isoform (Fig 8 F-H), which showed a clear trend for longer poly(A) tails proportional to the 3'-UTR length, which was strongest for *C. elegans* samples. 3'-UTR length was also the best predictor for median poly(A) tail length per isoform, compared to RNA stability, expression and other features.

Poly(A) tail length profiles were compared between different 3'-UTR isoforms identified for the same gene in HeLa S3, iPSC, organoids and *C. elegans* samples (Figure 13 A). 980 out of 4342 genes in HeLa S3 samples, expressed with more than 5 counts, were detected with more than one annotated 3'-UTR isoform. This fraction was comparable in iPSC and organoid samples where 674 out of 3264 and 1005 out of 4897 genes were identified with alternative 3'-UTRs. Alternative polyadenylation was less prevalent in *C. elegans*, where only 104 out of 529 or 84 out of 409 genes had multiple 3'-UTR isoforms, which could also be impacted by the generally shorter *C. elegans* 3'-UTRs. Most genes undergoing alternative polyadenylation produced two 3'-UTR isoforms (80% for human samples, 95% for *C. elegans*). Comparing poly(A) tail length profiles between 3'-UTR isoforms of the same gene revealed that in many cases alternative 3'-UTR profiles were associated with differences in poly(A) tail length profiles (Figure 13 A). 329 of 980 alternatively polyadenylated genes in HeLa S3 showed differences in poly(A) length profiles, in iPS cell this number was 290 genes and 547 genes in organoids. Again, fewer cases were found for *C. elegans* samples where only 9 or 16 genes were detected with 3'-UTR related differences in poly(A) tail length. Two example genes from HeLa S3 illustrated differences in poly(A) tail length (Figure 13 B): For the IMP4 proximal (shorter) 3'-UTR isoforms the median poly(A) tail length was 71 nt, while the longer distal isoform had a length of 126 nt, while the difference in 3'-UTR length was around 800 nt. For the AK2 gene, three 3'-UTR isoforms were found which had median poly(A) tail length of 70, 85 and 104 nt.

Ordering median poly(A) tail length per isoform by shorter proximal or longer distal isoforms for all genes showed that longer poly(A) tails were generally associated with more distal polyadenylation sites, except for the *C. elegans* adult sample were only few genes were detected

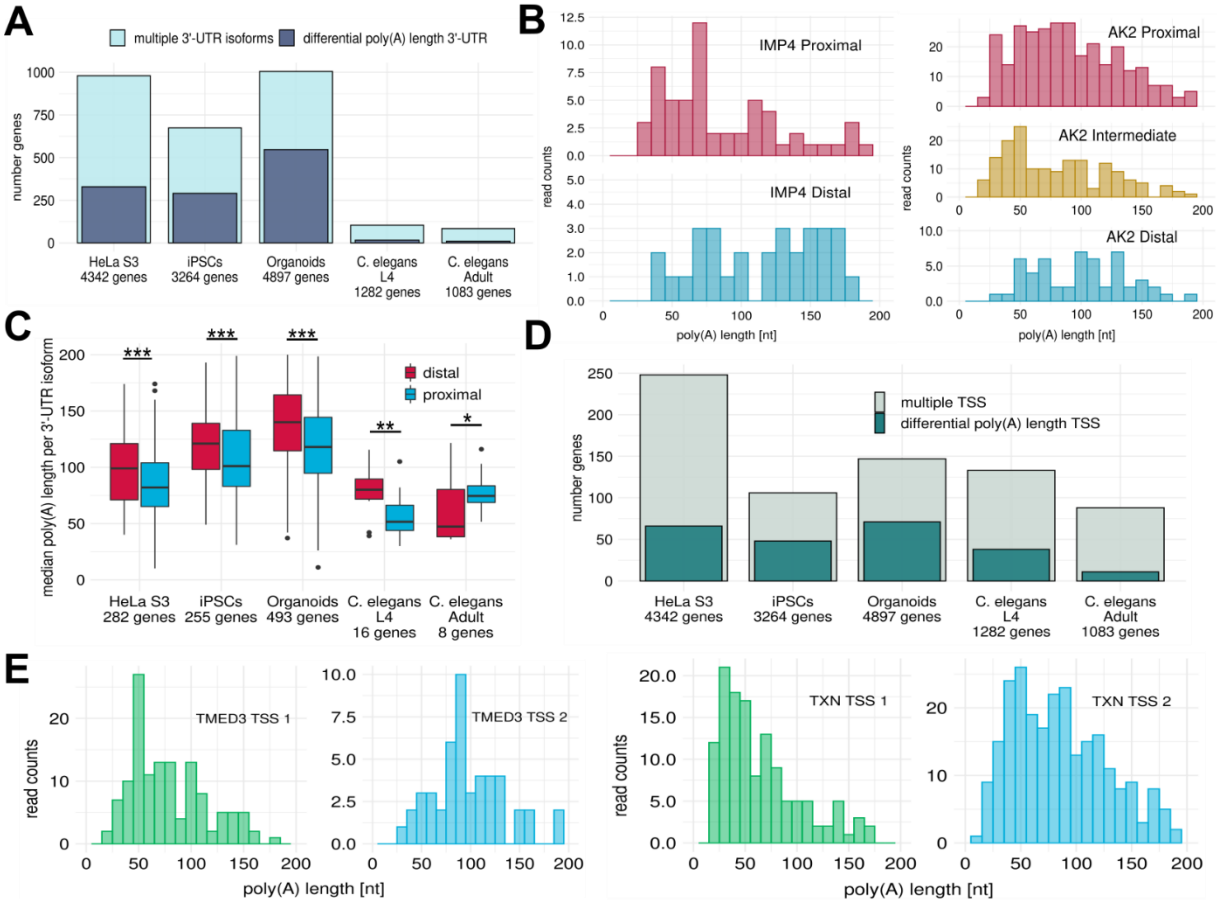

Figure 13 **Alternative polyadenylation and transcription start site usage (TSS)**
A) Number of genes with multiple identified 3‘-UTR isoforms and of those number of genes with significant differences in poly(A) tail length profiles of alternative 3’-UTR isoforms. **B)** Poly(A) tail length profiles of IMP4 and AK2 3’-UTR isoforms with which have significant differences in poly(A) tail length distributions. **C)** Median poly(A) tail length distributions for proximal and distal 3’-UTR isoforms. **D)** Number of genes with multiple transcription start sites (TSS) and of those the number of genes with significant differences in poly(A) tail length profiles between TSS isoforms. **E)** Poly(A) tail length distributions of example genes with significant differences in poly(A) tail length profiles between 3’-UTR isoforms.

with significant differences in poly(A) profiles (Figure 13 C). The average difference in global poly(A) length was 25 nt between proximal and distal isoforms of all samples.

A similar analysis was performed for transcription start sites (TSS), where FLAM-Seq reads were grouped based on previously annotated transcription start sites from human CAGE or *C. elegans* SAGE data to identify RNA isoforms differing in their transcription start sites. Between 100 and 250 genes were identified with alternative transcription start site usage, and of those around 50 genes were found were alternative TSS isoforms for a gene had significant differences in median poly(A) tail length, except for *C. elegans* adult samples, were this number was lower (Figure 13 D). In Hela S3 samples, two TSS isoforms were detected for the TMED3 gene, with median tail length of 72 nt and 93 nt. A comparable effect was found for the TXN

gene were two TSS isoforms differed by 30 nt (Figure 13 E). No significant enrichments were found between distinct combination of alternative TSS and 3'-UTRs which could be expected in cases where promoter choice impacts 3'-UTR cleavage site selection.

4.1.6 Poly(A) tails contain non-A nucleotides with a preference for cytosines

Previous research showed that poly(A) tails are not exclusively consisting of adenosines and tail modifications have important biological functions for instance with regard to RNA stability [371].

Frequencies of non-A nucleotides within poly(A) tail were calculated for HeLa S3, iPS and organoids and *C. elegans* samples. RNA and cDNA standards were taken as a control to account for potential mismatches introduced by enzymatic library preparation for instance through PCR steps or sequencing errors. Non-A frequencies were calculated as average cytosine, guanine and uridine counts per tail (Figure 14 A) or as fraction to all sequenced tail nucleotides (Figure 14 B). Cytosine frequencies were highest with a frequency of 0.2 % in human samples, around 0.12% in *C. elegans* and only around 0.03% in synthetic RNA / cDNA standards, indicating that cytosines were greatly enriched over the baseline technical error. Uridines were the second most enriched nucleotide occurring at a frequency of around 0.05% in human samples, 0.075% in *C. elegans* and only 0.01% in synthetic standards. Guanines occurred at a frequency of around 0.05%, but a similar frequency was found in controls such that detected guanines are potential artifacts of library preparation or sequencing. Normalization to total sequenced nucleotides (Figure 14 B) or averaging for each sequenced poly(A) tail did not greatly impact the reported frequencies per nucleotide, which hinted at a uniform distribution of non-A nucleotides over poly(A) tails. To further investigate whether non-A modifications are occurring only for subsets of genes or on a genome-wide level, counts of poly(A) tails containing non-A nucleotides were compared to total counts for each gene in each sample. A linear trend was observed between poly(A) tails containing U, G or C nucleotides and the total number of sequenced tails with Pearson correlation coefficients ranging from r = 0.75 for *C. elegans* adult to r = 0.92 for HeLa S3 samples (Figure 14 C). Poly(A) tails with non-A nucleotides were detected for 2900 to 9700 expressed genes with corresponded to 36–71% of all genes in FLAM-Seq samples. For genes with non-A containing poly(A) tails, on average 25% of all sequenced molecules contained at least one non-A nucleotide.

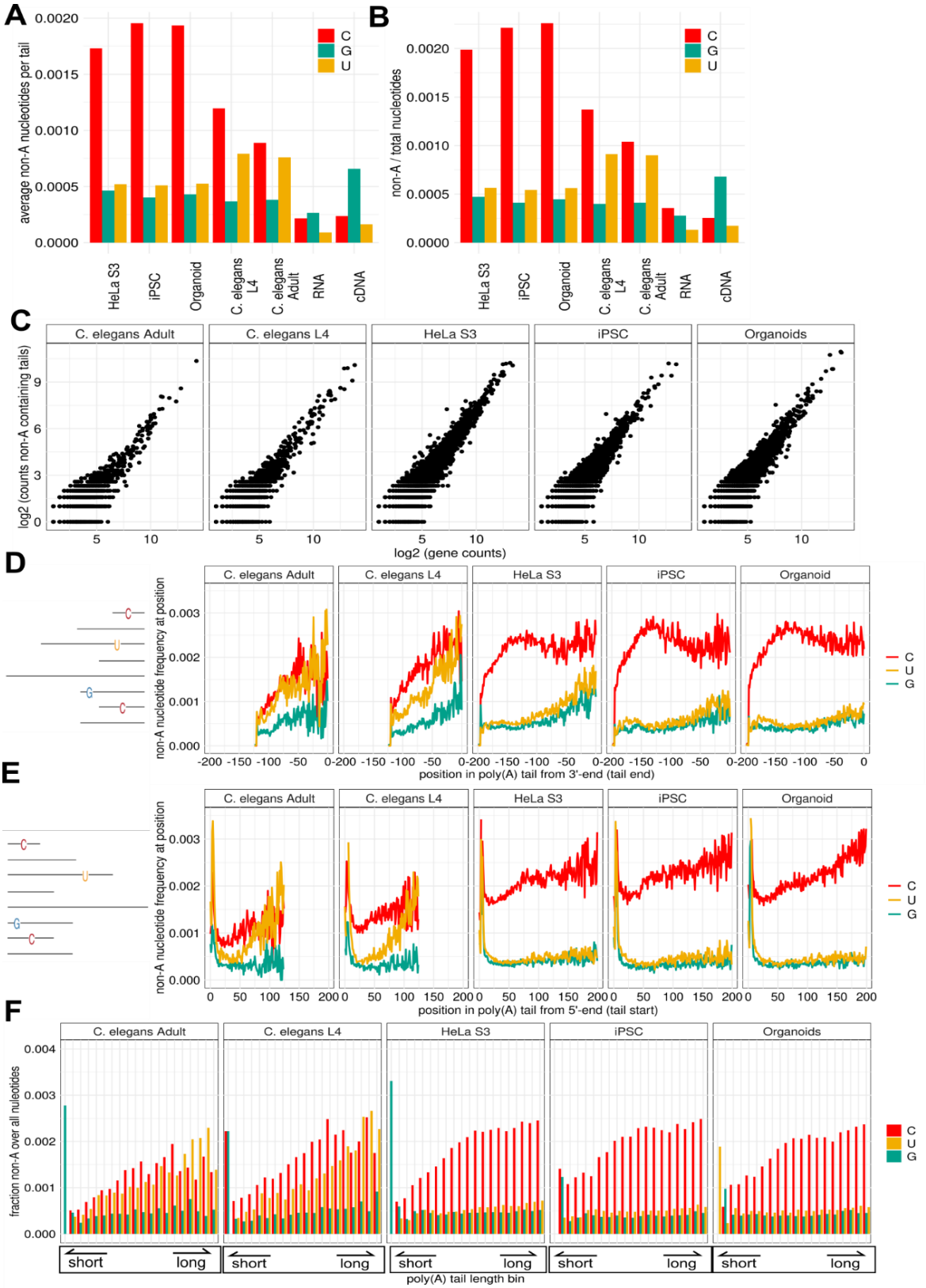

Figure 14 **Poly(A) tails contain non-A nucleotides**

A) Average frequencies of detected C, U and G nucleotides normalized to total nucleotides for each sequenced poly(A) tail. **B)** Non-A nucleotide frequencies normalized to total sequenced nucleotides in each sample. **C)** Number of detected molecules per gene which contain non-A nucleotides in poly(A) tail compared to total number of detected molecules per gene. **D)** Non-A nucleotide frequencies by position in poly(A) tail for tails aligned at their 3‘-end (tail end) or **E)** aligned at their 5‘-ends (tail start). **F)** Frequencies of non-A nucleotides in poly(A) tails where tails are binned by 10 nt bins.

Modifications such as guanylation and uridylation have been reported to occur at poly(A) tail 3'-ends [227]. The FLAM-Seq protocol is yet limited in investigating non-A nucleotides at the very 3'-end positions. This is because a synthetic GI-tail is added before reverse transcription which does now allow for distinction of endogenous guanosines. Second, the primer for reverse transcription has an overhang of three Ts, which selects for RNAs with poly(A) tails ending in 3 As.

To investigate positional preferences of non-A nucleotides within poly(A) tails, poly(A) tail sequences were first aligned at their 3'-ends. Nucleotide frequencies were then calculated for each position moving towards the start of the poly(A) tail across all tails for each sample (Figure 14 D). For cytosines in human samples we observed an increase in frequency towards the 'middle' of the tail, and a slight drop towards the start. Also G and U frequencies slightly increased towards the tail start.

For *C. elegans* a universal increase in non-A frequencies towards the tail start was observed. Reads were next aligned at their 5'-ends, i.e. the poly(A) tail start downstream of the 3'-UTR (Figure 14 E). Non-A frequencies were in all cases increased at the first position, which is likely a consequence of the FLAMAnalysis pipeline not correctly trimming remaining nucleotides of the 3'-UTR end encoded by the genome at the poly(A) tail start. Cytosines increased towards the tail end, while Gs and Us remained constant, except for *C. elegans* samples where uridine frequencies also increased. In summary, non-A nucleotides appeared to be more enriched within poly(A) tail bodies. Poly(A) tail ends were characterized by less non-A nucleotides and poly(A) tail starts by high non-A frequencies, which could be leftover 3'-UTR nucleotides not properly trimmed.

As an orthogonal analysis, poly(A) tails were binned by tail length and non-A frequencies were computed for each bin (Figure 14 F). For human samples, an increase in frequency was observed exclusively for cytosines which occur more frequently in longer poly(A) tails. An increase in uridine frequencies for longer poly(A) tails was seen for *C. elegans*. The lowest bin with tails less that 10 nt had high non-A frequencies, yet the number of detected tails was lowest in this bin which may limit statistical power.

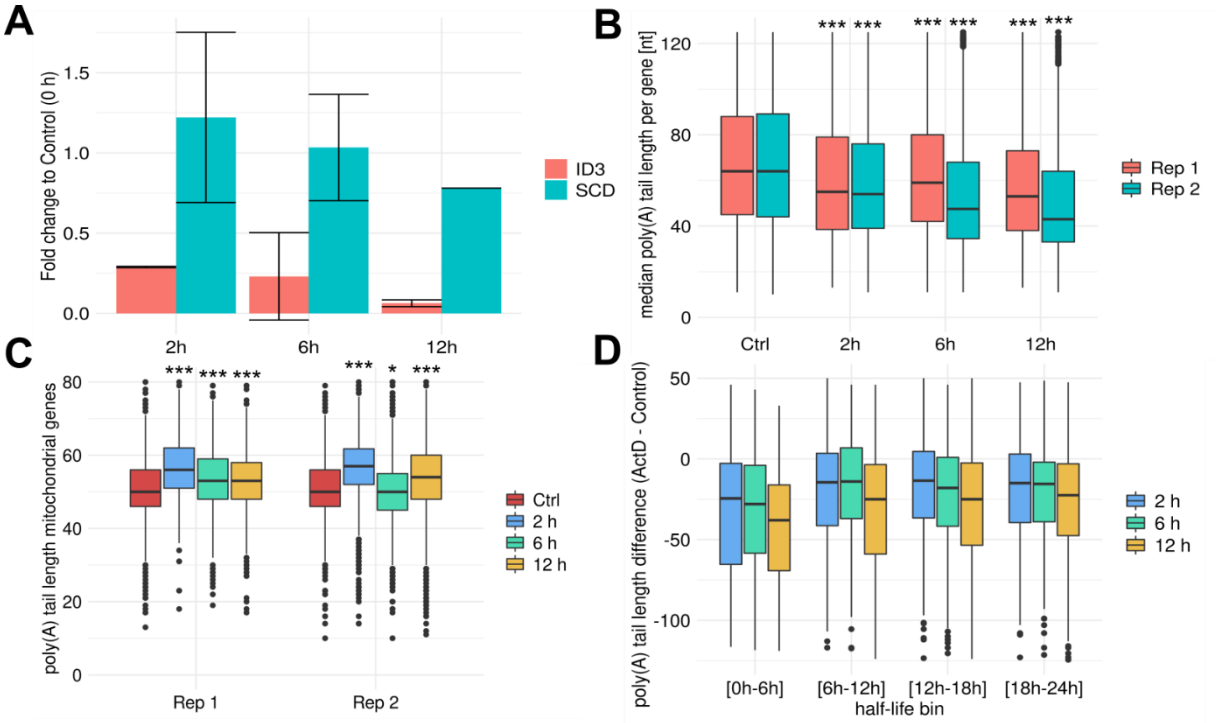

Figure 15 **Transcription inhibition using actinomycin D**
A) Gene expression fold-changes comparing ID3 and SCD gene expression for different timepoints after transcription inhibition to control (0 h ActD). GAPDH was used as a reference gene in qPCR quantification. **B)** Median poly(A) tail length per gene for replicates after transcription inhibition for different timepoints and control. Statistical significance of differences compared to control was calculated using Wilcoxon test. **C)** Poly(A) tail length of mitochondrial genes for replicates and different timepoints after transcription inhibition. Statistical significance of differences compared to control was calculated using Wilcoxon test. **D)** Difference in median poly(A) tail length per gene between ActD-treated and control timepoints for genes grouped by half-life bins with an interval of 6 h (x-axis).

4.1.7 Transcription inhibition leads to accumulation of shorter poly(A) tails

Inhibition of transcription is a useful method for investing RNA dynamics and decay and usually applied in combination with gene expression analysis of multiple timepoints after inhibition. To investigate the potential of FLAM-Seq in resolving changes in poly(A) tail length over time, transcription was inhibited using Actinomycin D, which intercalates DNA and thereby inhibits transcription by all RNA polymerases[15]. Transcription inhibition was performed in HEK Flp-In 293 T-rex cells up to 12 h in replicates. The effect of transcription inhibition was validated by comparing gene expression of less stable transcripts with shorter half-lives with the expression of housekeeping genes such as GAPDH. Fold-changes of ID3 (half-life $t_{1/2}$ = 0.8 h) and SCD ($t_{1/2}$ = 18 h) were measured in relation to GAPDH expression ($t_{1/2}$ = 21 h) (Figure 15 A). ID3 expression was reduced to 6% after 12 h of transcription inhibition compared to GAPDH, while SCD was expressed at around 78% of the control, while no reduction was observed after 2 h and 6 h. Comparing poly(A) tail length per gene between

control and timepoints of transcription inhibition showed progressive shortening of poly(A) tail upon longer inhibition periods.

Median poly(A) tail lengths were in good agreement between replicates of control and 2 h timepoints which differences while the observed variability was larger between replicates of 6 h and 12 h timepoints. Average median poly(A) tail length was 98 nt for control, 73 nt for 2 h, 72 nt for 6 h and 63 nt for 12 h time points, which indicates progressive shortening of tails (Figure 15 B). As a control, poly(A) tails of mitochondrial transcripts, which have a steady state poly(A) tail length of around 50 nt, were studied over time (Figure 15 C). Observed median tail length for mitochondrial transcripts varied from 50 nt to 57 nt between timepoints and replicates. The 2 h timepoints showed the longest mitochondrial tail length in both replicates. Whether differences of mitochondrial tail length were yet truly biological or defined the error margin of the FLAM-Seq protocol was unclear. The trend for progressive shortening of tails after transcription inhibition could not be seen for mitochondrial transcripts.

Differences in deadenylation rate were shown to directly impact RNA decay rates [265], in turn differences in poly(A) shortening across timepoints should be observable for genes with differences in RNA stability. To investigate the shortening of unstable transcripts over time, genes were binned by transcript half-lives per gene and the difference in median poly(A) tail length per gene between 0 h control and all other timepoints were determined (Figure 15 D). Poly(A) tail differences were largest for genes with a half-life of up to 6 hours, which was expected given the described relationship and the differences were conversely smallest for very stable transcripts, with little difference between 6 h and 12 h timepoints. FLAM-Seq hence enabled investigation of genome-wide poly(A) tail dynamics over time after inhibiting transcription.

4.2 Genome-wide nuclear deadenylation of mRNAs

The previous chapter described FLAM-Seq as a versatile tool for exploring RNA biology and gene regulation through sequencing of complete RNAs which uncovered important elements of RNA 3'-end processing and the tight coupling of (alternative) polyadenylation and poly(A) tail length control. This chapter investigates poly(A) tail metabolism first in context of nascent RNAs and pre-mRNA splicing, revealing that polyadenylation generates genome-wide long poly(A) tails of more than 200 nt. Since steady state poly(A) tail length distributions were much shorter in all profiled biological samples, the question of how tails are shortened to reach steady state length remained and was investigated by metabolic labeling and biochemical fractionation experiments to explore temporal and spatial features of deadenylation. Those experiments uncovered a fast nuclear deadenylation step. Finally, different experimental strategies were applied to perturb known deadenylase complexes which could be involved in nuclear deadenylation, and poly(A) tail profiles were measured in subcellular fractions to identify the enzyme(s) responsible for nuclear deadenylation.

4.2.1 Unspliced mRNAs have long poly(A) tails

In vitro experiments using reconstituted components of the cleavage and polyadenylation machinery [163] as well as metabolic labeling experiments of total

polyadenylated RNA [372,373] previously revealed synthesis of poly(A) tails with a length of around 250 nt. Those assays were yet not able to resolve individual genes and to further investigate this nascent poly(A) tail synthesis on a genome-wide scale, unspliced RNA molecules were extracted from FLAM-Seq datasets and the poly(A) tail length of unspliced reads was analyzed. For this, a reference of unambiguous intron annotations was curated. The reference contained introns of protein-coding genes that did not overlap with any exon annotations to exclude that reads, which may have come from alternative isoforms of the same gene, are identified as 'unspliced'. Overlapping FLAM-Seq alignments with intron and 3'-UTR annotations provided a stringent filter for identification of unspliced, intronic reads, which could also be visualized in a genome browser instance of the organoids AHSG locus (Figure 16 A). Comparing the poly(A) tail length of unspliced to spliced reads showed a median poly(A) tail length per gene of 256 nt for unspliced and 124 nt for spliced reads, which supports the initial hypothesis of synthesis of long poly(A) tails for this individual AHSG gene.

Comparing the fraction of detected intronic reads between HeLa S3, iPSCs and organoids showed a more than 5-fold increase in detected unspliced reads from around 0.03% to a

maximum of 0.22% of all sequenced reads (Figure 16 B) which corresponded to a total number of 77 to 1,313 unspliced molecules. The observed differences in relative detected intronic reads could be related to differences in splicing kinetics between cancer cell lines, stem cells and organoids. Poly(A) tail length of intronic reads had a median of 151 nt for HeLa S3, 208 nt for iPSCs and 232 nt for organoids (Figure 16 C). Many intronic reads for HeLa S3 samples were much shorter than 200 nt, which either hinted at artefacts in our computational pipeline, (cytoplasmic) transcripts with retained introns, or indeed poly(A) synthesis of shorter poly(A) tails.

The raw read length of unspliced reads was compared against total reads per sample to ensure that intronic reads had comparable sequencing properties. The cumulative raw read length distributions (Figure 17 A) were almost identical between bulk and intronic reads, but stark differences in read length were observed between replicates for the organoid samples as observed before (Figure 5 A). This may also explain the variability in detected unspliced reads Figure 17 B) for organoid replicates.

The intron reference used for extracting unspliced reads stringently excluded genes with ambiguous assignments or introns overlapping with exons of other isoforms. Additionally, for some genes the introns closest to the transcript ends (3'-UTR ends) had genomic distances of several kbp, which made detection of those introns unlikely given the observed read length limitations. To quantify the fraction of genes in each FLAM-Seq sample for which intronic reads could in principle be detected, genes with a maximum intron distance of 3 kbp from the transcript end were counted relative to all expressed genes per sample. Around 50% of expressed genes were represented in the intron annotation, independently of the FLAM-Seq sample. This fraction represents the upper bound of detectable genes with intronic reads (Figure 17 B) for later analysis. Genes with associated intronic reads were compared and between 1% and 7% of all detected genes had associated intronic reads across FLAM-Seq samples. Merging all datasets, this number increased to around 8% (Figure 17 C). This analysis showed a trend for more detected genes with intronic reads proportional to the total number of reads, i.e. sequencing depth in the sample. This hinted at a random sampling process which would be characteristic of genome-wide synthesis of long poly(A) tails. Intronic reads were next merged for all FLAM-Seq samples and downsampled to a given percentage of the total number of reads. The fraction of genes with intronic reads was then calculated based on the downsampled reads, which revealed an almost linear relation between the number of unspliced reads and the number of detected genes with unspliced reads (Figure 16 D). The downsampling analysis also suggested that deeper sequencing would detect more genes with associated unspliced reads.

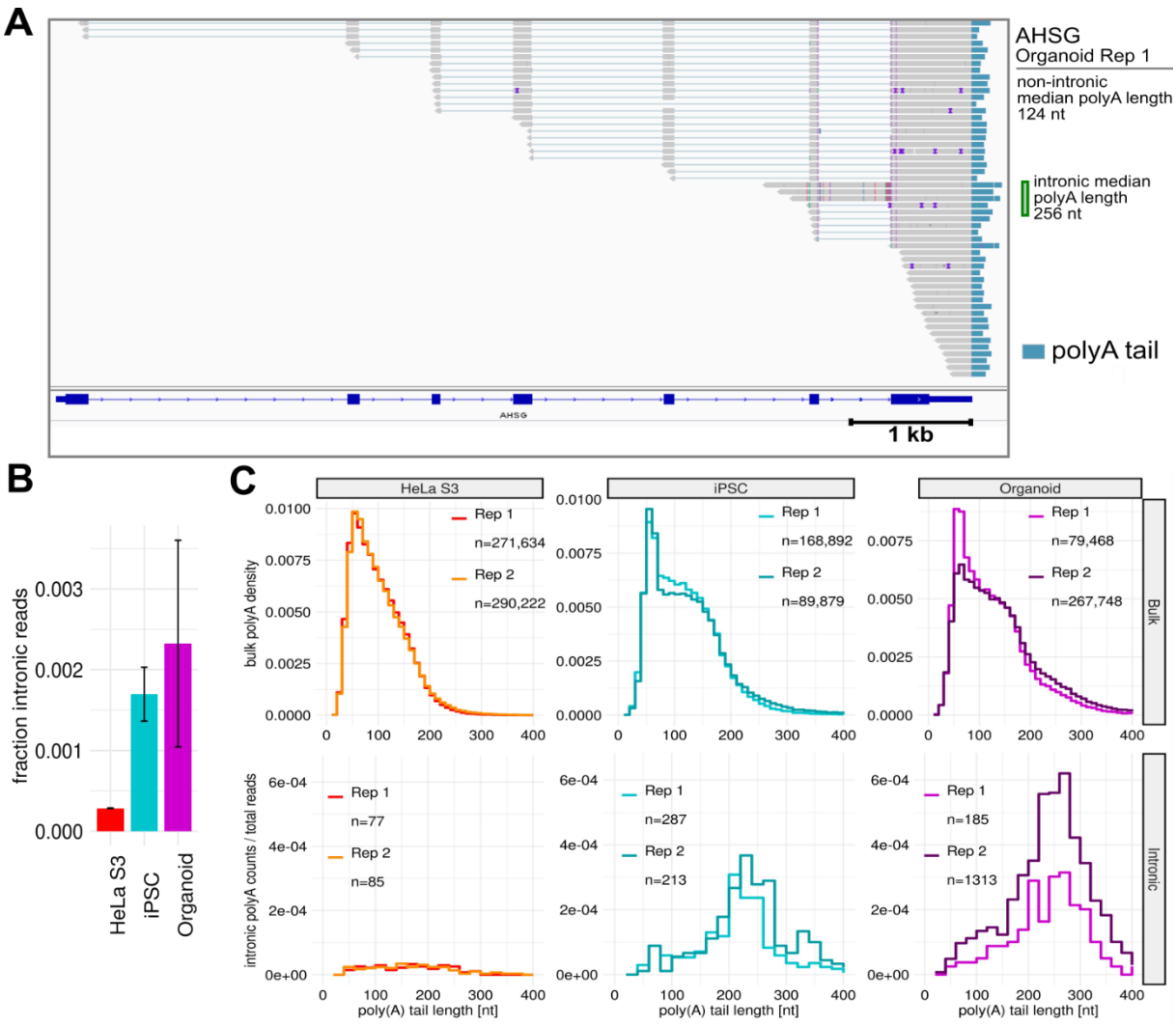

Figure 16 **Poly(A) tail length profiles of unspliced intronic reads**
A) IGV Genome Browser shots for AHSG locus in FLAM-Seq organoids dataset showing alignments of individual reads. Poly(A) tail length for individual reads were appended in blue. Number of detected intronic reads = 3; Number of spliced spliced reads = 317. **B)** Fraction of unspliced, intronic reads per sample. Error bars indicate standard error of the mean between replicates. **C)** Poly(A) tail length profiles of HeLa S3, iPS cells and organoid FLAM-Seq replicates. Top row: Poly(A) tail length density profiles of all sequenced reads ('bulk'). Bottom row: Poly(A) tail length frequencies of intronic, unspliced reads normalized to total reads per replicate. Read per dataset are indicated in the legend.

Comparing the genes for which unspliced reads were detected between HeLa S3, iPSCs and organoids showed little overlap, which supported the notion of random sampling of genes with associated intronic reads (Figure 16 E).

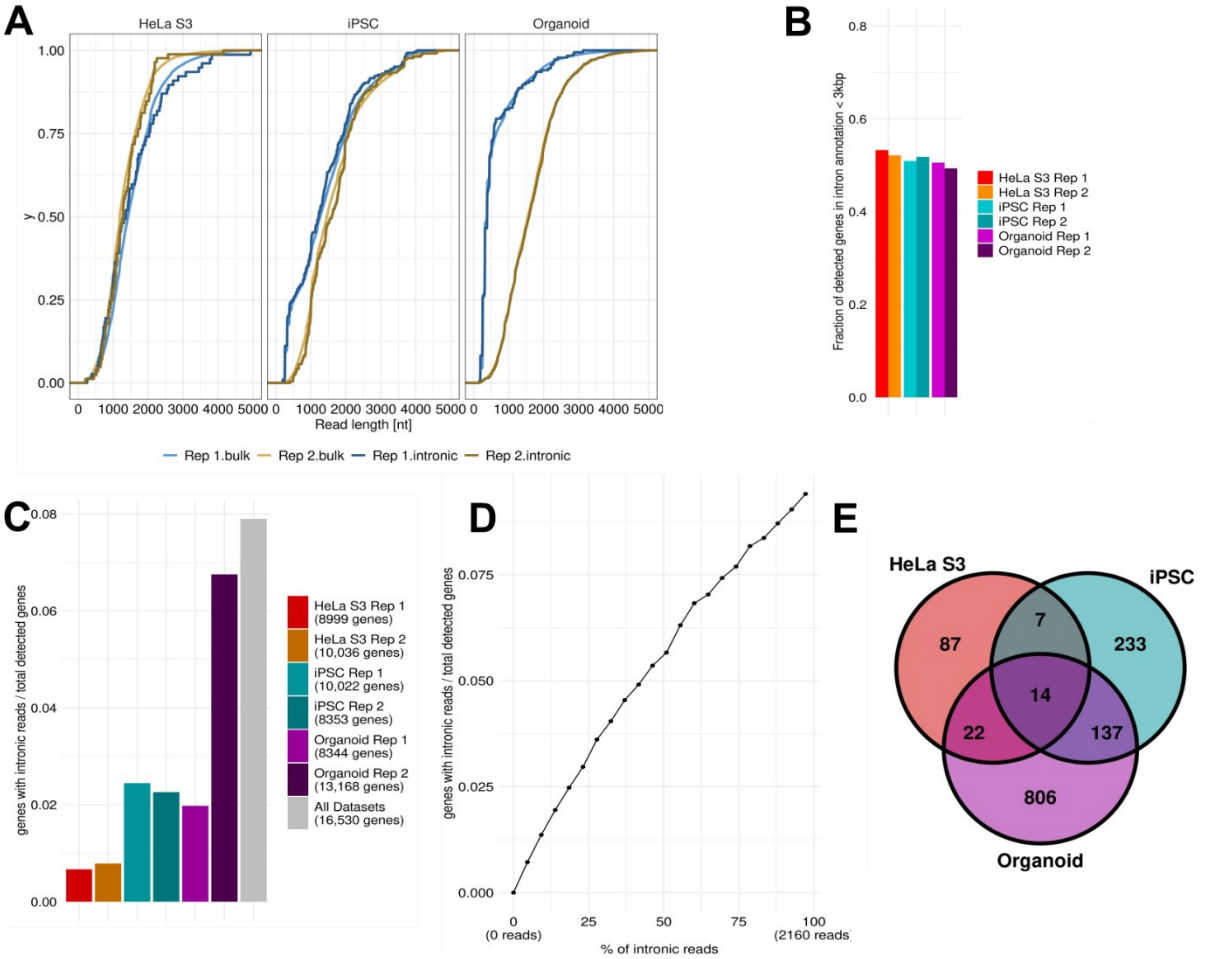

Figure 17 **Analysis of intronic reads detected in FLAM-Seq data**
A) Raw sequencing read length distribution of total ('bulk') and intronic reads as cumulate density distributions for FLAM-Seq samples. **B)** Fraction of annotated genes with unambiguous intron annotations for identification of unspliced reads (intronic sequences less than 3 kbp from 3'-UTR end). **C)** Fraction of genes with identified unspliced, intronic reads normalized to total detected genes in each FLAM-Seq sample. **D)** Downsampling of reads from merged FLAM-Seq datasets and quantification of unspliced reads as a fraction of total reads detected. **E)** Overlap of genes with unspliced, intronic reads as Venn diagram between HeLa S3, iPSC and organoid datasets.

Intron length is an important parameter, which implicates for instance splicing kinetics [374]. Unspliced reads were hence investigated in context of their intron length: first, the length of introns which directly overlap with aligned reads from the curated intron reference used for identification of unspliced reads were compared to all introns the curated reference (Figure 18 A). No major differences were noticeable, except at the extremes of the intron length distribution. The background distribution also contained many short introns of less than 100 nt and some exceptionally long introns which were also found less frequently in intronic reads.

As a second control, intron length was compared between all introns annotated in the Gencode v28 annotation and introns of genes with associated unspliced reads (Figure 18 B). No major difference in intron length could be identified here as well.

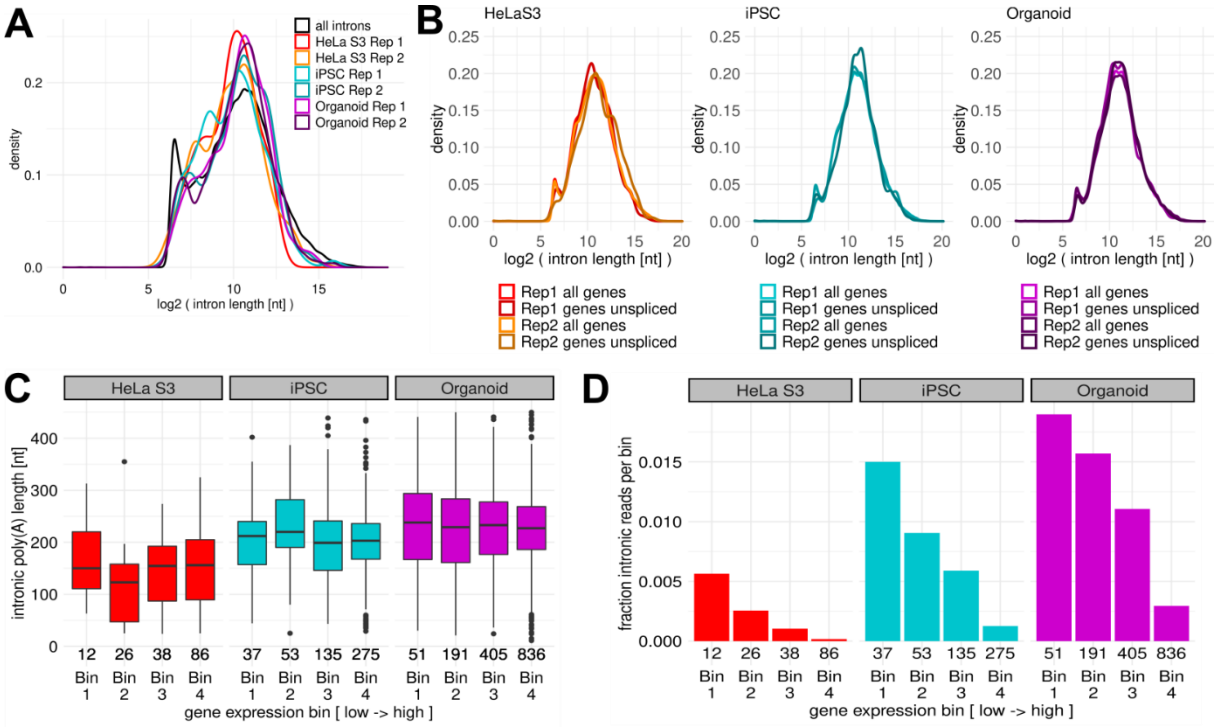

Figure 18 **Intron length and expression features of genes with intronic reads**
A) Distribution of intronic read length of introns in unspliced reads for each sample against all introns in annotation. **B)** Length distributions of all annotated Gencode introns of genes with detected intronic reads against all expressed genes. **C)** Poly(A) tail length distribution of unspliced, intronic reads binned by total expression counts of respective genes. Number of reads in each bin are displayed above the x-axis. **D)** Fraction of intronic reads by total reads per gene with genes binned by total expression counts of associated genes. Number of reads in each bin are displayed above the x-axis.

Genes were next binned by expression into four groups and poly(A) tail length distributions of intronic reads were compared between gene expression bins. No major biases in the poly(A) tail length of unspliced reads could be identified with respect to expression levels of the associated genes (Figure 18 C). Median intronic poly(A) tail length varied slightly for the second bin of HeLa S3 and iPSCs, which may also be related to the relative low number of unspliced reads per bin. A similar analysis was performed by calculating the fraction of intronic reads by total reads for each gene expression bin (Figure 18 D). The fraction of unspliced reads decreased for higher expressed genes independently of the FLAM-Seq sample, which may hint at more efficient RNA processing for higher expressed genes.

In summary the presented analysis showed that unspliced, intronic reads could be identified in FLAM-Seq datasets, which had overall long poly(A) tails of more than 200 nt in iPSC and organoid datasets and around 150 nt for HeLa S3 cells. Investigating genes with associated unspliced reads hinted at a random sampling process by which unspliced reads are detected, which suggested that the detected unspliced reads were representative of the genome-wide

synthesis of long poly(A) tails. Comparison of intron length and gene expression did not show any biases for distinct molecular properties differentiating genes with intronic reads, which supported the conclusion of unbiased sampling of unspliced reads.

4.2.2 Splicing inhibition causes an increase in unspliced reads and poly(A) lengthening

Splicing can be inhibited by treating cells with small molecule inhibitors of spliceosome assembly such as Pladeinolide B (PlaB), which inhibits the U2 snRNP component SF3b. Splicing inhibition was performed in HeLa S3 cell lines under the hypothesis that the fraction of detected unspliced reads increases if the computational pipeline correctly identifies *bona fide* unspliced reads. Nuclei from PlaB-treated and control datasets were extracted for preparation of FLAM-Seq sequencing libraries to enrich for unspliced RNA from nuclei. Extracting RNA directly from nuclei resulted in a 10-fold increase in unspliced reads (Figure 19 A) comparing Hela S3 bulk and nuclei preparations. Inhibiting splicing with spliceosome inhibitor PlaB further increased the fraction of intronic reads in nuclei threefold to around 1.3% of total sequenced reads, which validated that the performed analysis accurately identifies unspliced reads. Poly(A) tail length distributions of all reads were slightly shifted towards shorter tails upon PlaB treatment, although a longer poly(A) tail length mode persisted (Figure 19 B).

Poly(A) tail length distributions of intronic reads were slightly shifted towards longer intronic tails upon splicing inhibition from 205 nt in control to 242 nt in PlaB samples. Interestingly poly(A) tail length of intronic reads from nuclear HeLa S3 preparations were longer than those for RNA extractions from whole cells (Figure 19 C), which first hinted at synthesis of long poly(A) tails also in HeLa S3 cancer cell lines, and second shows that HeLa S3 cytoplasm contained RNAs with retained introns. The fraction of genes with associated intronic reads was compared to all expressed genes which showed that 1–4% of all detected genes had associated unspliced reads, which was comparable to bulk FLAM-Seq samples (Figure 16 F). The minimal overlap of genes with intronic reads (Figure 19 D) further illustrated the random sampling process for intronic read detection also upon splicing inhibition in nuclear preparations. Binning genes by expression and investigating poly(A) tail length of unspliced reads by expression bin showed that unspliced reads poly(A) tail length was uniformly around 205 nt in control and 245 nt upon PlaB inhibition (Figure 19 E). This difference was statistically significant for each expression bin showing that the PlaB treatment increased the poly(A) tail length of unspliced mRNAs. For each bin, the fraction of unspliced reads for each gene expression bin was computed (Figure 19 F). The fraction of unspliced reads decreased for higher expressed genes as observed before, and splicing inhibition further increaseed the fraction

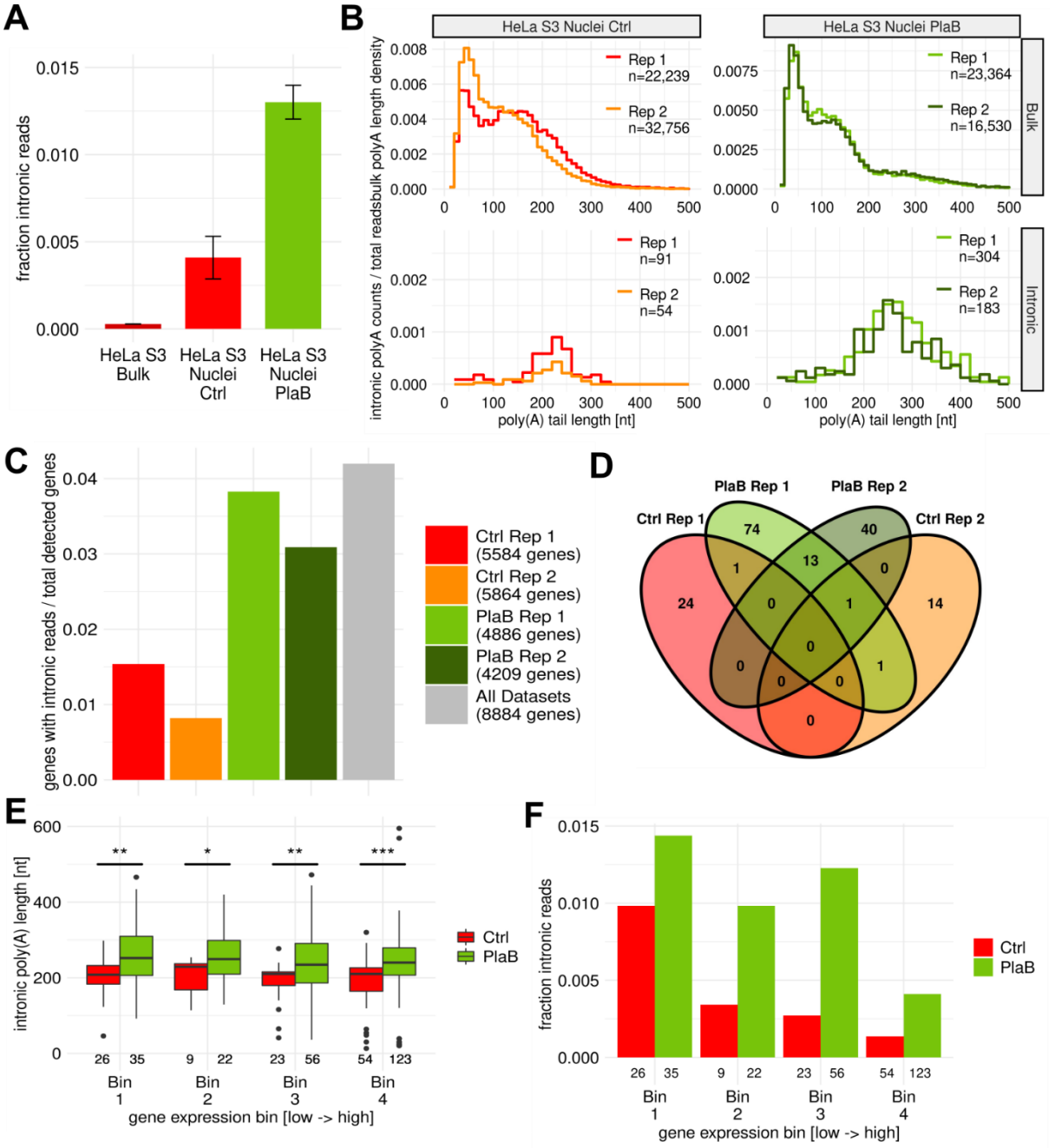

Figure 19 **PlaB splicing inhibition and effects on poly(A) tail length**
A) Fraction of intronic reads by total reads per replicate for HeLa S3 bulk, nuclei control and nuclei from cells treated with PlaB. **B)** Poly(A) tail length density distributions of nuclei from HeLa S3 cell lines treated with PlaB and control for total sequenced reads (top) and detected unspliced, intronic reads as fraction of total reads (bottom). **C)** Fraction of genes with detected unspliced, intronic genes by all genes detected per sample. **D)** Overlap of genes with detected intronic reads between replicates of PlaB splicing inhibition experiments. **E)** Poly(A) tail length of intronic, unspliced reads for PlaB-treated and control HeLa S3 samples binned by total expression counts of associated genes. Number of reads in each bin are displayed above the x-axis. **F)** Fraction of intronic reads to total reads for each gene expression bin for control and PlaB-treated samples.

of unspliced reads as expected. The second highest expression bin thereby did not follow this linear trend upon PlaB treatment, which could be an effect of distinct gene sets which are particularly affected by splicing inhibition.

Splicing inhibition through PlaB caused changes in median poly(A) tail length per gene. Differences in poly(A) tail length were binned and average expression of genes in each bin was plotted on histograms (Figure 20 A). The average difference across all genes was a 11 nt decrease in poly(A) length upon treating HeLa S3 cells with splicing inhibitor PlaB. Genes with mild changes in median poly(A) tail length per gene had the highest expression, although several highly expressed genes had a significantly increased poly(A) tail length upon splicing inhibition. This could hint at genes which respond to changes in RNA processing with overall increased poly(A) tail length which may stabilize existing transcripts. To evaluate the relationship between changes in poly(A) tail length and changes in expression levels upon PlaB treatment, differences in poly(A) tails length were compared against fold changes (Figure 20 B). No correlation was observed, yet some genes were identified which responded to increased expression upon splicing inhibition with poly(A) tail lengthening (MYC, IER3), or decreases in expression with longer (STX10, MAT2A) or shorter poly(A) tails (RBMX, KLF5). Closer investigation of the IER3 locus showed a consistent upregulation and poly(A) tail lengthening after inhibiting splicing (Figure 20 C). Investigation of molecular features which may explain the different behaviors of certain genes upon PlaB treatment revealed that genes with changes in poly(A) tail length are on average less stable (Figure 20 D). In particularly genes with increased poly(A) tail length upon splicing inhibition had overall short half lives. The latter group of genes also had longer 3'-UTRs (Figure 20 E).

In summary, splicing inhibition experiments showed first, that the developed computational pipeline accurately detected unspliced reads together with their poly(A) tail length. The analysis thereby validated the hypothesized synthesis of long poly(A) tails on a genome-wide level. No molecular differences could be found for genes with unspliced reads, which supported the notion that detected intronic reads were representative of genome-wide poly(A) tail biogenesis. Second, the FLAM-Seq analysis of nuclear RNA greatly facilitated detection of unspliced reads and third it was shown that inhibition of splicing led to global shortening of poly(A) tails for most genes, with some exceptions such as MYC. On the contrary, poly(A) tails of unspliced, intronic reads showed an increase in poly(A) tail length as a response to splicing inhibition.

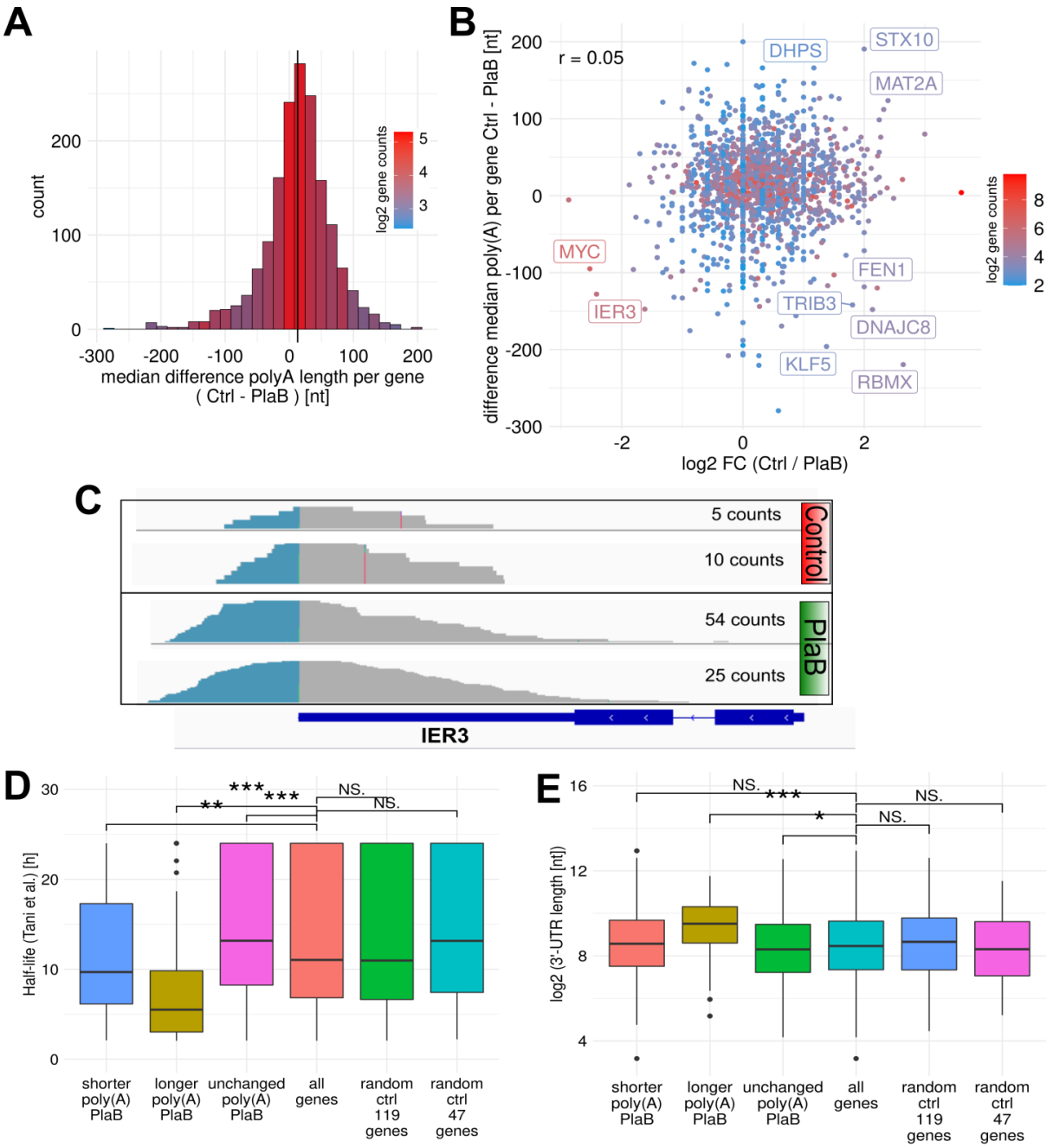

Figure 20 **Poly(A) tail length differences upon PlaB splicing inhibition**
A) Difference in median poly(A) tail length per gene between control and PlaB treated HeLa S3 cell lines and average gene expression per poly(A) difference bin. **B)** Difference in median poly(A) tail length per gene against gene expression fold-change between HeLa S3 control and PlaB-treated samples. **C)** Browser shotes of IER3 gene locus with aligned reads from FLAM-Seq HeLa S3 control and PlaB replicates. Poly(A) tails were appended to alignments in blue. **D)** Half-lives per gene (from Tani et al.) for genes binned by changes in median poly(A) tail length between control and PlaB treated HeLa S3 samples. Size-matched random gene sets were used as a control. Half-lives of binned genes were compared to all genes by Wilcoxon test. **E)** 3'-UTR length per gene for genes binned by changed in median poly(A) tail length per gene between control and PlaB treated HeLa S3 samples. Size-matched random gene sets were used as a control. Size-matched random gene sets were used as a control. Half-lives of binned genes were compared to all genes by Wilcoxon test.

4.2.3 Direct RNA sequencing of nascent, chromatin associated total RNA validates synthesis of long poly(A) tails beyond polyadenylated RNAs

The FLAM-Seq protocol relies on extraction of polyadenylated RNA before GI-tailing and library preparation. For the presented analysis of unspliced reads in FLAM-Seq data, this requires reads to be both unspliced and at the same time polyadenylated. Since splicing is for most genes co-transcriptional [38], most introns should be spliced before the poly(A) tail is added. Other studies yet concluded that post-transcriptional splicing is widespread [43] and splicing patterns of terminal introns, which are most likely to be covered by FLAM-Seq, may be kinetically distinct and coupled to polyadenylation.

To further address whether the analysis of post-transcriptionally spliced molecules limits the general hypothesis for synthesis of long poly(A) tails to post-transcriptional splicing, published Nanopore direct RNA sequencing datasets from K562 cell lines were analyzed. Nanopore sequencing was here performed on chromatin associated RNA which was purified by streptavidin pulldown after 8 minutes labeling of cells with 4-thiouridine, which is incorporated into newly synthesized RNA [43]. For Nanopore analysis, rRNA depleted total RNA, was either directly sequenced or polyadenylated *in vitro* to increase the fraction of non-polyadenylated RNAs. Obtained read ends were categorized for each read as ending in introns (‘intron’), ending at annotated polyadenylation sites (‘polyA’) or ending downstream of annotated polyadenylation sites (‘post_polyA’) for nascent mRNAs which are not cleaved yet. Reads were then categorized into spliced and unspliced groups (Figure 21 A), by processing alignments using the computational pipeline as above. The computational pipeline designed for FLAM-Seq datasets had yet limited sensitivity in identifying unspliced reads, since the high error rates of Nanopore sequencing led to a many fragmented alignments. The large number of reads that ended in intron annotations and were at the same time ‘spliced’ illustrate this limited sensitivity. In principle all reads ending in annotated introns should be regarded as ‘unspliced’ since RNAPII has not completed synthesis here. As stated by Drexler et al., the experimental protocol without the poly(A) tailing step (‘no tailing’) was mostly enriched for polyadenylated molecules which ends aligned to annotated polyadenylation sites. This is comparable to the results from FLAM-Seq, since poly(A) tail length was similarly around 200 nt for newly synthesized, chromatin associated RNA, both for spliced and unspliced reads. *In vitro* poly(A) tailing of total RNA enriched mostly for read ends aligning to intronic sequences but also a large number aligning downstream of annotated polyadenylation sites (‘post poly(A)’).

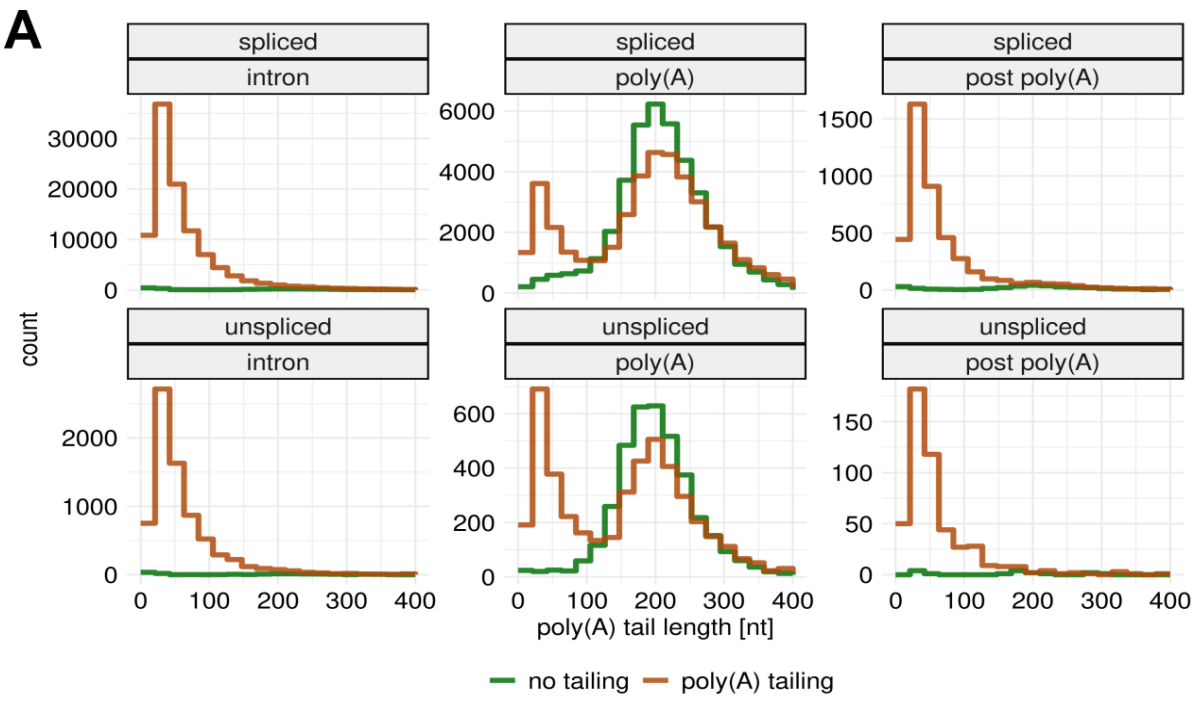

Figure 21 **Validation of poly(A) tail length for unspliced reads from Nanopore direct RNA sequencing A)** Poly(A) tail length distributions of reads from 'no tailing' and 'poly(A) tailing' nascent RNA preparation methods from Drexler et al.. Reads were grouped by read ends aligning in introns / gene bodies ('intron') or annotated polyadenylation sites ('poly(A)') or downstream of polyadenylation sites ('post poly(A)'). Reads were classified as 'spliced' or 'unspliced' bins by the computational pipeline used for FLAM-Seq annotations.

Both spliced and unspliced read ends aligning at introns or 'post poly(A)' sites had poly(A) tail length of on average less than 50 nt, which were likely the product of *in vitro* tailing and indicated absence of endogenous poly(A) tails. Poly(A) tails of reads at poly(A) sites were either around 50 nt which would correspond to the absence of endogenous poly(A) tails or around 200 nt which resembles the length profiles observed in FLAM-Seq for unspliced reads. This bimodal distribution argued in favor of a model in which poly(A) tails are either completely absent or synthesized as long tails without evidence for synthesis of intermediate tail length.

The analysis of Nanopore mRNA sequencing datasets validated that long poly(A) tails were detected for completely transcribed and cleaved nascent RNAs, independent of their splicing status. This suggested that long poly(A) tails were also synthesized upon co-transcriptional splicing. The computational analysis was yet less specific with respect to identification of unspliced reads, likely driven by higher error rates of Nanopore sequencing. Poly(A) tails detected for non-cleaved RNAs were most likely resulting from *in vitro* polyadenylation such that 'endogenous' addition of a poly(A) tail can be excluded here.

4.2.4 Rapid shortening of poly(A) tails revealed by metabolic labeling of RNA

Steady state poly(A) tail length distributions were shorter than the described length of 200 or more nucleotides at the point of poly(A) synthesis. This posed the question of how poly(A) tails converge towards steady state distributions over time and whether this shortening process is gene specific. Two orthogonal genome-wide methods for measuring RNA dynamics over time were combined with FLAM-Seq to quantify poly(A) tail length over time. First, metabolic labeling of RNA using 4-thiouridine (4sU) and pulldowns of labeled, biotinylated RNA was used in conjunction with FLAM-Seq to profile poly(A) tails of labeled RNA fractions over time[53,375]. Second, the SLAM-Seq protocol, which combines metabolic labeling of RNA with chemical derivatization of 4sU, was used together with an adapted FLAM-Seq library preparation procedure. RNA synthesized within the respective labeling periods could then be detected based on T-to-C mutations in sequencing reads, which are introduced during cDNA synthesis through incorporation of complementary cytosines at positions of 4sU derivatization.

Metabolic labeling using 4sU was performed for 0, 10, 15, 20, 45 and 90 minutes in replicates. Replicates were then pooled for the streptavidin pulldown steps performed for each timepoint, which was necessary given the high input requirements of the FLAM-Seq protocol. Labeling was performed in HEK Flp-In T-rex cell lines. RNA concentrations in pulldown fractions were expected to be proportional to the labeling durations, since more RNA is produced for longer labeling timepoints and the average RNA half-life is with around 4 h [120] longer than the labeled timepoints such that RNA decay should have little impact. RNA concentrations ranged from 10 to 420 ng/µL, which corresponded to around 1-10% of the total cellular RNA pool as quantified by the supernatant RNA concentrations (corresponding to a concentration of around 4000 ng/µL) (Figure 22 A). The fraction of biotinylated RNA was proportional to the labeling time and additionally compared by dot blots, streptavidin-HRP incubation and chemiluminescence detection (Figure 22 B). Dot blots also showed the expected increase in biotinylated RNA. After validating 4sU incorporation proportional to labeling periods, FLAM-Seq libraries were prepared from labeled pulldown (PD) fractions and unlabeled supernatant (SN) fractions. A FLAM-Seq library was also obtained for the labeled pools of the 0 min timepoint. The 0 min sample contained a small fraction of RNA (ca. 25% of 10 min labeling timepoint), which was most likely due to RNA which unspecific bound to streptavidin beads. Poly(A) tail length profiles of pulldown fractions were bimodal, with peaks around 50 nt and 150 nt and little differences between individual labeling timepoints (Figure 22 C). The supernatant poly(A) profiles were shorter than the labeled fraction, while the 0 min labeling

control was slightly longer than the supernatant. Control poly(A) distributions may hint at preferential background binding of longer poly(A) tails to streptavidin beads. Nonetheless, poly(A) distributions of all labeled pulldown fractions were longer than the 0 min control. Poly(A) tail length distributions of supernatant fractions for individual timepoints were highly reproducible, which was indicating the absence of biases in poly(A) tail length quantification for individual labeling timepoints (Figure 22 D).

To ensure that metabolic labeling did not bias quantification of long tails, intronic reads were extracted by applying the developed pipeline on merged pulldown and supernatant samples. The fraction of unspliced reads in merged pulldown samples was three times higher than in supernatant samples (supernatant: 0.04%, pulldown: 0.16%), which was expected given that RNA splicing operates within timescales of minutes [46]. Occupying 0.05% of total reads, the fraction of unspliced fractions in supernatants of metabolic labeling was comparable to 0.03% unspliced reads found in bulk HeLa S3 FLAM-Seq samples (Figure 16 B). Poly(A) tail length of unspliced reads was around 200 nt (Figure 22 F). This validated the synthesis of long poly(A) tails as observed for other human model systems before. As additional proof for the absence of biases between labeled and supernatant samples, poly(A) tail length was quantified for mitochondrial genes, which uniformly showed an average poly(A) tail length of around 50 nt (Figure 22 G), which corresponds to the poly(A) tail length found in bulk HeLa S3 RNA preparations (Figure 6 F) and has been reported previously [151]. No length difference was observed between labeled and unlabeled mitochondrial poly(A) tails, which indicated a lack of deadenylation on the investigated timescales.

Median poly(A) tail length per gene was compared between pulldown and supernatants for the investigated timepoints (Figure 22 H). Average difference poly(A) tail length between newly synthesized and pre-existing RNA was 25-32 nt, with no clear ranking of poly(A) length differences by the labeling time. The difference was 9 nt when comparing the 0 min control timepoint. Poly(A) tail length was also investigated for different gene sets since the sequencing depth was insufficient for exploring individual genes. Comparing poly(A) tails of immediate early genes (IEGs) and ribosomal genes showed that ribosomal genes had relatively short poly(A) tails after 10 minutes of labeling with an average length of around 110 nt, whereas IEGs had much longer tails of around 190 nt after 10 minutes labeling which were progressively shortened. This trend was strongest for lncRNAs, which had poly(A) tails longer than 200 nt after 10 min labeling (Figure 22 I).

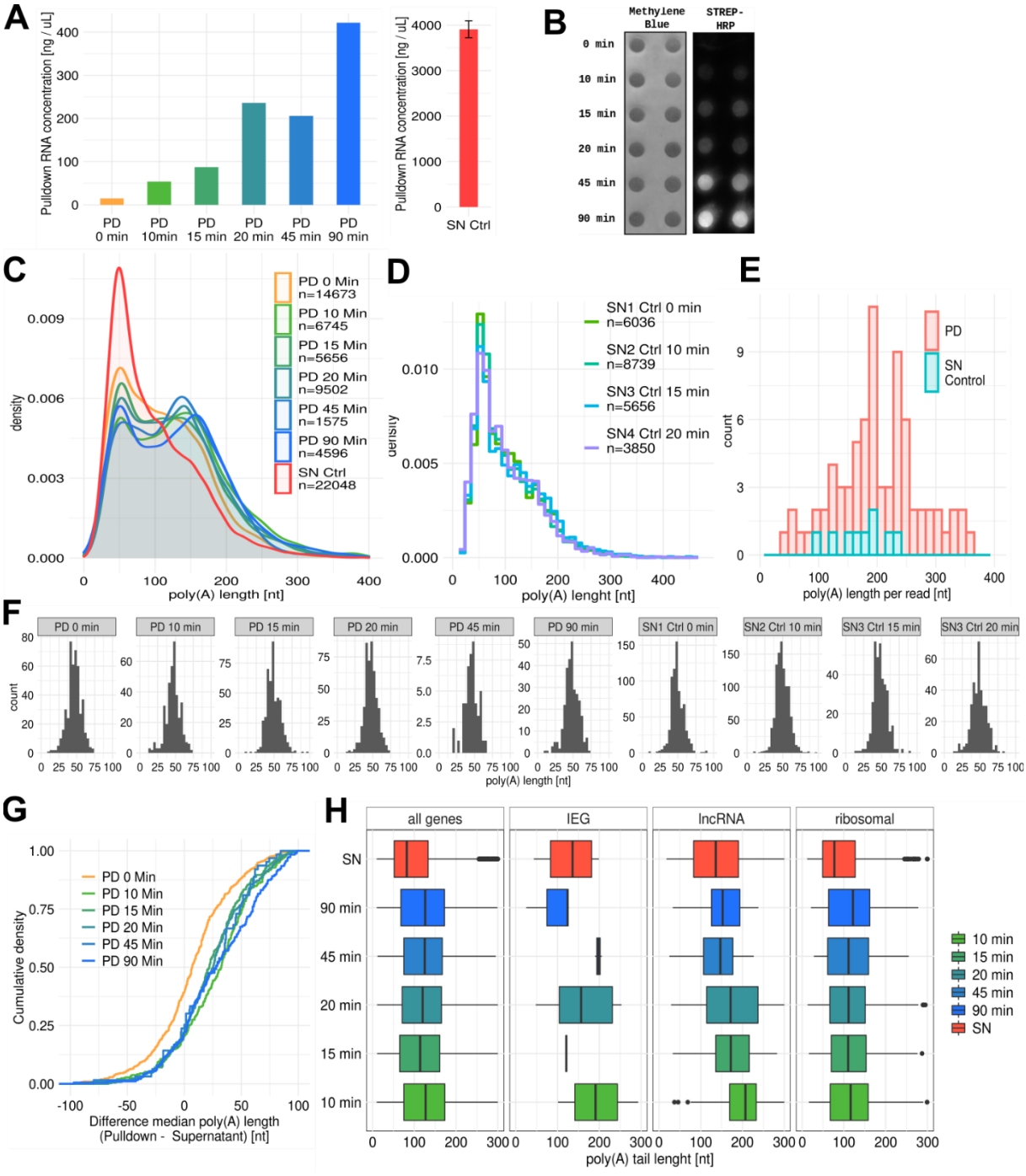

Figure 22 **RNA metabolic labeling and pulldown reveals poly(A) tail dynamics of newly synthesized RNA** **A)** RNA concentrations of pulldown (PD) fractions after biotinylation and streptavidin pulldown. RNA was labeled for indicated time intervals. Right bar indicates average concentrations of supernatant (SN) fractions. **B)** Dotblots for biotinylated RNA from pulldown (PD) fractions. Left: methylene blue staining for RNA quantity, Right: ECL detection of Strep-HRP antibody labeling. **C)** Poly(A) tail length density profiles of FLAM-Seq samples for RNA metabolic labeling for indicated labeling timepoints and a merged supernatant distribution. **D)** Poly(A) tail length density distributions for supernatant fractions. **E)** Poly(A) tail length distribution of intronic reads in pulldown (PD) and supernatant (SN control) fractions. **F)** Poly(A) tail length distributions for mitochondrial genes in PD and SN fractions. **G)** Difference in median poly(A) tail length per gene between individual pulldown corresponding supernatant fractions. **H)** Poly(A) tail length distribution of different gene sets in pulldown fractions for different labeling timepoints and supernatant fraction.

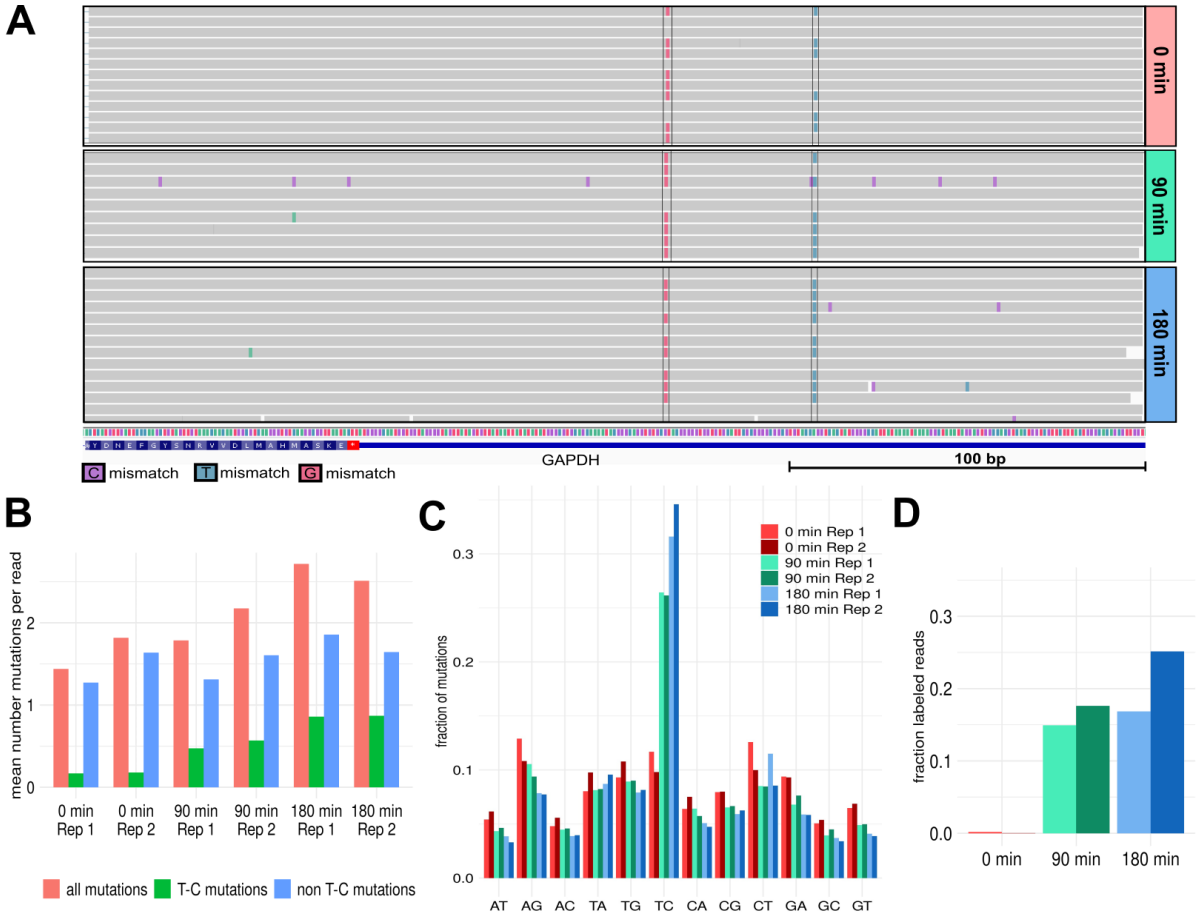

Figure 23 **SLAM-Seq and poly(A) profiling as orthogonal approach for analysis of poly(A) tail dynamics A)** Browser shot for GAPDH locus and different 4sU labeling intervals. Mismatches bases in read alignments are visualized. Vertical bars indicate likely heterozygous single nucleotide variants. **B)** Average mutations per read for SLAM-Seq replicate samples grouped by all mutations, T-to-C mutations only and non-T-to-C mutations. **C)** Frequencies of nucleotide conversions for SLAM-Seq datasets. **D)** Fraction of labeled reads in SLAM-Seq datasets for replicates and different timepoints.

SLAM-Seq was used in combination with FLAM-Seq library preparation for quantification of poly(A) tail length of newly synthesized RNA without requiring separation of labeled and unlabeled RNA. The SLAM-Seq method relies on labeling cells with 4sU, which is incorporated into newly synthesized RNA. RNA is then extracted and incorporated 4sU is chemically derivatized by addition of iodoacetamide. RNA is then reverse transcribed into cDNA, which leads to introduction of T-to-C conversions when reverse transcribing RNA at positions with derivatized 4sU, where guanosines instead of adenosines are incorporated. The FLAM-Seq protocol was adapted for these experiments: for chemical derivatization of 4sU labeled RNA, SLAM-Seq requires harsh treatment at 50°C and basic pH which could potentially lead to RNA hydrolysis. Since hydrolysis may lead to mis-quantification of poly(A) tail length, poly(A) selection and GI-tailing was performed before introducing T-to-C conversions which required harsh incubation of RNA. Possible RNA hydrolysis then leads to a loss of the GI-tail, which would exclude those molecules from reverse transcription, that

requires the GI-tail for binding the oligo-dC primer. Newly synthesized RNA was identified based on T-to-C mutations in aligned reads with respect to the reference genome (Figure 23 A). T-to-C-mutations were increased for longer labeling timepoints and were randomly scattered throughout individual reads. This was in contrast to detected polymorphisms, which were present at identical positions on a larger fraction of reads. SLAM-Seq was performed in HeLa S3 cell lines with 0 min, 90 min and 180 min 4sU labeling in replicates. Investigating the mutation profiles and mean number of mutations per read in each dataset showed an overall increase from around 1.6 mutations per read in control to 2.6 mutations after 180 min labeling. As expected, the increase was mostly driven by more T-C mutations which increased from 0.17 for 0 min to 0.85 after 180 min (Figure 23 B). Calculating statistics for different classes of observed nucleotide conversions ('mutations') showed that all possible conversions were covered at a comparable level after 0 min labeling. Each observed conversion made up between 5-10% of all conversions, which fluctuated around the expected 8.3% for 12 different possible mutations (Figure 23 C). T-to-C conversion were most enriched upon 4sU labeling for 90 and 180 min, which was expected. 4sU labeling had else no apparent effect on distorting the ratios between the detected conversions compared to 0 min labeling.

A statistical model, comparable to the GRAND-SLAM approach[356], was implemented which identified reads coming from labeled RNA by computing the log-likelihood of observing the detected number of T-C conversion under a labeling process or a background model, where T-C mutations were assumed to come from sequencing errors only. The model identified between around 16% of reads as labeled after 90 min labeling and 22% after 180 min labeling, while less that 1% labeled reads were found for 0 min controls (Figure 23 D), which showed that the computational model was highly specific in detecting labeled RNA. Comparing poly(A) tail length of labeled reads showed an increase in poly(A) tail profiles of newly synthesized RNA compared to total reads, which represented the steady-state poly(A) tail length distribution (Figure 23 A), with reasonable agreement between replicates. Comparing total poly(A) tail length distributions in between labeling timepoints yet shows that poly(A) profiles differed between samples: 0 min and 90 min profiles were in good agreement but showed increased poly(A) tail length distributions comparing to bulk HeLa S3 profiles (Figure 7 A). 180 min labeling samples had shorter poly(A) tail profiles which were more reflective of bulk HeLa S3 profiles. The differences in global poly(A) profiles were not expected since the SLAM-Seq approach should preserve the global structure of the RNA pool and poly(A) tail length. Those disagreements could hint at experimental problems in uniformly handling small quantities of

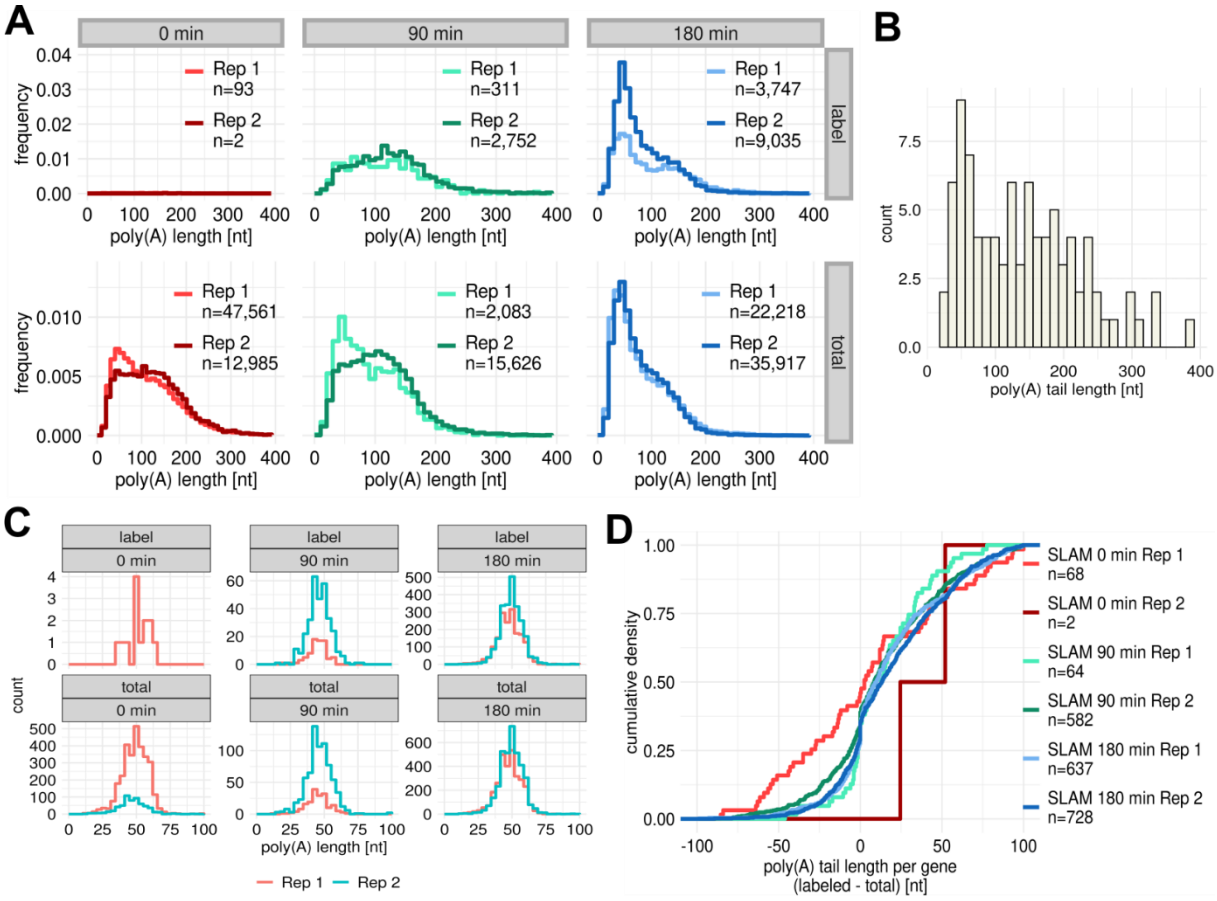

Figure 24 **SLAM-Seq poly(A) profiling for different labeling periods**
A) Poly(A) tail length density distributions for labeled and total reads for 4sU labeling intervals from SLAM-Seq datasets. **B)** Poly(A) tail length distributions of intronic reads detected in merged SLAM-Seq datasets. **C)** Poly(A) tail length of labeled and total read of mitochondrial in SLAM-Seq datasets for labeling timepoints. **D)** Difference in median poly(A) tail length per gene between labeled and total read fractions for SLAM-Seq labeling timepoints and replicates.

RNA in the modified SLAM-Seq / FLAM-Seq protocol or undesired effects of RNA degradation.

Extracting intronic reads from merged SLAM-Seq samples showed that poly(A) tails of intronic reads had a median length of 136 nt (Figure 23 B). This was much shorter than the 200 nt observed for other mammalian samples for unspliced reads, but in essence reflected the intronic poly(A) distributions measured for bulk RNA from HeLa S3 bulk cells (Figure 16 C). As an additional control, poly(A) tail length of mitochondrial genes was around 50 nt both for labeled and total read bins and across different labeling timepoints (Figure 23 C), which confirms the absence of differences in poly(A) also observed in pulldown experiments. Comparing median poly(A) tail length per gene between labeled and total reads showed that most genes had longer poly(A) tails after 90 and 180 min labeling compared to steady-state with a difference of 18 to 23 nt and no remarkable differences between 90 min and 180 min (Figure 23 D).

Metabolic labeling experiments in summary revealed that poly(A) tails were shorter than the 200 nt at their point of synthesis even after relatively short labeling times as 10 min. Beyond this, differences in poly(A) profiles between individual labeling timepoints were small, which suggested slow(er) deadenylation which may not be resolvable by the lower depth of FLAM-Seq. Those findings were in principle supported by combining FLAM-Seq and SLAM-Seq, yet some technical problems were observed here which were related unexpected differences in steady state poly(A) profiles between labeling timepoints, which limited the significance of results obtained from the SLAM-Seq approach.

4.2.5 Subcellular fractionation hints at nuclear deadenylation

Metabolic labeling experiments in combination with FLAM-Seq uncovered shortening of poly(A) tails within the first 10 minutes after completion of transcription, assuming global synthesis of long poly(A) tails. Since RNA exports operates on comparable time scales [376,377], the question remained to what extend the hypothesized shortening is a nuclear or cytoplasmic process.

To address this questions, HeLa S3 cell lines were biochemically separated into cytoplasmic, nucleoplasmic and chromatin fractions. A total of 6 biological replicates with each two technical replicates were prepared from untreated HeLa S3 cells (including control samples from non-induced shRNA expressing HeLa cell lines, s. below) across a time interval of more than a year to account for possible technical variation which is known to be inherent to biochemical fractionation protocols [378]. In brief, HeLa S3 cells were harvested and incubated with lysis buffer containing NP-40 detergent to dissolve the cell membranes. Cells were then centrifuged through a sucrose cushion to separate nuclei from cytoplasm and the cytoplasmic fraction was collected. Nuclei were then incubated with a second lysis buffer, dissolving the nuclear membrane, and separating nucleoplasm from chromatin pellets after centrifugation (Figure 25 A).

Input samples (total lysate), cytoplasmic and nucleoplasmic fractions were analyzed by Western Blot for potential cross-contamination of nuclear fractions with cytoplasmic components. GAPDH was used as a cytoplasmic marker, along with BCAP31, which is an ER marker protein [379,380] and more indicative of cytoplasmic contamination since rough ER is more likely to remain attached to isolated nuclei. TDP43 was used as a marker for both nucleoplasm and cytoplasm (Figure 25 B), since it shuttles between both compartments [381]. GAPDH signal was absent from any nucleoplasmic fractionation experiments and faint bands for BCAP31

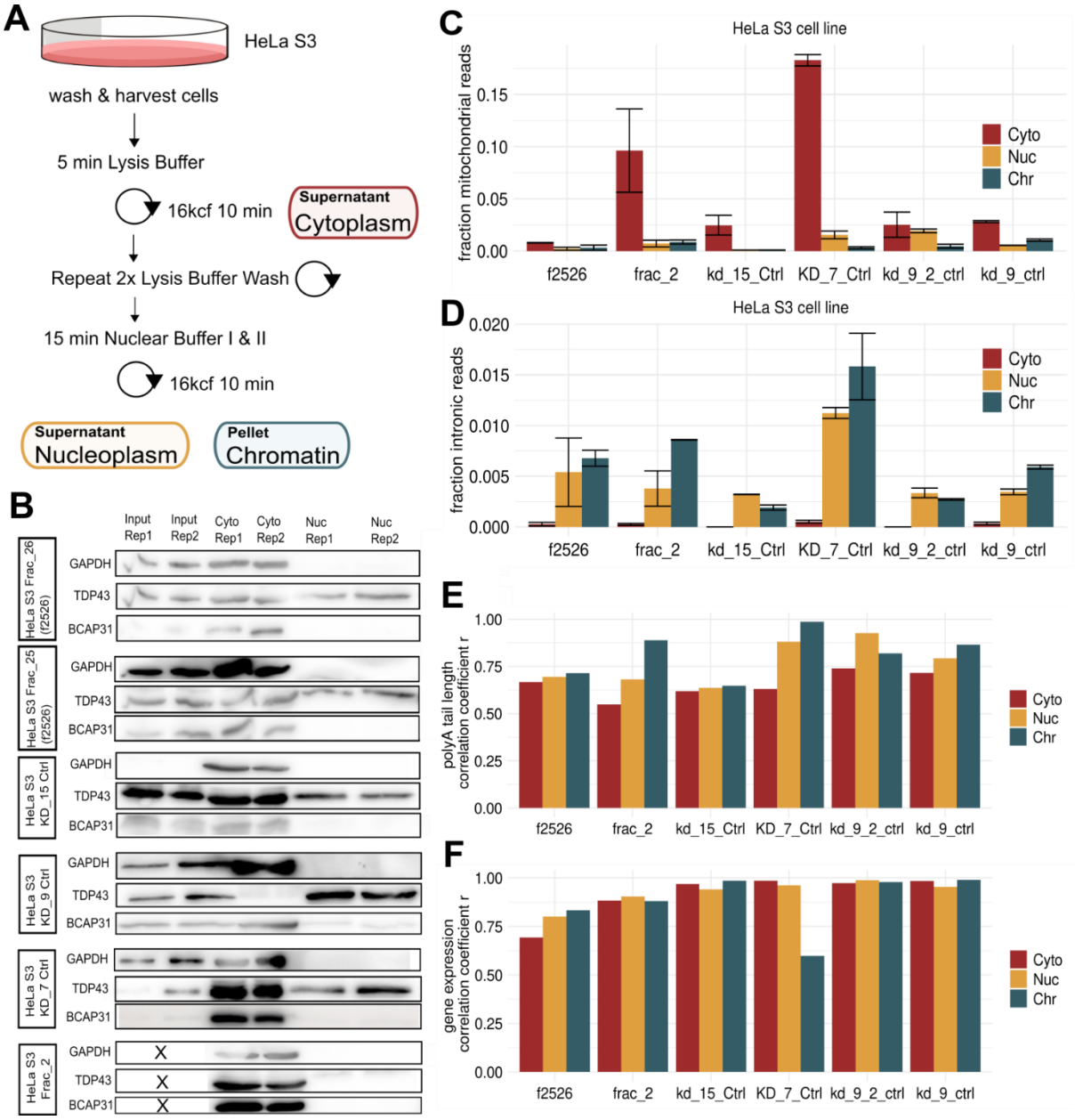

Figure 25 **Characterization of cytoplasmic, nucleoplasmic and chromatin fractions from HeLa S3 cells**
A) Schematic outline of experimental protocol for biochemical fractionation of HeLa S3 cell line in cytoplasm, nucleoplasm and chromatin fractions. **B)** Western blot analysis of subcellular fractions from HeLa S3 cell lines for markers GAPDH (cytoplasm), TDP43 (cytoplasm & nucleus) and BCAP31 (cytoplasm, ER). **C)** Fraction of mitochondrial reads in subcellular fractions for HeLa S3 replicates. Error bars denote standard deviation. **D)** Fraction of intronic reads in subcellular fractions for HeLa S3 replicates. Error bars denote standard deviation. **E)** Pearson correlation coefficients between median poly(A) tail length per gene for technical replicates of HeLa S3 samples, for genes with more than 10 counts. **F)** Pearson correlation coefficients between gene expression counts per gene for technical replicates of HeLa S3 samples, for genes with more than 10 counts.

was visible only in one set of biological replicates (KD_9 Ctrl). FLAM-Seq sequencing libraries were produced from RNA extracted in each fraction. To assess the purity of sequencing libraries with respect to potential contamination, the relative proportion of sequenced molecules from mitochondrial genes was investigated, which should only be found in cytoplasm. Between 2% and 17% of all cytoplasmic reads were mitochondrial transcripts, which underscored the inherent variability of the experiments (Figure 25 C).

The fraction of mitochondrial RNA in nucleoplasmic fractions was between 0.1% and 2% and 0.1% - 1% in chromatin fractions. In all cases, the nuclear fraction of mitochondrial RNA was smaller than the cytoplasmic fraction with a ratio of cytoplasmic to nuclear mitochondrial reads ranging from around 70% to less than 10%. The highest relative fraction of nuclear mitochondrial reads was observed for sample KD_9_2_Ctrl which also showed a faint BCAP31 band. This showed that nuclear mitochondrial reads were reflective of cytoplasmic contamination. To assess nuclear contamination in the cytoplasm, the fraction of intronic reads was analyzed for each fraction using the computational pipeline outline above. Between 0% and 0.06% of cytoplasmic reads was found to be unspliced. As expected, this number was much higher for the nuclear fractions, where between 0.17% and 1.85% of all reads were identified as being unspliced (Figure 25 D). Comparing median poly(A) tail length per gene between technical replicates showed good agreements, with Pearson correlation coefficients ranging from 0.52 to 0.98 (Figure 25 E). The same reproducibility was observed for gene expression counts between technical replicates with Pearson correlation coefficients ranging from 0.48 to 0.94 (Figure 25 F). Biochemical fractionation experiments in HeLa S3 cell lines were in summary shown to be mostly free of detectable cytoplasmic contamination with except of one replicate. Investigation of quality control parameters such as mitochondrial reads across different fractions yet illustrated the inherent variability of the experimental method which was mostly related to batch effects and day-to-day variation, since technical replicates were very reproducible. To further investigate experimental variables which best explain observed differences in poly(A) tail length profiles between replicates for a given fraction, a linear model was fitted where median poly(A) tail length observed for each fraction was modeled as a function of experimental variables such as ‘fraction mitochondrial reads’ or ‘RNA concentration’ obtained from each experiment (detailed description in 3.3.20). The factor which had greatest impact on describing median poly(A) tail length per gene was a poly(A) tail length scaling factor, which was calculated for all fractions of an experiment and describes the deviation from the average poly(A) tail length per gene calculated across each experiment and each fraction.

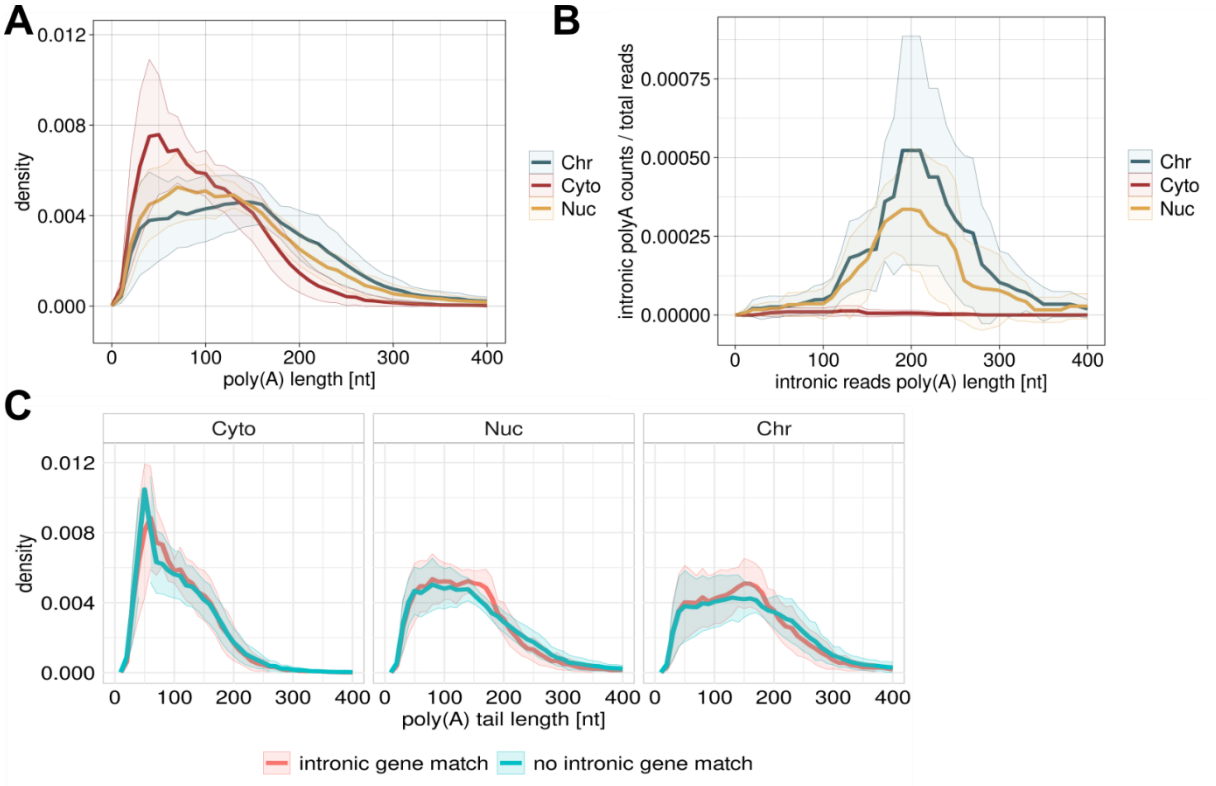

Figure 26 **Poly(A) tails in nuclear compartments are shorter than intronic poly(A) tails**
A) Average poly(A) tail length distributions for subcellular fractions of HeLa S3 replicates. Error margins refer to standard deviation across replicates. **B)** Poly(A) tail length distributions of intronic reads for HeLa subcellular fractions. Error margins refer to standard deviations across replicates. **C)** Poly(A) tail length distributions for genes with associated intronic reads ('intronic gene matched') and poly(A) tail length distributions of genes without intronic reads ('no intronic gene match').

For further analysis, the resulting poly(A) tail length distributions for all replicates were averaged and the standard deviation could be calculated for poly(A) length profiles of subcellular fractions. Comparing poly(A) tail length distributions between HeLa S3 fractions indicated progressive shortening of poly(A) tails (Figure 26 A). Median poly(A) tail length in cytoplasmic fractions was 80 nt, which was less compared to nuclear fractions with a median tail length of 134 nt in chromatin and 117 nt in nucleoplasm. Standard deviations for poly(A) tail distributions also showed that the differences in poly(A) tail profiles were unlikely to be random effects related to variability between biological replicates. Median poly(A) tail length per gene showed more pronounced differences between fractions with a median length of 108 nt in cytoplasm, 144 nt in nucleoplasm and 164 nt in chromatin fractions. Intronic poly(A) tail length profiles had a median length of 205 nt in chromatin and nucleoplasm, and poly(A) distributions where mostly indistinguishable (Figure 26 B). The observed length profiles also matched the intronic poly(A) tail length described above for nuclear RNA preparations from

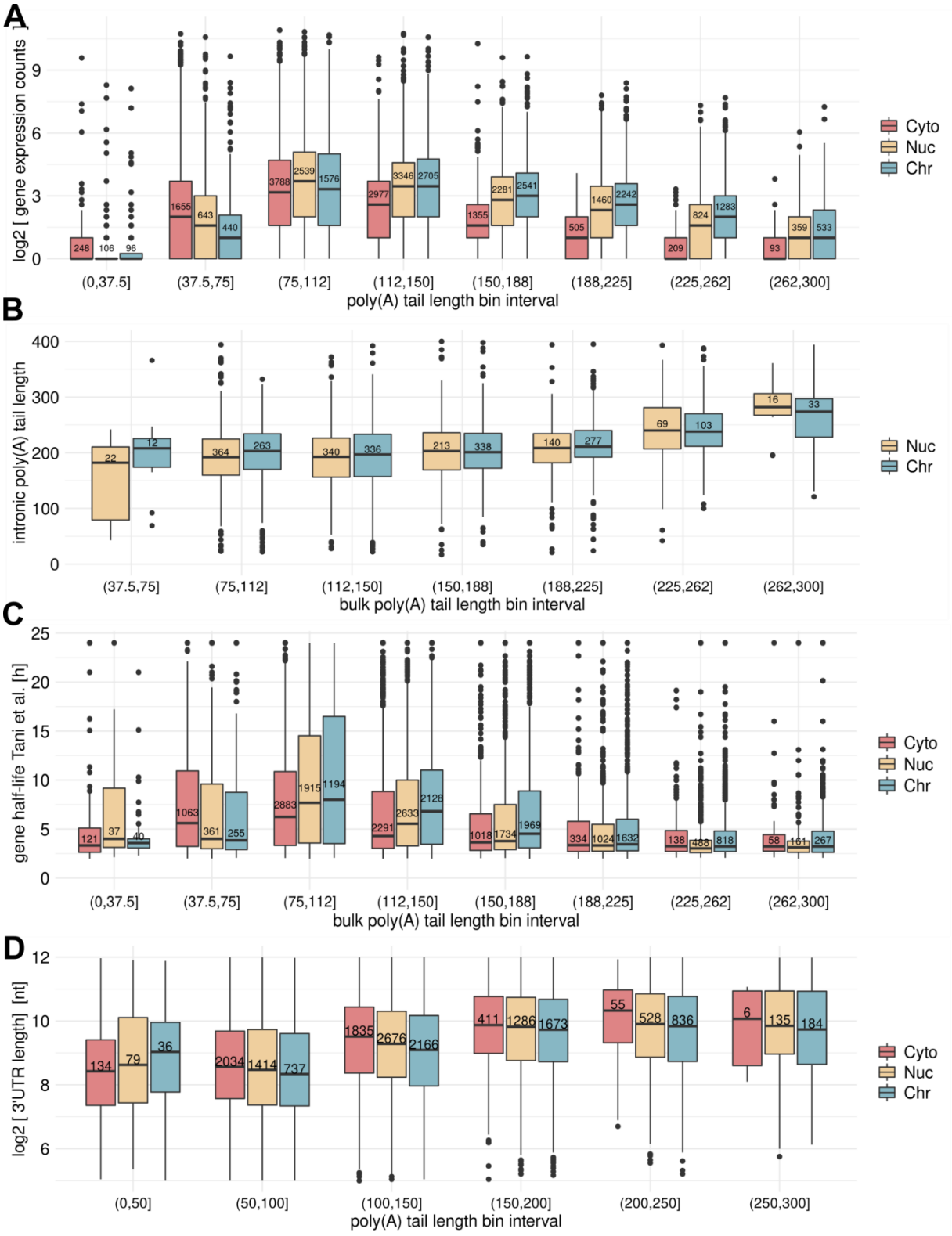

Figure 27 **Nuclear poly(A) tail length profiles correlate with similar molecular features as cytoplasmic poly(A) tails**

A) Gene expression counts for genes binned by median poly(A) tail length per gene in subcellular fractions. Numbers in bars represent reads for bin. **B)** Poly(A) tail length of intronic reads for genes binned by median poly(A) tail length across reads. Numbers in bars represent reads for bin. **C)** Half-lives of genes binned by median poly(A) tail length per gene. Numbers in bars represent reads. **D)** 3'-UTR length for genes binned by median poly(A) tail length per gene for HeLa S3 subcellular compartments. Numbers in bars represent reads.

HeLa S3 cell lines (Figure 19 B). Few intronic reads were detected in the cytoplasmic fractions which had a median poly(A) tail length of 130 nt. This supported the hypothesis that cytoplasmic transcripts with retained introns could in some cases be falsely annotated as nascent, unspliced reads. To ensure that poly(A) tail length distributions in subcellular fractions were comparable between genes with or without detected unspliced reads, those poly(A) tail length distributions were compared for each fraction (Figure 26 C). Poly(A) tail length was slightly increased for genes without detected intronic reads, standard deviations of length profiles overlapped.

To understand the molecular properties associated with poly(A) tail length in each subcellular fraction, genes were first binned by median poly(A) tail length and different molecular features were plotted for each bin. Comparing gene expression counts by poly(A) tail length bins revealed that highly expressed genes were associated with shorter poly(A) tails in cytoplasmic fractions compared to nuclear fractions, which is expected given the overall shifted poly(A) tail length distributions from chromatin to nuclear fractions (Figure 27 A). Comparing median poly(A) tail length per gene with the poly(A) tail length of unspliced reads for genes in each bin showed that intronic reads have universal long tails in nuclear fractions, also if the tails of spliced transcripts for the same gene were already short in the nucleus (Figure 27 B). This trend was different for genes with poly(A) tails which were longer than 250 nt. For those, the poly(A) tail length of unspliced reads was similarly increased. Half-lives for genes binned by poly(A) tail length showed that genes with most stable transcripts had short tails in the cytoplasmic fraction (Figure 27 C), while their poly(A) tail length was generally longer in the nuclear fractions. Genes with short poly(A) tails in nuclear fraction tended to also have less stable transcripts. 3'-UTRs were generally longer in cytoplasmic fractions for all poly(A) tail length bins except for very short poly(A) tails (Figure 27 D). In summary, nuclear poly(A) tails were found to be longer than cytoplasmic tails, but much shorter than the 200 nt at the point of synthesis without evidence that poly(A) profiles differ between genes with intronic tails versus genes without intronic tails. Investigating molecular features as expression, half-life, intronic tail length and 3'-UTR length showed a very similar relationship between those features and median poly(A) tail length per gene all subcellular fractions, but with a shift towards longer tails for all nuclear fractions.

To investigate nuclear poly(A) tails *in vivo*, similar biochemical fractionation experiments were performed on two hemispheres of a mouse brain: Cytoplasmic and nuclear fractions were separated using a Dounce homogenizer (Figure 29 A; Fractionation experiments by Maddalena Pacelli). Western Blots were performed to probe nuclear and cytoplasmic markers: Cytoplasmic

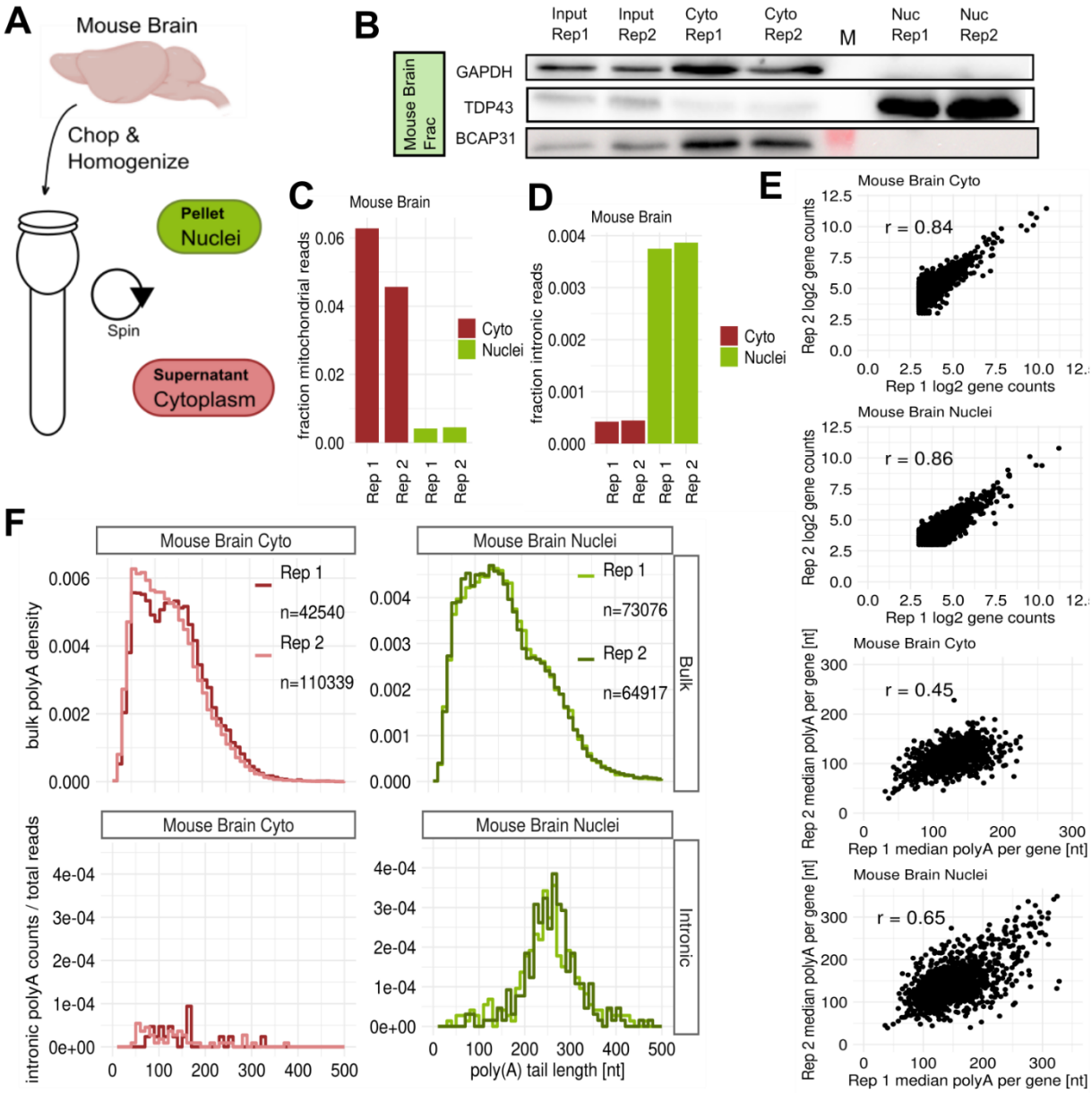

Figure 28 **Validation of nuclear poly(A) tail shortening *in vivo***
A) Schematic experimental outline for biochemical fractionation of mouse brain into nuclear and cytoplasmic fractions. **B)** Western Blot analysis of subcellular fractions from mouse brains for markers GAPDH (cytoplasm), TDP43 (cytoplasm & nucleus) and BCAP31 (cytoplasm, ER). **C)** Fraction of mitochondrial reads in subcellular fractions for mouse brain replicates. **D)** Fraction of intronic reads in subcellular fractions for mouse brain replicates. **E)** Correlation of gene expression counts and median poly(A) tail length per gene between replicates of mouse brain nuclear and cytoplasmic fractions. **F)** Poly(A) tail length distributions for mouse brain cytoplasmic and nuclear fractions for total ('bulk') and intronic reads.

marker GAPDH was not detected in the nucleus, and neither the ER marker BCAP31, which indicated absence of cytoplasmic contamination in the nucleus. TDP43 was detected both in nuclear and cytoplasmic fractions (Figure 29 B). The fraction of mitochondrial transcripts was around 10 times higher in cytoplasm compared to nuclear fractions, which indicates absence of cytoplasmic RNA contamination in nuclear fractions (Figure 29 C). The fraction of intronic reads was much higher in the nuclear fraction (Figure 29 D) which was also observed in HeLa S3 fractionation experiments. Comparing gene expression between replicates of cytoplasmic and nuclear fractions showed reproducible quantification of genes, with correlation coefficients of r = 0.84 and r = 0.86 (Figure 29 E). Comparing median poly(A) tail length per gene between replicates also showed decent agreement between replicates (r = 0.45-0.65). Comparing poly(A) tail length between nuclear and cytoplasmic fractions showed that the nuclear fraction had overall longer poly(A) tail length profiles with 146 nt compared to 119 nt in cytoplasm (Figure 29 F). Both cytoplasmic and nuclear distributions were much shorter than the poly(A) tail length of intronic reads detected in the nucleus, which had a median length of 250 nt, which was longer than intronic reads in HeLa S3 nuclear fractions and resembled the length observed for organoid FLAM-Seq data (Figure 16 C). Intronic reads detected in the cytoplasmic fraction were similarly shorter with a length of around 100 nt, also comparable to those observed in HeLa S3 bulk FLAM-Seq samples (Figure 16 C). *In vivo* analysis of poly(A) tail length in subcellular mouse brain fractions validated the results from HeLa S3 cell lines showing that poly(A) tails were drastically shortener already in the nucleus compared to poly(A) tails of intronic reads which is assumed to reflect tail length at the point of synthesis.

Investigating median poly(A) tail length for different classes of genes showed that lncRNAs had particularly long poly(A) tails in nuclear fractions, both for HeLa S3 (Fig 23 A) and mouse brain fractions (Fig 23 B). Bulk poly(A) distributions were not dominated by few highly expressed nuclear lncRNAs, since the observed nuclear poly(A) tail length per gene was on average longer than 200 nt for HeLa S3 cell lines and in also mouse brain fractionation experiments, which hinted at overall different deadenylation patterns for lncRNAs.

Median poly(A) tail length per gene was compared between individual subcellular fractions: comparing merged cytoplasmic and nucleoplasmic fractions showed that most genes had longer tails in the nucleoplasm and a higher dynamic range (Figure 30 A). Comparing poly(A) tails in chromatin and cytoplasmic fractions showed similar trends with most genes having longer poly(A) tails in the chromatin fraction. This was expected given that chromatin fractions were overall slightly longer than nucleoplasmic fractions. Comparison of chromatin and nucleoplasmic poly(A) tail length showed that nuclear compartments had a more linear

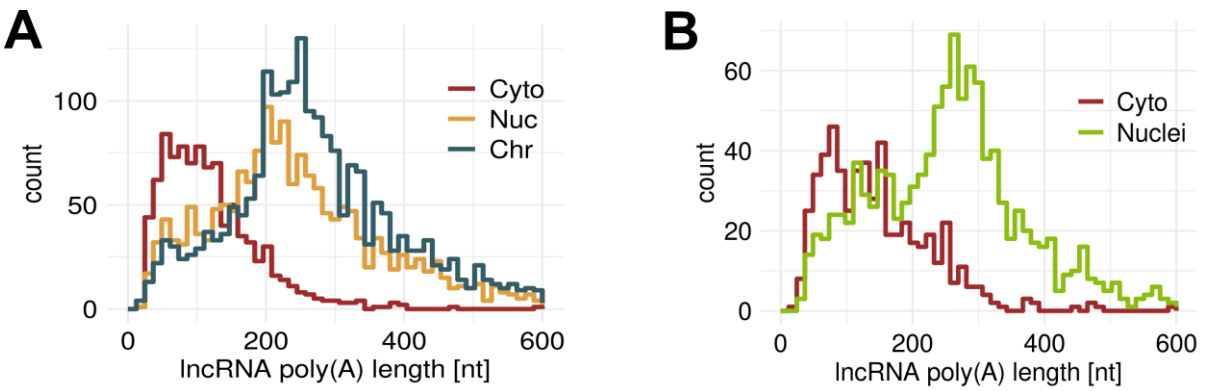

Figure 29 **Long non-coding RNAs have long poly(A) tails in the nucleus**
A) Poly(A) tail length distributions of lncRNAs in the HeLa S3 subcellular fractions. **B)** Poly(A) tail length distributions for lncRNAs in mouse brain nuclear and cytoplasmic fractions.

relationship, with an offset towards slightly longer chromatin poly(A) tails. Similar trends were found for mouse brain cytoplasmic and nuclear fractions with longer poly(A) tails per gene in in the nucleus. Investigating different gene sets such as ribosomal protein genes, lncRNAs and immediate early genes (IEGs) hinted at gene set specific poly(A) profiles across fractions: Ribosomal protein genes had shorter poly(A) tails in all fractions, while IEGs and in particular lncRNAs had longer poly(A) tails in nuclear fractions (Figure 30 B).

Other studies identified many lncRNAs to be highly enriched and retained in the nucleus [382] while this study found lncRNAs to have mostly long poly(A) tails in the nucleus. As a next step the general relationship between poly(A) tail length and transcript enrichment between nucleus and cytoplasm was investigated, under the hypothesis that poly(A) tail length is indicative of export or enrichment in the nucleus. Cytoplasmic-to-nuclear ratios for each gene were calculated from ENCODE data for HeLa subcellular fractions [245] and those were compared to cytoplasmic-to-nuclear ratios computed from FLAM-Seq HeLa fractionation data which showed decent agreement (r = 0.5). Since many genes were represented with only few counts in FLAM-Seq samples, median poly(A) tail length per gene for each fraction was compared to cytoplasmic-to-nuclear ratios inferred from ENCODE datasets. Poly(A) tail length was more correlated to cytoplasmic-to-nuclear ratios in nuclear fractions than in cytoplasmic fractions and genes with long poly(A) tails tended to be more enriched in the nucleus than in cytoplasm (Figure 30 D). LncRNAs were found to be most enriched in nuclear fractions, while IEGs had intermediate localization between cytoplasm and nucleoplasm and poly(A) tail length profiles. Transcripts of ribosomal protein genes were found to be strongly enriched in the cytoplasm.

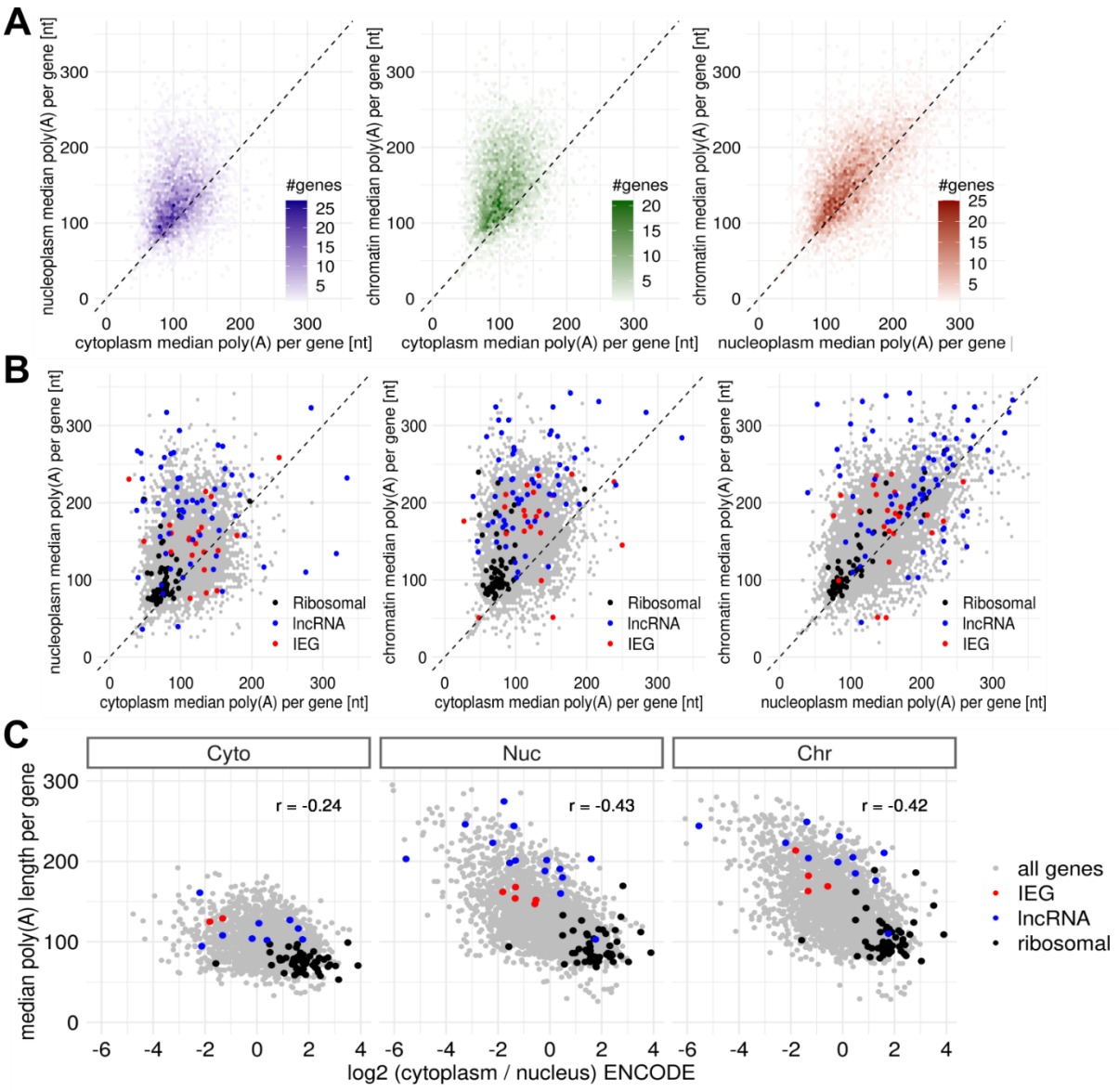

Figure 30 **Gene-specific features of poly(A) tail profiles in the nucleus**
A) Median poly(A) tail length per gene compared between subcellular HeLa S3 fraction for genes with > 5 counts. **B)** Gene sets of ribosomal protein genes, lncRNAs and immediate early genes (IEGs) were plotted on top of median poly(A) tail length per gene between subcellular compartments in HeLa S3 fractions. **C)** Cytoplasmic-to-nuclear ratios and median poly(A) tail length with IEGs, lncRNAs and ribosomal genes highlighted.

4.2.6 Perturbation of deadenylase enzyme complexes in subcellular fractions

A number of deadenylase enzyme complexes have been identified with different roles in mRNA deadenylation [254]. The CCR4-NOT complex was identified as being responsible for the complete removal of poly(A) tails [245], while the PAN2-PAN3 complex has been proposed to act upstream in trimming of longer poly(A) tails [250], possibly in the nucleus [214]. To investigate which enzymes are involved nuclear deadenylation in mammals different experimental strategies for RNA knockdown were applied in combination with subcellular fractionation and FLAM-Seq. CNOT7 and CNOT8, which encode the Caf1a and Caf1b subunits that were shown to be involved in removal of poly(A) tail regions less bound by Pab1 in yeast [247] and were targeted for knockdown. Further targets were PAN2 and PAN3 as well as the deadenylase PARN, which is involved for instance in nuclear telomere biogenesis [257].

Different strategies were evaluated for perturbing expression of deadenylase complexes on RNA level including RNA CRISPR-Cas bases systems as Cas13b and CasRx, siRNAs and stable, inducible shRNA expressing cell lines. RNA Knockdown efficiencies were each quantified by Nanostring or qPCR measurements. Western Blot validation of knockdown efficiencies on a protein level could not be performed through the lack of reliable antibodies against CNOT7/8 and PAN2/3.

Cas13b was identified as an RNA guided programmable RNase for efficient transcript cleavage [383,384]. Flp-In T-rex 293 cells were transfected with plasmids expressing guide RNAs and Cas13b for 24 h. Each two guide RNAs against CNOT7, CNOT8, PAN2 and PAN3 were transfected for 24 h before RNA was extracted and analyzed by Nanostring, a multiplexed assay for RNA quantification as well as qPCR measurements [385]. Cas13b transfections had opposing effects from the expected downregulation of RNA: genes targeted by Cas13b were upregulated compared to RNA counts in control samples or samples in which a different gene was targeted. CNOT7 Nanostring counts increased from around 500 in control and non-targeted samples to 690 when targeted by Cas13b (Figure 31 A). CNOT8 counts respectively increased from around 350 to 1450 in when targeted by Cas13b, for PAN2 counts increased from around 200 to 500 counts, for PAN3 the effect was milder with an increase from 200 to 300 counts when targeted by Cas13b. Genes not targeted by a Cas13 guide RNA remained unchanged compared to Non-targeting control expression.

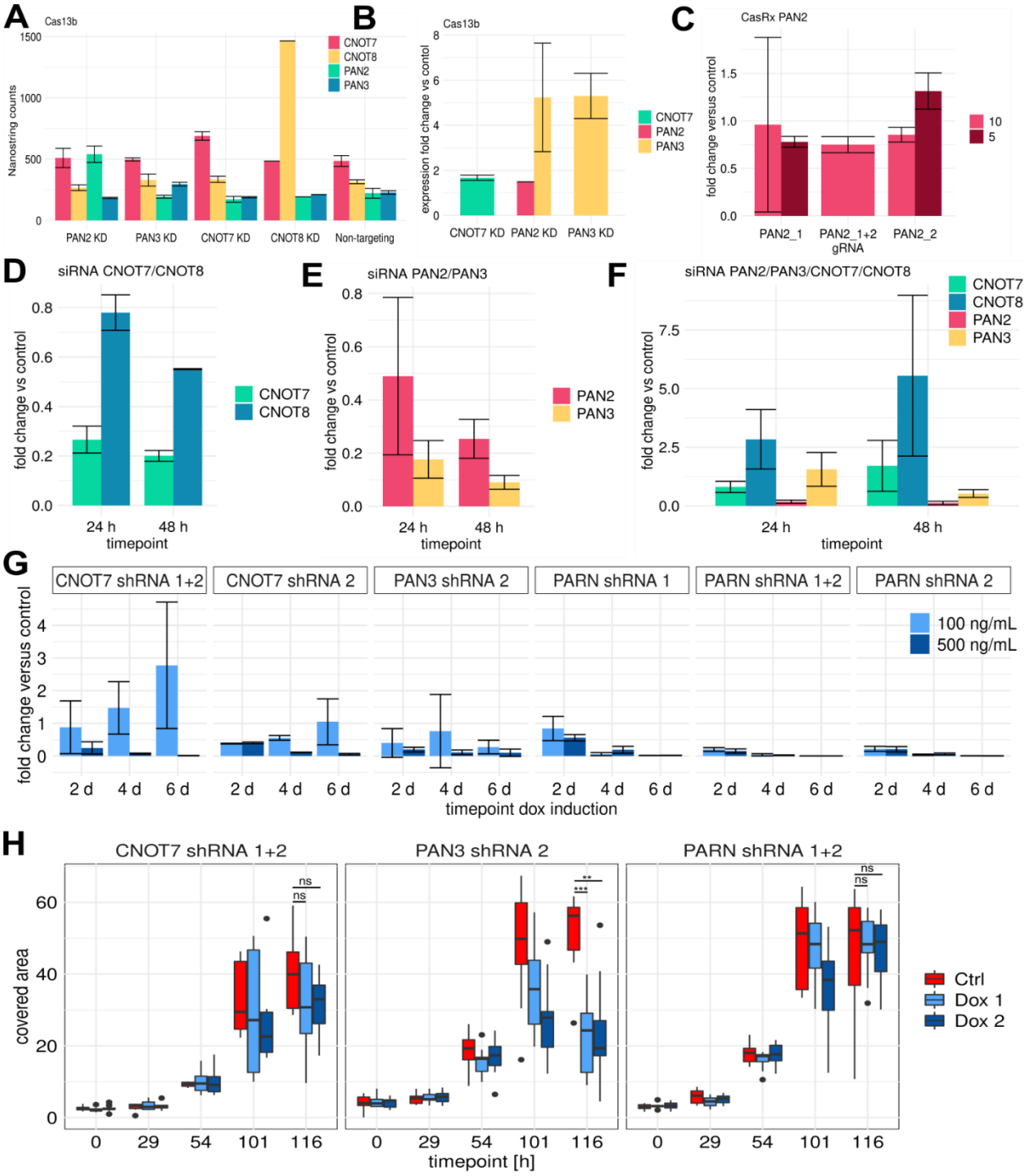

Figure 31 **Validation of knockdown strategies for deadenylase enzymes and phenotyping**
A) Nanostring counts for Cas13b knockdowns. Guide RNA target genes are shown on the x-axis. Nanostring assayed genes are shown in the legend for each sample. Error bars denote standard deviation for Nanostring counts. **B)** qPCR fold changes against control for Cas13b knockdown for different guide RNA samples. Target genes for qPCR quantification are shown on the legend. Error bars denote standard deviation. **C)** qPCR fold changes against control of different guide RNAs against PAN2 using CasRx system. 5/10 on legend refer to transfected volume of virus in µL. **D)** qPCR fold change versus for 24 h and 48 h siRNA knockdowns using siRNAs against CNOT7 and CNOT8. qPCR targets are shown on legend. **E)** As before for siRNAs targeting PAN2/PAN3 **F)** As before for siRNAs targeting CNOT7,CNOT8,PAN2 and PAN3. **G)** qPCR fold change against (non-induced) control for samples from cell lines with doxycycline (dox) inducible expression of shRNAs from 2 days to 6 days using 100 ng/mL and 500 ng/mL dox. Error bars denote standard deviation. **H)** Cell covered area (%) from imaging of cell growth curves for shRNA knockdown of CNOT7, PAN3 and PARN shRNA induced cell lines. Each series comprises 9 images for cell density for each timepoint. Differences in covered area were compared by a two-sided Student's t-test between Ctrl and Dox induction series.

This unexpected upregulation was validated by qPCR (Figure 31 B), where CNOT7 was upregulated 1.7-fold upon CNOT7 targeting by Cas13b, PAN2 was similarly upregulated 1.5-fold, while PAN3 was upregulated around 5-fold both when targeting PAN3 with a guide RNA but also as a side effect when targeting PAN2 with a guide RNA. Another RNA targeting Cas-based system was tested with CasRx [350], for perturbing PAN2 expression using two guide RNAs alone or in combination and transducing different amounts of virus (Figure 31 C). PAN2 could not be downregulated consistently using CasRx PAN2 and fold changes fluctuated around 0.75 and 1.5.

Next, siRNAs were tested for knockdown of deadenylase enzymes, since efficient depletion had been reported before for PAN2-PAN3 and CNOT7-CNOT8 [245]. Double knockdown of CNOT7 and CNOT8 for up to 48 h led to reduction in CNOT7 mRNA expression levels to 20% of control siRNA transfection, while CNOT8 levels were reduced to 78% (Figure 31 D). Double knockdown of PAN2 and PAN3 led to reduction of PAN2 expression to 25% of control for PAN2 and 9% for PAN3 (Figure 31 E). Simultaneous transfection of four siRNAs against CNOT7, CNOT8, PAN2 and PAN3 again led to reduction on PAN2 and PAN3 levels to 8% and 16% of control, while CNOT7 and CNOT8 were upregulated between 1.7 to 5.5-fold (Figure 31 F).

As a third option, stable HeLa S3 cell lines expressing shRNAs against CNOT7 (two cell lines expressing one shRNA or a combination of two shRNAs), PAN3 (one cell line expressing one shRNA) and PARN (two cell lines expressing one shRNA and one cell line expressing two shRNAs) were engineered (Ivano Legnini, Max Delbruck Center), with shRNA expression under control of a doxycycline inducible promoter (Tet-On system). 100 ng/mL and 500 ng/mL doxycycline (dox) were tested for shRNA induction and knockdowns were monitored up to 6 days (Figure 31 G). For CNOT7, 500 ng/mL dox was required for sustained knockdown which dropped to around 5% for the cell line expressing CNOT7 shRNA2 and around 1% for CNOT7 shRNA 1+2 cell line. PAN3 expression could be reduced to around 10% for 500 ng/mL dox. PARN expression was reduced to 1% in the HeLa S3 cell lines expressing two shRNAs. 100 ng/mL dox generally showed less efficient induction of knockdowns compared to 500 ng/mL and in one case an upregulation of the CNOT7 target similar to the effects observed for Cas13-mediated knockdowns.

During optimization of knockdown experiments, a reduction in cell growth was noted, particularly when inducing shRNA expression in the PAN3 shRNA2 cell line. To quantify this effect, shRNA-inducible cell lines were seeded, and shRNA expression was induced using

doxycycline. Non-induced cell lines served as a control. Cell growth was monitored for up 5 days (116 hours) by taking series of microscopy images which were processed using a custom image analysis pipeline that quantified the cell-covered area in each image. The area was plotted over time for different CNOT7, PAN3 and PARN cell lines (Figure 31 H). After 5 days, statistically significant differences were found in covered cell area between control and dox induction series for PAN3 shRNA expression, while no significant differences were observed for CNOT7 and PARN cell lines. To further investigate whether dox induction was relevant predictor of growth rates, a simple exponential growth process was assumed, and a linear model predicting covered area from the variables timepoint, dox induction and the interaction of both was fitted. Only the timepoints from 0 h to 101 h were considered for the analysis, since cell growth reached a plateau for the 116 h timepoint, which could not be captured by an exponential model. Linear models for PARN and CNOT7 growth curves returned only timepoints as significant predictor for the covered area. For the PAN3 cell line, the covered area ('growth curve') had statistically significant associations with dox induction and the interaction between dox and the timepoint for predicting the covered area, which further supported the notion that PAN3 knockdown but not PARN or CNOT7 knockdowns, significantly decreased cell growth.

Engineered cell lines with shRNA expression under control of tetracycline-controlled transcriptional activation enabled long term induction and RNA knockdown for several days and culturing of large amounts of cells which were required for obtaining sufficient input RNA for FLAM-Seq library preparation after biochemical fractionations. Cell lines with inducible expression of shRNAs targeting PARN, CNOT7 and PAN3 were treated with doxycycline for 5 days and for the PAN3 cell line additionally for only 3 days. Cells were fractioned into chromatin, nucleoplasm and cytoplasm and FLAM-Seq libraries were prepared from RNA extracted from each fraction. Knockdown efficiencies were monitored by qPCR (Figure 32 A): CNOT7 was downregulated to 12% of control expression after 5 days of dox induction in cell lines expressing CNOT7 targeting shRNAs. PAN3 expression was not affected upon CNOT7 knockdown. In the PAN3 shRNA cell line, PAN3 expression was downregulated to 7% of control expression after 5 days dox induction while CNOT7 was mildly downregulated to 84% of baseline. The second PAN3 induction experiment, in which dox induction was performed for 3 days, showed PAN3 knockdown to around 3%, which showed a degree of variability for maximum knockdown efficiency compared to 5 days knockdown, but is in line with previous time course experiments showing that after 4 days PAN3 levels were maximally reduced (Figure 31 G). Surprisingly PAN2 was also downregulated to around 30% of control expression in this cell line, albeit not being targeted by the expressed PAN3 shRNA. This hinted at possible

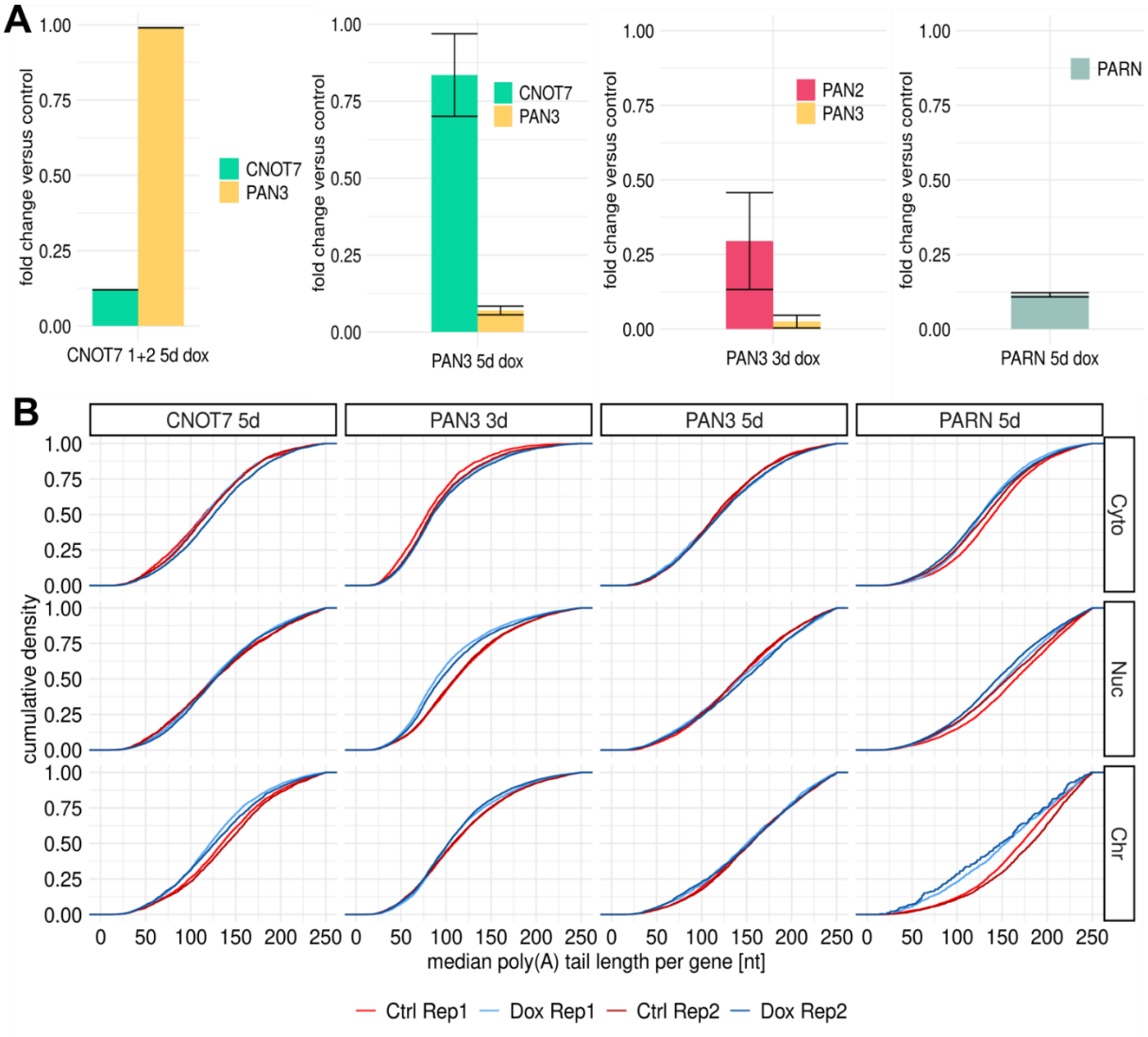

Figure 32 **Depletion of deadenylase enzymes impact poly(A) tail length in subcellular fractions**
A) qPCR fold changes of target genes compared to non-induced control cells for CNOT7 shRNA1+2 cell line with 5 days dox induction, PAN3 shRNA with 5d days dox induction, PAN3 shRNA with 3d days dox induction and PARN shRNA cell line with 5 days dox induciton. **B)** Cumulative poly(A) tail length distributions for replicates from control and dox induction and cytoplasmic (Cyto), nucleoplasmic (Nuc) and chromatin (Chr) biochemical fractions for shRNA cell lines. shRNA expression and dox induction time is shown in column labels.

feedback regulation upon PAN3 depletion. PARN expression could be reduced to around 12% of control expression.

Poly(A) tail length distributions per gene were compared between dox and control replicates for subcellular fractions for each shRNA cell line (Figure 32 B). Depletion of CNOT7 had only minor effects on cytoplasmic and nucleoplasmic poly(A) tail length, median differences in poly(A) tail length were 0.5 and 6 nt. Chromatin tails were slightly shorter upon dox induction with a median difference of 15 nt to control. 3 days PAN3 mRNA knockdown also led to shortening of poly(A) tails upon dox induction in nuclear fractions on average around 13 nt.

The observed effects were different after 5 days dox induction and PAN3 knockdown: Nucleoplasmic tails were here 12 nt longer than in non-induced control cell lines, also cytoplasmic and chromatin tails were slightly longer by 7 or 5 nt. Upon PARN knockdown, poly(A) tails were shorter in dox induced cell lines in all fractions. Here the observed differences ranging from 7 to 18 nt were strongest.

In summary it was found that perturbations of deadenylase enzyme complexes CCR4-NOT, PAN2-PAN3 and PARN was technically challenging and might have resulted in a number of unexpected effects: RNA-targeting Cas approaches led upregulation of many targeted genes for the Cas13b system. CasRx did not show the expected knockdown effects. siRNA transfections efficiently depleted target RNAs for up to 48 hours, yet double knockdowns of CCR4-NOT and PAN2-PAN3 components led to a stark increase in CNOT subunit expression, possibly through feedback regulation. Interestingly, the upregulation of CNOT7 mRNA levels was also observed in the cell line expressing an shRNA against CNOT7 under 100 ng/mL dox induction.

shRNA cell lines induced with concentrations of dox led to an efficient long-term knockdown and was hence applied in combination with biochemical fractionations. Results from FLAM-Seq analysis of biochemical fractionations showed in most cases a shortening of poly(A) tails upon knockdown of the deadenylase enzymes, while all observed effects were relatively small with a maximum observed difference in poly(A) tail length between dox induction and control of 18 nt in PARN chromatin fractions. In summary it was not possible to directly identify the enzyme(s) responsible for nuclear shortening of poly(A) tails. Results obtained from mRNA knockdowns using shRNA cell lines showed overall opposite (PARN, CNOT7) or inconsistent (PAN3 3 days, PAN3 5 days) changes in nuclear poly(A) tail length than those that would be expected for an enzyme responsible for shortening poly(A) tails in the nucleus.

5 Discussion

Poly(A) tails are essential elements for regulating gene expression and impact most steps of RNA metabolism such as splicing, export, RNA decay and translation. Decades of research have contributed to a detailed mechanistic understanding of how poly(A) tails are synthesized after transcription, how poly(A) tails impact translation efficiencies and how poly(A) tail deadenylation rates affect RNA decay.

Poly(A) tails were early on recognized as universal mRNA features and labeling experiments showed that tails become shorter over time. Studies in yeast mutants were indispensable for revealing poly(A) tail function implicated in transient processes such as RNA export and revealed different nuclear decay pathways which involve poly(A) tail hyperadenylation. Recent systems biology approaches systematically modeled the relationship between RNA deadenylation and decay uncovering deadenylation rates as highly predictive of RNA half-life [265].

Poly(A) tail length has experimentally been quantified either for individual genes or by extraction of bulk polyadenylate of the whole RNA population. To generalize observations from individual genes, methods are required which enable genome-wide quantification of poly(A) tails to gain a more comprehensive systems-biology perspective on polyadenylation.

5.1 High-throughput sequencing of full-length mRNA molecules

With the advent of microarrays and high-throughput sequencing, several methods were developed to quantify poly(A) tail length on a genome-wide scale [157,227,299]. More recently Nanopore direct-RNA sequencing was used for quantification of tail length [340]. The available approaches had yet important limitations which motivated the development of a novel method for high-throughput sequencing of complete mRNAs including their poly(A) tails.

Microarrays were used in combination with poly(A) tail length dependent elution from poly(U)-columns to profile yeast poly(A) tails [299]. The precision of poly(A) tail length quantification is thereby limited by the number of eluted fractions analyzed, which enabled binning of transcripts into ‘long’ or ‘short’ bins but was unable to directly quantify tail length. Short read Illumina sequencing enabled more quantitative poly(A) length analysis: The PAL-Seq method [157] quantifies poly(A) tail length by incorporation of biotinylated thymidines during cluster generation on an Illumina flow cell, such that the net biotin residues per RNA molecule are proportional to the poly(A) tail length. The biotin signal for each cluster is then detected by the

Illumina sequencer optics after addition of fluorescently labeled streptavidin. The fluorescent signal can then be compared to signals from a standard curved of known length to infer the poly(A) length from the fluorescent signal. The approach requires manipulations of the Illumina cluster generation steps, which are difficult to perform even for experienced users, and does not directly provide the poly(A) tail sequence.

The actual tail sequence is highly relevant for controlling transcripts since many studies showed that modifications of the poly(A) tail by guanines or uridines can stabilize or mark RNAs for decay [371]. The TAIL-Seq method directly sequences poly(A) tails and provides single nucleotide resolution for the last nucleotides of a given tail. For many tails non-A modifications were found at terminal tail positions [227], although the quantification was biased since internal poly(A) positions could not be measured by this protocol. TAIL-Seq directly applies paired-end Illumina sequencing to interrogate the poly(A) sequence from one end and the transcript body from the other end. Since the sequencing quality drastically diminishes when sequencing longer homopolymer sequences such as poly(A) tails, a specialized computational pipeline is used to extract the poly(A) tail part from reads. The read length limits the maximum detectable tail length to around 230 nt. TAIL-Seq also requires one specific model version of Illumina sequencers, which limits applicability.

A novel sequencing method was hence required which is simple, applicable for users with little experience in next generation sequencing and provides poly(A) tail length and sequence. Many genes produce alternative RNA isoforms with great differences in stability [386] and localization [387]. Since poly(A) tails are important indictors of RNA stability, it was important that poly(A) tails could be measured in the context of individual RNA isoforms. As such the method was required to deliver full-length transcripts which greatly facilitates isoform-level analysis.

The developed FLAM-Seq method utilizes PacBio sequencing [321] to sequence complete polyadenylated RNAs and is based on a simple laboratory protocol for generating sequencing libraries. The protocol is further compatible with Nanopore sequencing. The currently high error rates of Nanopore sequencing yet complicated the analysis of resulting datasets, in particular the quantification of poly(A) tails, but sequencing quality is expected to improve with future generations of the Nanopore sequencers. Other experimental methods for improving Nanopore sequencing quality, such as rolling circle amplification [388], could also be used together with FLAM-Seq.

The FLAM-Seq protocol begins with poly(A) selection from RNA to deplete abundant non-coding RNAs such as ribosomal RNA. rRNA depletion can also be performed by other methods, for instance RNAseH directed rRNA cleavage [304]. rRNA removal has been shown to be compatible with GI-tailing and FLAM-Seq library preparation albeit being less efficient and resulting in fewer usable reads (Ivano Legnini, Max Delbruck Center, data available at protocol exchange 10.21203/rs.2.10045/v1). Chang et al. omitted the poly(A) selection step for the TAIL-Seq protocol as it could select for longer tails, which more efficiently bind to oligo-dT beads. Indeed, shorter tails appeared more enriched when using FLAM-Seq with rRNA depletion instead of poly(A) selection.

GI-tailing was then used to introduce a universal priming site for reverse transcription that allows for introduction of a PCR handle in order to retain the complete poly(A) tail in the cDNA. The PAIso-Seq method [283], which also used PacBio to measure poly(A) tails, annealed for this step an adapter to the RNA poly(A) tail which was extended in 3'-direction to add a PCR handle by Klenow polymerase.

TAIL-Seq and PAL-Seq further fragment RNA which is not performed in the FLAM-Seq protocol. FLAM-Seq is yet to some degree limited in transcript coverage by the efficiency of reverse transcription and PCR amplification of long transcripts. A template switch step was included in the protocol which should enrich for full-length cDNAs [359], but we noticed that FLAM-Seq coverage typically dropped after 2 kb from the transcript end. IsoSeq, which is the PacBio RNA-Seq protocol [328], produced longer reads with less drop in coverage by including additional size selection steps to better represent long transcripts. Size selection is also possible for FLAM-Seq and may help to recover more full-length cDNA amplicons.

The first FLAM-Seq protocol contained a large number of concatemer sequences related to the template switch oligo (TSO) used for reverse transcription, which could be efficiently alleviated by chemically modifying the 5'-end using non-natural nucleotides [360]. Despite those improvements, concatenated sequences were in some cases detected in libraries produced from lower input material. This was surprising given that template switch reverse transcription can be performed from ultra-low RNA quantities (ca. 10 pg) without producing detectable library artefacts [389]. The fact that FLAM-Seq reverse transcription is primed on the GI-tail may yet compromise reverse transcription, since the efficiency of the GI-tailing step is not clear. Some reads also contained several poly(A) stretches or their own reverse complement, which could be a consequence of (adapter) ligation reactions. On average 40% of all sequenced reads contained a detectable poly(A) tail, showing that this step is important for further experimental

optimization of the method. FLAM-Seq quantified poly(A) tails for thousands of genes, which suggests unbiased sampling of transcripts comparable to regular RNA-Seq applications. The number of molecules sequenced for each gene was yet much lower than for PAL-Seq or TAIL-Seq methods. FLAM-Seq typically produced 100.000 – 500.000 reads per sample, while PAL-Seq and TAIL-Seq produced more than 50 million reads. Sequencing depth is currently limited by the PacBio sequencer output, which is also expected to increase in the future.

FLAM-Seq read starts were mapped to annotated human and *C. elegans* transcription start sites to estimate how many reads covered full-length transcripts, which was around 50% of reads. *In silico* control trimming of reads led to the expected reduction in coverage. SAGE annotations [352] used for *C. elegans* transcription start sites did not show the expected drop when trimming reads. This could be due to the fact that a large number of SAGE peaks was found in annotations without metrics to assess their relevance, i.e. filtering noisy start sites. Many *C. elegans* transcripts also undergo trans-splicing, where a splice leader sequence is spliced to the 5'-transcript end. Splice leader sequences were yet only detected for 5% of all reads (analysis by Ivano Legnini, Max Delbruck Center), which is much less than the reported 70% of genes [367] and could be related to coverage at the 5'-ends.

The ability of the FLAM-Seq approach to reliably quantify poly(A) tail length was validated by sequencing cDNA and RNA standards with known poly(A) length, which showed overall accurate quantification of tails. A slight shift towards shorter than expected poly(A) tails was observed, which could have been caused by PCR bias in preferentially amplifying shorter sequences or may reflect actual differences in the length of chemically synthesized oligos used as poly(A) standards. An RNA standard was used to assess potential errors possibly introduced by the enzymatic steps of the FLAM-Seq protocol such as reverse transcription. For most reads of the RNA standard, tails had the expected length of 50 nt, but the overall standard deviation was much higher than for the cDNA standards. The reason for this were a number of reads with very long tails that could be the product of multiple splint ligations of oligo(A) to the RNA 'body' of the standard (Legnini et al. 2019 [358]), resulting in oligo(A) concatemers.

Two complementary algorithms were combined to quantify poly(A) tail length from FLAM-Seq datasets, one based on seed extension and the second on a sliding window approach which both produced similar results and were in good agreement with manual annotation of poly(A) tail length. Other methods have been applied for identification poly(A) sequences from sequencing reads, for instance the Hidden Markov models which were used in the TAIL-Seq computational pipeline and trained on RNA standards [227], thereby providing a statistical model

for poly(A) quantification. Poly(A) tail profiles measured by FLAM-Seq were further compared to poly(A) profiles obtained from PAT assays which is an orthogonal electrophoresis-based method for quantifying tails of individual genes. Most genes measured by PAT assays showed good agreement with FLAM-Seq profiles, except for the BTF3 gene. For BTF3, poly(A) tails measured by PAT assay were longer, which was not directly captured by FLAM-Seq and could indicate slight biases in amplifying long poly(A) tails by FLAM-Seq.

5.2 Reproducible profiling of poly(A) tails in different model systems

FLAM-Seq was applied to quantify poly(A) tails for HeLa S3 and iPS cell lines, brain organoids and *C. elegans* L4 stage and adult worms. Poly(A) tail length distributions were characteristic for each biological sample and varied depending on the model system. The longest poly(A) tails were found in brain organoid samples, while *C. elegans* adult stage poly(A) tails were shortest. The comparison between iPS and organoids as well as *C. elegans* L4 und adult showed changes in poly(A) tail length across different developmental timepoints. Technical replicates further enabled assessment of technical noise in measuring tails for different biological samples. For all samples, a clear trend was observed that more counts per genes increased the correlation in median poly(A) tail length per gene between replicates. This confirmed reliable quantification of poly(A) tails for biological samples but also illustrated the uncertainty for poly(A) estimates of lowly expressed genes. Since many genes have broad poly(A) tail length distributions relatively large standard deviations were expected when comparing medians between replicates.

HeLa poly(A) tail distributions were compared between FLAM-Seq, TAIL-Seq and PAL-Seq methods. Reproducibility between methods was generally rather low, even when restricting the analysis to highly expressed genes. Despite biases of different methods in quantifying tails, transcriptomic differences between HeLa lines from different labs may also impact this comparison [390]. FLAM-Seq had the highest dynamic range in quantifying poly(A) tails and produced the overall longest poly(A) tail profiles compared to TAIL-Seq and PAL-Seq.

Poly(A) tail length was for all samples negatively correlated with gene expression, which has been reported by other high-throughput poly(A) tail sequencing studies [157,158,227]. For HeLa cell lines a negative correlation was also observed between median tail length and RNA stability per gene and translational efficiency. Those features are likely connected as highly expressed genes tend to be more stable and produce more protein. Earlier mechanistic studies in yeast showed that poly(A) tail length impacts translation initiation rates [220], yet this association was not found under steady state conditions. The exception here is early development [157] where

poly(A) tail length is correlated to translation rates in different model systems. The expression of poly(A) binding proteins (PABP) has in this context been proposed in mediating poly(A) tail length dependent translation rates [228], which is explaining the coupling observed in developing systems, where less poly(A) binding protein is available and long tails compete more efficiently for PABPs. Whether the coupling between translation rates and poly(A) tail length can be found iPS cells or organoids would be an important follow-up question, which would require determination of translation rates, for instance by ribosome profiling [391].

Gene ontology (GO) term analysis uncovered different housekeeping functions such as 'translation initiation' for which genes with short tails were enriched. On the other hand, certain regulatory cell functions such as 'cell development' were associated with longer tails. The GO term analysis yet needs to be interpreted with caution as formulation of a null hypothesis between genes of interest and background controls is difficult. Terms such as 'blood vessel morphogenesis' were enriched for genes with long poly(A) tails, but their relevance for a HeLa cancer cell line is for instance unclear, which makes distinctions between actually relevant and unrelated GO terms often difficult.

One important application of FLAM-Seq are comparisons of poly(A) tail length distributions between experimental conditions and different samples to identify genes for which poly(A) tails change. To enable this type of comparison, development of statistical models was required that can incorporate errors in measuring poly(A) tails, such that observed differences in tail length can be interpreted with respect to the uncertainty in tail quantification. Standard deviations of poly(A) tail length for highly expressed genes were slightly lower than for lowly expressed genes and increased proportional to tail length, which was also observed for synthetic cDNA. Variation in poly(A) tail length quantification between different genes had both technical and biological components: Technical variation was related to sequencing depth and could for instance be accounted for by shrinkage of variance, as performed in certain differential gene expression models as DESeq2[317]. Yet, comparing standard deviations in median poly(A) length in HeLa S3 cell lines with synthetic cDNA spike-ins showed that variation increased more than expected for longer poly(A) tails, which hinted at a biological component contributing to the spread of poly(A) tail length distributions.

Different statistical models were designed to identify significant differences in poly(A) tail length distributions between genes, also in the absence of replicates, since the low counts obtained from FLAM-Seq required in many cases merging of replicate datasets. To understand limitations in power for comparing poly(A) profiles, poly(A) tail length distributions were

simulated and tested for significant differences using three different statistical tests. A resampling based test was considered most sensitive and differences in poly(A) tail length could be reliably resolved for genes with more than 10 counts and differences in poly(A) tail length of around 25-30 nt. It is unlikely that a poly(A) tail length difference of few nucleotides would greatly impact mRNA behavior between experimental conditions, such that sensitivity on this level may not be required. Poly(A) length differences of 25-30 nt could on the other hand have important biological consequences, as for instance one poly(A) binding protein occupies a footprint on the poly(A) tail of comparable size [392]. Other statistical models which model the variability in median poly(A) tail length across replicates could be an important alternative in particular if PacBio sequencing depth increases in the future.

264 genes were identified with longer poly(A) tails in organoids compared to iPS cells, which hinted at specific regulation of poly(A) tail length between different developmental timepoints. In particular genes specifically expressed in organoids had longer poly(A) tails. Post-transcriptional gene regulation in neuronal systems was shown to involve extensively long 3'-UTRs [100,101], which could also be shown by FLAM-Seq as organoid genes had long 3'-UTR profiles.

The first aim of this thesis was development of a novel method for genome-wide analysis of poly(A) tail length in context of complete mRNA molecules, including computational models for data analysis. FLAM-Seq enabled analysis of poly(A) tail length for thousands of genes in independent HeLa S3 and iPS cells, brain organoids and *C. elegans* adult and L4 stage samples. Quantification of poly(A) tail length was validated by manual tail annotation, using synthetic standards and comparisons between technical replicates. Various statistical models for comparing differences in poly(A) tail length between experimental conditions were optimized, uncovering many genes with significant differences in poly(A) tail length between iPS and organoids. The only current disadvantage of FLAM-Seq compared to other methods is the low sequencing depth of PacBio sequencing. This makes comparisons on a single-gene level challenging, in particular for lowly expressed genes, but future generations of PacBio or Nanopore sequencers are expected to increase sequencing depth.

5.3 Regulation of polyadenylation, poly(A) tail length and nucleotide content

One advantage of the FLAM-Seq method is the possibility investigate poly(A) tails in context of other gene regulatory features such as individual 3'-UTR isoforms, facilitated by long reads. Although this type of analysis was in some cases possible using previous methods [157], FLAM-Seq is the preferred method. FLAM-Seq reads cover the exact cleavage site at the 3'-UTR end

and as such does not require a precompiled annotation database such as RefSeq or Gencode databases for identification of 3'-UTR isoforms. This is an important advantage since 3'-UTR annotations were in many cases imprecise and did not match with the 3'-UTR ends identified by FLAM-Seq. This can impact gene quantification using RNA-Seq [393] and in particular single-cell RNA-Seq applications, which are biased towards sequencing of 3'-UTR ends and could greatly benefit from more accurate 3'-UTR annotations provided by FLAM-Seq.

FLAM-Seq also alleviates the problem of internal priming for identification of transcript ends. Internal priming can occur when oligo-dT primers are used for reverse transcription that prime cDNA synthesis from A-rich transcript regions, thereby creating 'false' transcript ends [362,394]. As FLAM-Seq provides the poly(A) sequence, reads can be directly compared to the genomic sequence to decide whether the poly(A) part of a read resulted from an A-rich transcript region or non-templated poly(A) tail.

Inspecting FLAM-Seq alignments in a genome browser revealed than many read ends did not align at a single coordinate for a given 3'-UTR end, but within windows of several nucleotides. This effect was quantified and found for several hundred 3'-UTRs. This 'microheterogeneity' has been found by other studies in different model systems [94,395,396], and could be related to the exact sequence context which may specify the exact positioning of the cleavage and polyadenylation machinery binding to the nascent transcript.

Besides the canonical AAUAAA polyadenylation signal, other variants have been described [61]. The frequencies of polyadenylation sites identified close to the 3'-ends in FLAM-Seq samples matched previous reports [94], for instance AUUAAA was identified as second most prevalent polyadenylation signal in human datasets and around 10% of mRNAs were found without any polyadenylation signal. The frequencies in polyadenylation site usage differed between human and *C. elegans*, which could be related to differences in the molecular architecture of the cleavage and polyadenylation complex or auxiliary factors. The rankings of *C. elegans* PAS site usage was also reproduced from previous studies [396], which illustrated the accuracy of the analysis. In this context, the 3'-UTR length distribution of *C. elegans* samples matched with a median of around 150 nt the length profiles identified by other studies annotating 3'-UTRs [394,396]. Annotated human 3'-UTRs have a median length of ca. 1000 bp [71], yet the 3'-UTRs annotated from FLAM-Seq had an average length of around 550 nt. This difference could have two reasons: first, highly expressed genes have shorter 3'-UTRs [92], which are likely more represented in our analysis through the shallow sequencing depth of FLAM-Seq. Alternatively,

our computational 3'-UTR annotation pipeline could be biased towards shorter 3'-UTRs as it selects for the closest coding sequence end, which may not reflect the true end in each case.

3'-UTR length differences for the same gene between iPS cells and brain organoids showed that most 3'-UTRs were extended in organoids, which is in line with previous studies reporting extensively long 3'-UTR in brain tissue [100]. Although alternative 3'-UTR isoforms were reported for more than a third of all *C. elegans* genes [397] and isoform switches were extensively characterized during early development [398], only a handful of 3'-UTR switches were found between L4 and adult stages.

In all sequenced samples, 3'-UTR length was coupled to poly(A) tail length, with longer 3'-UTRs having longer poly(A) tails. Longer 3'-UTRs are statistically more likely to contain AU-rich elements or miRNA binding sites, which would render them less stable. Longer poly(A) tails might hence be a consequence of differential stabilities of different 3'-UTR isoforms which has been observed comparing proximal and distal alternative 3'-UTRs [91].

Investigation of poly(A) tail length in context of alternative polyadenylation confirmed this trend as longer distal 3'-UTRs had longer poly(A) tails than proximal 3'-UTRs of the same gene. Similar cases of co-regulation have been observed at promoters, where for some genes alternative transcription start sites were associated with differences in poly(A) tail length. How differences in 3'-UTRs lengths are mechanistically linked to differences in poly(A) profiles could depend on different factors that affect poly(A) distributions for each isoform: one option is the connection between stability and deadenylation rates, as long 3'-UTRs are more likely to contain destabilizing motifs which can recruit *trans*-acting factors such as RBPs and deadenylase complexes that promote deadenylation. The observed differences could also be related to differences in production rates, but little is known about isoform specific differences in transcription.

Annotations of transcription start sites from FLAM-Seq data on the other hand was challenging since many FLAM-Seq reads did not cover the full transcript. In those cases, existing gene and transcription start site annotations had to be utilized for mapping FLAM-Seq reads. For more accurate transcript start annotations, it may be helpful to modify the FLAM-Seq protocol for enrichment of long transcripts as described earlier, for instance by size selection which may increase coverage at exact transcript starts.

Poly(A) tails contain other nucleotides besides adenosines, which were shown to impact stability of transcripts. Terminal guanosines ('mixed tailing') stabilize mRNAs by blocking the

action of deadenylases [271], while terminal uridine marks RNA for decapping and decay [277]. Analysis of non-A nucleotides was previously restricted to terminal poly(A) tail positions since TAIL-Seq was not able to faithfully resolve internal poly(A) sequences. FLAM-Seq enabled nucleotide-resolution for sequencing of poly(A) tails and was used to determine the global non-A nucleotide frequencies within poly(A) tails. Those were highest for cytosines, followed by uridines and guanosines. Guanosine frequencies were yet not enriched over controls and could have been artefacts introduced by reverse transcription or other processing steps. Occurrence of non-A nucleotides was not limited to terminal positions and slightly more increased within tails, although the three positions at the end of the poly(A) tail were also selected to be exclusively adenines through the design of the reverse transcription primers that enhance the protocol efficiency. Non-A nucleotides were found for most genes independent of the biological sample and poly(A) tail length. Cytosines had not been described in endogenous poly(A) tails before but the canonical poly(A) polymerase (PAP) has been demonstrated to incorporate cytosines *in vitro* [273] which indicated that non-A nucleotides could in part result from unspecific nucleotide incorporation during poly(A) synthesis. Uridines or guanosines are incorporated into poly(A) tails by TUT and TENT classes of non-canonical poly(A) polymerases. Whether a similar class of enzymes directs addition of cytosines is unclear, but some enzymes have been identified in adding CCA, which modify tRNAs but could also act on poly(A) tails [399]. How non-A nucleotides impact transcript processing has not been investigated in detail, but it was for instance shown that guanosines can alter the intrinsic structure of the poly(A) tail which inhibits the Pan2 enzyme [272].

5.4 Synthesis of long poly(A) tails in the nucleus

Detailed *in vitro* studies have uncovered a mechanism for poly(A) tail biogenesis of long poly(A) tails. Poly(A) synthesis involves binding of the nuclear poly(A) binding protein (PABPN1) to the elongating poly(A) tail which is synthesized by the nuclear poly(A) polymerase (PAP). The successive binding of multiple PABPN1 molecules leads to a steric collapse of the resulting complex at a poly(A) tail length of ca. 250 nt, which marks the poly(A) tail length in mammals at the point of synthesis [163]. Similar conclusions were derived from radioactive labeling of RNA in cell lines and comparison of the bulk polyadenylate fraction against standards of known length [149,165]. Both analyses yet were not able to define the poly(A) tail length of newly synthesized RNAs on a genome-wide scale. By extracting unspliced, polyadenylated reads from FLAM-Seq datasets it was possible to investigate poly(A) tails of newly synthesized RNAs for thousands of genes. Global synthesis of long poly(A) tails could be observed in iPS and organoid samples but also in HeLa S3 nuclei and *in vivo* in mouse brains.

In HeLa S3 bulk samples, many unspliced reads were detected with shorter poly(A) tails, but biochemical fractionation experiments showed that that those unspliced reads mostly occurred in the cytoplasm and may represent retained introns, which are frequently observed in cancer transcriptomes [400]. The fraction of unspliced reads was higher in iPS and organoid samples compared to HeLa cell lines, which could be related to faster RNA splicing kinetics in the cancer cell line. RNA processing rates have been correlated to cell growth in yeast [401], and similar relationships could also play a role in mammalian cell growth, that might explain why less unspliced reads are found in highly proliferative cancer cell lines.

The number of detected unspliced reads was overall low, as such it was important to investigate in how far the conclusions drawn from few intronic reads were representative of genome-wide RNA processing. Genes for which unspliced reads were identified were carefully investigated, but no evidence was found that only distinct gene sets were represented in the fraction of unspliced reads. Unspliced reads appeared to be randomly ('Poisson') sampled such that the resulting number of unspliced reads was mostly limited by sequencing depth. This also became apparent when inhibiting splicing, which led to a 2- to 3-fold increase in detected genes with unspliced reads. The second limitation was the annotation of introns which unambiguously define unspliced reads: the curated intron annotations did not cover all genes through conflicts between overlapping coding regions of different isoforms, such that an upper bound of 50% of genes was calculated for which unspliced reads could in principle be detected.

The computational approach for identifying unspliced reads was validated by splicing inhibition experiments in combination with FLAM-Seq RNA profiling of nuclear RNA, which greatly increased the fraction of unspliced reads. As expected, splicing inhibition led to accumulation of unspliced reads, which validated the computational approach. Interestingly, despite the inhibition of splicing for three hours, no evidence for unspliced reads with shortened poly(A) tails was found. Poly(A) tails of unspliced pre-mRNAs could in principle be shortened even if splicing is blocked. Yet, since all unspliced reads retained long poly(A) tails, this suggested a link between splicing and subsequent shortening of the poly(A) tail. Splicing could hence be a pre-requisite for deadenylation. Mammalian splicing leads to deposition of exon junction complexes (EJCs) close to splice sites, which are important for instance in triggering nonsense-mediated decay [117]. Nuclear deadenylases could be recruited through EJCs and start deadenylation of unspliced reads. Another option is that unspliced reads remain associated to the transcription site leading to compartmentalization that excludes unspliced reads.

Unspliced reads have longer poly(A) tails upon splicing inhibition, with a length difference of ca. 40 nt in HeLa S3 nuclei. This observation posed the question why tail length is increased when inhibiting splicing. The common model for length control of poly(A) tails proposes that tail length is limited by the nuclear poly(A) binding protein (PABPN1), which probably does not differentiate whether splicing is globally inhibited or not. PABPN1 yet does no prevent distributive poly(A) tail synthesis of the canonical poly(A) polymerase which could cause the observed increase in tail length of unspliced reads by continuous addition of adenosines with lower efficiency when inhibiting splicing. Many nuclear decay pathways, such as PAXT [182] or PPD [205] involve hyperadenylation of poly(A) tails by nuclear poly(A) polymerases such as TRF4 (PDPD5), which precedes decay of the transcripts by the nuclear exosome. The increase in poly(A) tail length of unspliced transcripts could also have resulted from adenylation by other nuclear poly(A) polymerases that could mark unspliced transcripts for nuclear decay. Many nuclear RNA decay targets, including mRNAs have been described which are targeted by the TRAMP complex in yeast [199] and the PAXT pathway [213]. Whether unspliced mRNAs are a distinct target class has not been explored in detail so far. The nuclear fate of unspliced transcripts and regulation through the poly(A) tail are an interesting subject for follow-up studies, since regulated splicing may provide means to actively control how mRNA is released into the cytoplasm [52,54].

Global poly(A) tail length decreased after inhibition of splicing, which was likely a consequence of inhibiting production of mature mRNAs. Interestingly, a second mode of longer poly(A) tails appeared after splicing inhibition, which could reflect the accumulated unspliced reads, including a number of reads, that may not have been annotated as ‘unspliced’ through the lack of intron coverage (‘false negatives’). For certain genes, for instance IER3 or MYC, splicing inhibition led to increased poly(A) tail length and transcript upregulation. Genes with longer poly(A) tails also had predominantly shorter half lives. Those findings yet need to be interpreted with caution, since RNAs with short half-lives are degraded faster, such that their steady state poly(A) tail length distributions will be more influenced by newly synthesized RNAs with long(er) tails and may not reflect direct regulation of poly(A) tail length.

Since FLAM-Seq involves a poly(A) selection step, unspliced reads are both polyadenylated and have a poly(A) tail. This configuration can only be found for mRNAs which are post-transcriptionally spliced. Post-transcriptional splicing has been shown to be less frequent that co-transcriptional splicing [38], but more recently long read sequencing studies found a large fraction of genes with terminal introns which are post-transcriptionally spliced, in particular in mammals [43]. Co- and post-transcriptional splicing may also co-exist since splicing order also

depends on intron position within a transcript and can be mechanistically linked to other maturation steps as polyadenylation [50].

To exclude that the presented findings were limited to post-transcriptionally spliced genes, previously published Nanopore long read sequencing datasets of nascent, chromatin associated total RNA were re-analyzed. Despite some technical limitations in identification of unspliced reads and quantification of poly(A) tail length caused by lower Nanopore sequencing quality, it could be shown that nascent mRNAs have either no poly(A) tails or long tails after cleavage at annotated polyadenylation sites. This observation could describe the kinetics of the polyadenylation reaction, involving a fast poly(A) synthesis step and a lag time between cleavage and poly(A) synthesis. The stringent RNA isolation used by Drexler and colleagues, which involved chromatin fractionation and 4sU labeling for few minutes, further suggested that polyadenylated RNAs remain for some time physically associated with chromatin.

Other direct RNA sequencing studies also found mostly long poly(A) tails for incompletely spliced reads identified from direct mRNA sequencing studies [340], although those analysis could not directly distinguish unspliced reads form retained introns.

5.5 Rapid deadenylation of poly(A) tails in the nucleus

Genome-wide analysis of unspliced RNAs in mammalian model systems showed that poly(A) tails of unspliced molecules were mostly longer than 200 nt in cell lines, organoids and brain samples. This was is in line with previous *in vitro* experiments that revealed a mechanism for synthesis of long poly(A) tails [163] and radioactive metabolic labeling [149,165]. The steady state poly(A) tail length distributions were shorter, with medians of 90 to 140 nt per gene, and comparable length profiles found by other studies [157,227]. The discrepancy between poly(A) tails during synthesis and steady-state tail length required for progressive deadenylation, which had been investigated by several studies for individual genes [246,250,265,402].

FLAM-Seq was used for resolving deadenylation after inhibiting transcription. Poly(A) tail distributions shifted towards shorter tail length for increasing time of transcription inhibition using actinomycin D. This ‘ageing’ of poly(A) tails upon inhibiting transcription has been observed already in the 1970s [149] and was validated by FLAM-Seq and other methods [265] on a genome-wide level. Shortening is explained by continuous deadenylation, but RNA decay must be taken into account as well to explain steady state poly(A) distributions. Mitochondrial poly(A) tail dynamics were different from those of nuclear encoded genes since no differences in poly(A) tail length profiles were observed after inhibiting transcription. Although dedicated

enzymes for deadenylation of mitochondrial transcripts have been identified, for instance PDE12 [152], the kinetics of this process in context of mitochondrial RNA decay are little understood, but the FLAM-Seq analysis suggest that progressive shortening of mitochondrial poly(A) tails as for mRNA seems unlikely.

A recent study investigated poly(A) tail length for different timepoints after synthesis by combining metabolic labeling with PAL-Seq for genome-wide quantification of poly(A) tails [265]. The authors concluded that deadenylation rates are highly indicative of mRNA turnover and identified deadenylation as the limiting factor for RNA decay, which had been proposed before for individual genes. The study by Eisen et al. was yet limited to the analysis of mRNA poly(A) tails that were already exported to the cytoplasm, which the authors defined as the poly(A) tail length after 45 min metabolic labeling. How well this time interval reflects the default timing for export yet remained an open question: cytoplasmic export has been described to occur in the range of 5-40 min for reporter systems in cell lines [376] or on average 15 min in *Drosophila* genome-wide studies [377], which illustrates a degree of uncertainty associated with timing for export. Nevertheless, Eisen et. al found poly(A) tails with a median length of 133 nt after 45 min labeling. This was already shorter than the tail length observed right after synthesis, which is 200 nt or more.

Metabolic labeling and pulldowns using 4sU were hence performed for earlier timepoints, ranging from 10 min to 90 min labeling, to investigate poly(A) tail length dynamics directly after synthesis of poly(A) tails. Surprisingly, even after 10 min, poly(A) tails were shortened, which implied a rapid deadenylation step after poly(A) tail synthesis. Poly(A) tails of unspliced reads were as expected around 200 nt in length and more abundant in labeled RNA fractions. The differences in poly(A) tail length profiles for longer labeling timepoints were small and could not be reliably compared, also through the lack of technical replicates which had to be merged for obtaining sufficient material for FLAM-Seq experiments. Similar results were obtained by combining SLAM-Seq, an orthogonal method for genome-wide identification of newly synthesized RNAs with FLAM-Seq for simultaneous quantification of poly(A) tail length. Labeled reads could be reliably identified from FLAM-Seq datasets using a statistical model which compared observed T-to-C conversions per read against a background error model. After labeling reads for 90 min or 180 min, poly(A) tails were as expected shorter than 200 nt, but longer than pre-existing poly(A) tails. The median difference between newly synthesized and pre-existing RNA was with 20 nt per gene comparable to the observed difference of 30 nt for 4sU pulldown experiments. SLAM-Seq experiments yet suffered from

differences in overall poly(A) tail distributions between individual samples, which was likely related to RNA degradation through harsh reaction conditions.

Poly(A) tails after 10 min 4sU labeling were markedly shorter than the poly(A) tails from re-analyzed Draxler et al. Nanopore direct RNA sequencing data of nascent, chromatin associated RNAs, where 4sU labeling is performed for 8 minutes. Reasons for this discrepancy could be that deadenylation of poly(A) tails requires dissociation of mRNA from the transcription site into the nucleoplasm, which could happen within the first 10 minutes after synthesis.

Time estimates for RNA export range from 5 to 40 min, which suggests that deadenylation observed after 10 minutes of labeling could be a nuclear event. To further investigate this hypothesis, HeLa S3 cell lines were biochemically separated into cytoplasmic, nucleoplasmic and chromatin fractions. Nuclear fractions were carefully investigated for cytoplasmic contamination by Western Blot and the abundance of mitochondrial transcripts. Despite the absence of cytoplasmic contamination, experimental variability in poly(A) tail length profiles for the same fraction from different experiments could not be avoided, which has been described before and is inherent to biochemical fractionation experiments [378]. To deal with this variation, a number of experimental replicates were produced in order to adequately take into account and model the underlying variability for fractionation experiments. In total 12 replicate samples were analyzed for each subcellular fraction. The largest source of variability was linked to global biases in median poly(A) tail length per gene between experiments, which each affected all fractions of an experiment. This bias could be modeled as a scaling factor when computed across all fractions of each experiment. The effect might be linked to differences in RNA input or biases when using different PCR cycles, which could impact global poly(A) tail quantification.

Poly(A) tails of mature nuclear RNAs were shorter than poly(A) tail distributions of unspliced reads, which were mostly found in the nucleus. This supported the hypothesis that poly(A) tails are rapidly shortened directly after synthesis in the nucleus on a genome-wide level. This analysis could be validated *in vivo* by extracting RNA from mouse brain nuclei (experiment by Maddalena Pacelli). Chromatin associated RNA had shorter poly(A) tails, which was not directly observed in the analysis of the Drexler et al. Nanopore datasets which combined chromatin fractionation with further 4sU pulldowns, revealing longer poly(A) tails of chromatin associated RNA. This difference could again be explained by the additional 4sU labeling step performed by Drexler et al. which could have enriched for more nascent RNA. The fraction of unspliced reads found between nucleoplasm and chromatin fractions was very similar, while a

higher fraction could be expected for the chromatin fraction. This could also hint at a mix of chromatin and nucleoplasmic RNA, which is not totally unexpected given the experimental separation procedure of chromatin and nucleoplasmic fractions, which relies on different solubility of chromatin-associated RNA.

The steady state poly(A) tail length varied for individual genes between subcellular fractions. The comparison between ribosomal protein genes, lncRNAs and immediate early genes (IEGs) showed that those gene classes had different poly(A) tail length distributions in the nucleus, with much shorter poly(A) tails of ribosomal housekeeping genes. Similar differences were seen before when comparing the poly(A) tail length for different 4sU labeling timepoints. It was yet difficult to discern whether the steady state poly(A) tail distributions were the consequence of differences in nuclear deadenylation rates, export, or decay rates. Nuclear decay and export rates are difficult to quantify on a genome-wide level and the lack of reliable experimental data made the distinction of individual contributions impossible. It was yet clear that nuclear deadenylation affected most genes, but exceptions were found for many lncRNAs, which had poly(A) tail length distributions of around 200 nt in the nucleus that resembled those of unspliced RNAs. This illustrated that nuclear deadenylation, similar to cytoplasmic deadenylation, is likely under active regulation and gene specific. lncRNAs are often actively retained in the nucleus [143], and distinct sequence elements have been identified being involved in retaining lncRNA transcripts in the nucleus [382]. Whether poly(A) tails are required for mRNA export is unclear. At least a number of histone RNAs are not polyadenylated, but efficiently exported, thereby involving different protein adapters which renders them special to the majority of RNAPII transcripts.

The comparison between poly(A) tail length and transcript enrichment in nuclear or cytoplasmic fractions uncovered a correlation suggesting that genes with shorter nuclear tails are more enriched in the cytoplasm, hence could be exported more efficiently. Nuclear deadenylation could in this regard play a role in enhancing or regulating RNA export or serve as a quality control step in RNA maturation required for export. Follow-up experiments investigating the role of nuclear deadenylation and export could involve measurement of transcript enrichment after successful perturbation or inhibition of nuclear deadenylation. A recent study that investigated deadenylation in the nucleus for a small number of genes came to similar conclusions [248], uncovering nuclear deadenylation for serum-induced genes.

Comparing poly(A) tail length per gene between different biochemical fractions showed that poly(A) tails are mostly shortened moving from chromatin and nucleoplasm to the cytoplasm

with few genes having longer tails in the cytoplasm. As described earlier, cytoplasmic polyadenylation is used as a mechanism to activate deadenylated mRNAs for translation throughout development [343] and can affect protein production for instance in neurons [403]. Comparing poly(A) tails in cytoplasmic and nuclear fractions yet did not show any evidence for global adenylation of tails in the cytoplasm. This does not preclude that cytoplasmic adenylation also operates outside of development impacting steady-state poly(A) tail length, but little is known about how those processes could be regulated.

5.6 Identification of enzymes responsible for nuclear deadenylation

Metabolic labeling experiments and biochemical fractionations suggest that poly(A) tails are already shortened before being exported to the cytoplasm. This requires for deadenylase enzymes to be active in deadenylation in the nucleus. The two major deadenylase enzymes have established nuclear functions since the Pan2-Pan3 complex has been linked to nuclear deadenylation in yeast [214] and CCR4-NOT has been shown to act as transcriptional activator or repressor [342]. The deadenylase PARN also has nuclear function, for instance in telomer RNA biogenesis [257]. The CCR4-NOT complex is a multi-subunit deadenylase, assembled of different subunits with a variety of functions, ranging from transcriptional activation [404] to translational repression [405]. The Ccr4 and Caf1 subunits are the deadenylase subunits, which correspond to CNOT7 (Caf1a)/CNOT8 (Caf1b) and CNOT6 (Ccr4a)/CNOT6L (Ccr4b) in human. In cell lines, CNOT7/CNOT8 knockdown has been shown before to cause an increase in poly(A) tail length [245]. Another study showed that the yeast Caf1 subunit mostly leads to trimming of specific poly(A) tails not bound by the poly(A) binding protein Pab1, whereas the Ccr4 subunit had more universal deadenylase activity that was stimulated by Pab1 [247]. The PAN2-PAN3 complex is composed of two PAN3 and one PAN2 subunit, where PAN2 harbors the deadenylase activity. PAN3 is essential in maintaining deadenylase activity since PAN3 depletion also leads to global poly(A) tail lengthening [406]. Perturbing the major eukaryotic deadenylation complexes was hence a first step in identification of enzymes involved in nuclear deadenylation.

PAN2, PAN3, CNOT7 and CNOT8 were initially targeted for mRNA knockdown by the Cas13b system [384] and the more compact and efficient CasRx system [350]. None of the programmable Cas-based RNA targeting methods could efficiently deplete the target RNAs. While PAN2 levels did not change for CasRx mediated knockdowns, Cas13b mediated knockdown led to higher expression levels of the targeted genes, which was also observed for the CNOT7 shRNA cell line, induced with a lower concentration of doxycycline. Upregulation

could be caused by indirect effects which sense transcript depletion and enhance gene expression through feedback loops. Since knockdowns were performed only for 24 h, investigation of longer time intervals could be useful in understanding this effect. Through the lack of specific antibodies, knockdown efficiencies could not be validated on the protein level. siRNA mediated knockdown led to an efficient reduction both in CNOT7/CNOT8 and PAN2/PAN3 expression. This has been reported before by Yi et al., who could also show that PAN2 and CNOT7 mRNA reduction to around 20% led to a complete disappearance of protein expression by Western Blot. Finally, stable shRNA inducible cell lines were used for CNOT7, PAN3 and PARN knockdowns over several days, since engineered HeLa cell lines allowed to obtain the large number of cells required for biochemical fractionation experiments. shRNA knockdowns were highly efficient for all targets.

Yi et al. reported a profound lengthening of poly(A) tails after CNOT7/CNOT8 co-depletion. For PAN2/PAN3 knockdowns the effect was only detectable for very long poly(A) tails and no difference was observed upon PARN depletion. Comparing the differences in poly(A) tail length distributions for HeLa S3 cell lines in different subcellular fractions showed little effects of shRNA-mediated CNOT7, PAN3 or PARN knockdowns, which were performed for 3 or 5 days. Different biological and technical reasons could have influenced the weak molecular phenotypes: Knockdowns were not validated on protein level, hence protein expression could have been less reduced compared to measured RNA levels, which was measured to validate shRNA efficacy. This seemed yet unlikely since efficient mRNA knockdowns were already observed after 2 days, and only few proteins have half-lives of more than 4-5 days [118], so that a general reduction in protein levels was expected. Another option is that compensatory mechanisms could have overwritten the molecular effects of shRNA knockdowns. Perturbing highly conserved and essential cellular mechanisms such as deadenylation is unlikely to be tolerated by cells over extended periods of time without impacting their viability. Since several deadenylase enzymes exist this may also provide buffer capacities and compensate for mutual losses in activity, which could have been observed in this study. Parallel siRNA knockdowns of PAN2/PAN3 and CNOT7/CNOT8 also led to upregulation of CNOT7/CNOT8, although knockdowns of CNOT7/CNOT8 alone were efficient, which could hint at induction of compensatory mechanisms.

One approach for dealing with compensatory actions could be a rapid depletion of deadenylases directly at the protein level. This could for instance be achieved through auxin inducible degron technology [407]. For this method a short degron tag is appended to a protein of interest, which is bound by the TIR1 protein (or engineered variants thereof) upon addition of auxin or chemical

derivatives. TIR1 is exogenously expressed for instance in cell lines, and forms an E3 ligase complex leading to ubiquitination of the degron tag and decay of the tagged protein by the proteasome, which can be achieved in less than one hour after auxin addition [407]. Biological reasons could also affect the absence of a clear phenotype: Yi et al. showed that PAN2-PAN3 and PARN knockdowns had very little or no detectable effects upon siRNA-mediated knockdown and strong lengthening of poly(A) tails was only observed upon simultaneous knockdown of CNOT7 and CNOT8. In this study only the CNOT7 subunit was depleted, which may by itself be insufficient for abolishing deadenylation.

Different models have been suggested regarding the coordinated action of deadenylase complexes. Yamashita and colleagues proposed a model in which poly(A) tails are initially trimmed by the PAN2-PAN3 complex, and then completely deadenylated by the CCR4-NOT complex [250], although in this study only beta-globin was examined. How general this model is remained questionable, since the knockdown experiments performed by Yi et al. did not show strong effects upon PAN2-PAN3 knockdown, where an enrichment of long poly(A) tails would have been expected. It must be noted that the experimental approaches used by the two studies differed: While Yi et al. used siRNAs and expression of dominant negative mutants, Yamashita and colleagues overexpressed deadenylase subunits leading to gain of function phenotypes. PAN2-PAN3 was initially considered the top candidate gene for nuclear deadenylation, since the complex has been proposed to mediate the first phase of deadenylation, and since a role in nuclear deadenylation was suggested for the yeast Pan2-Pan3 complex [214]. Five days of shRNA expression did cause a mild increase in nuclear poly(A) tail length of around 12 nt per gene, but the effect could not be validated by repeating the shRNA induction experiments for three days only.

Despite the targeted experimental design, a broader approach for identification of enzymes responsible for nuclear deadenylation could be applied, for instance by performing siRNA or CRISPR-Cas screens with higher throughput, which would enable screening of more deadenylase candidate genes. Another study identified CNOT1 as the CCR4-NOT subunit connected to nuclear poly(A) shortening [248], although the mechanism is unclear since CNOT1 is the scaffold protein without reported deadenylase function.

5.7 Outlook

The work presented in this thesis addressed two important issues in understanding the role of poly(A) tails as dynamic eukaryotic gene regulatory elements which impact different aspects of RNA metabolism: The first part addressed the need for simple assays that enable poly(A) tail length quantification on a genome-wide scale, with low error rates and together with of full-length transcript information. For this, a new sequencing protocol, termed FLAM-Seq, was implemented, including software and statistical tools for data analysis. FLAM-Seq makes used of third generation sequencing technology which is currently transforming the genomics field and could become the standard application in the future. FLAM-Seq enables precise mapping of RNA isoforms, and it was shown that poly(A) tail length is strongly associated with 3'-UTR isoforms and alternative polyadenylation, which highlighted the relevance of investigating complete RNA molecules. The correlations of 3'-UTR and poly(A) tail length are conceptually important for understanding RNA stability: other studies showed that deadenylation rates are the main determinants of RNA stability [265] and *trans*-acting factors, such as destabilizing RNA-binding proteins, which mediate this effect through modulation of deadenylation rates [408]. Poly(A) tail length is hence a hub integrating those signals that contribute to RNA decay. Further studies are yet required to monitor adaptation of deadenylation rates over time for different stimuli, and for different RNA isoforms to understand the impact of gene regulatory elements that differ between isoforms. Experiments involving metabolic labeling of RNA have been applied in this study but were limited by sequencing depth in modeling rates for individual genes.

The second part of this work addresses nuclear poly(A) tail biogenesis and poly(A) dynamics right after transcription. Nuclear mRNA processing is complex and in part difficult to investigate since separation of nuclei is tedious in context of more elaborate experiments. Nuclear RNA decay has also been mostly investigated for non-coding RNA, such as snoRNAs, which are highly abundant in the nucleus, but is essential to degrade products from pervasive transcription of the genome, which could have detrimental effects on a cell when reaching the cytoplasm and possibly being translated.

As hypothesized for many decades, it could be shown that poly(A) tails are universally synthesized at a length of 200 nt or more in all investigated mammalian model systems, and fractionation and labeling experiments showed that deadenylation has a nuclear component which involves shortening of poly(A) tails. Why poly(A) tails are shortened immediately after synthesis is still enigmatic but splicing inhibition experiments showed that deadenylation could

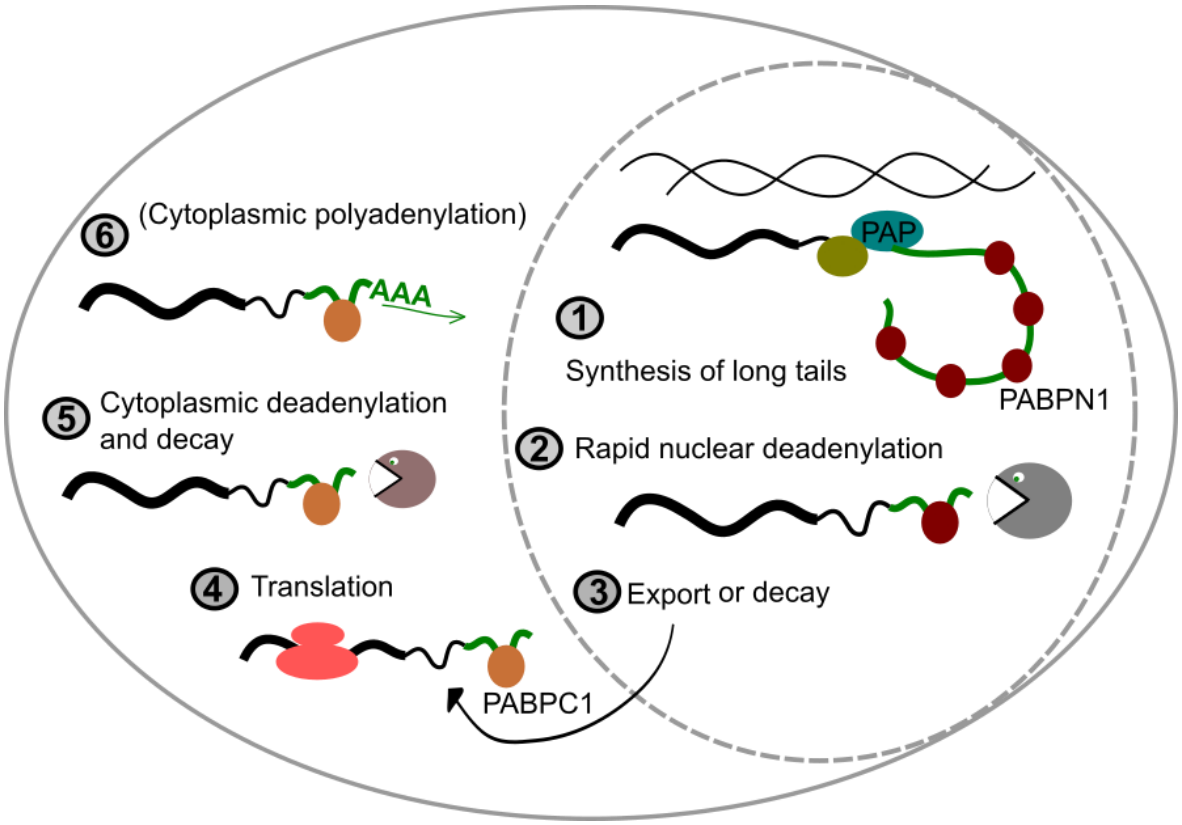

Figure 33 **A unifying model for poly(A) tail metabolism**

be linked to completion of pre-mRNA splicing, which could trigger shortening of tails. Further experiments to investigate the role of poly(A) tails in nuclear trafficking could involve (inducible) nuclear expression of constructs with different poly(A) tail length and tracking of transcripts, for instance using imaging techniques as single-molecule FISH. Since many nuclear lincRNAs were found not to have undergone nuclear deadenylation, this may hint at length dependent export mechanisms. Here it would be interesting to investigate whether nuclear RNA export is impacted upon depleting the enzyme(s) which are responsible for nuclear deadenylation. Nuclear deadenylation may also serve in 'presetting' transcript properties before those reach the translation machinery. If and how poly(A) shortening is mediated during multiple rounds of translation is not well understood in detail, although poly(A) tails are important in specialized decay processes such as non-stop decay, [136] by which parts of the poly(A) tail are translated which leads to RNA decay. A unifying model of poly(A) tail length metabolism is shown in Figure 33 which is summarizing the biological questions addressed in this thesis.

Humans have few more genes than less complex life forms as for instance the nematode *C. elegans* which requires for elaborate regulatory mechanisms layers which control how, when

and where genes are expressed. This work tries to contribute to our understanding of gene regulation, by uncovering new roles for poly(A) tails in controlling gene expression and also by providing the community with new tools for investigating poly(A) tails in order to test new biological hypothesis.

Bibliography

1. Timofeeff-Ressovky, N. W., Zimmer, K. G. & Delbrück, M. Über die Natur der Genmutation und der Genstruktur. *Math. Klasse, Fachgr. VI, Biol. Bd. 1, Nr. 13, 189-245* (1935).
2. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561–563 (1970).
3. Brockdorff, N., Bowness, J. S. & Wei, G. Progress toward understanding chromosome silencing by Xist RNA. *Genes Dev.* **34**, 733–744 (2020).
4. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **21**, 381–395 (2011).
5. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nat. Rev. Mol. Cell Biol.* **20**, 590–607 (2019).
6. Field, A. & Adelman, K. Evaluating Enhancer Function and Transcription. *Annu. Rev. Biochem.* **89**, 213–234 (2020).
7. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: Function, expression and evolution. *Nat. Rev. Genet.* **10**, 252–263 (2009).
8. Cramer, P. *et al.* Structure of eukaryotic RNA polymerases. *Annu. Rev. Biophys.* **37**, 337–352 (2008).
9. Harlen, K. M. & Churchman, L. S. The code and beyond: Transcription regulation by the RNA polymerase II carboxy-terminal domain. *Nat. Rev. Mol. Cell Biol.* **18**, 263–273 (2017).
10. Lipovsek, D. & Plückthun, A. In-vitro protein evolution by ribosome display and mRNA display. *J. Immunol. Methods* **290**, 51–67 (2004).
11. Veloso, A. *et al.* Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Res.* **24**, 896–905 (2014).
12. Fuchs, G. *et al.* 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. *Genome Biol.* **15**, R69 (2014).
13. Tennyson, C. N. & Worton, R. G. The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. *Nature* **9**, (1995).
14. Proudfoot, N. J. Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science.* **352**, 1291 (2016).
15. Bensaude, O. Inhibiting eukaryotic transcription: Which compound to choose? How to evaluate its activity? *Transcription* **2**, 103–108 (2011).
16. Grimm, R. A. *et al.* Actinomycin D in the treatment of advanced breast cancer. *Cancer Chemother. Pharmacol.* **4**, 195–197 (1980).
17. Devarkar, S. C. *et al.* Structural basis for m7G recognition and 2′-O-methyl discrimination in capped RNAs by the innate immune receptor RIG-I. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 596–601 (2016).
18. Galloway, A. & Cowling, V. H. mRNA cap regulation in mammalian cell function and fate. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1862**, 270–279 (2019).
19. Craig Venter, J. *et al.* The sequence of the human genome. *Science (80-.).* **291**, 1304–1351 (2001).
20. Sakharkar, M. K., Chow, V. T. K. & Kangueane, P. Distributions of exons and introns in the human genome. *In Silico Biol.* **4**, 387–393 (2004).
21. Jorquera, R. *et al.* SinEx DB: A database for single exon coding sequences in mammalian genomes. *Database* **2016**, 1–8 (2016).
22. Wahl, M. C., Will, C. L. & Lührmann, R. The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell* **136**, 701–718 (2009).
23. Mercer, T. R. *et al.* Genome-wide discovery of human splicing branchpoints. *Genome Res.* **25**, 290–303 (2015).
24. Hir, H. Le, Saulière, J. & Wang, Z. The exon junction complex as a node of post-transcriptional networks. *Nat. Rev. Mol. Cell Biol.* **17**, 41–54 (2016).
25. Effenberger, K. A., Urabe, V. K. & Jurica, M. S. Modulating splicing with small molecular inhibitors of the spliceosome. *Wiley Interdiscip. Rev. RNA* **8**, (2017).
26. Zhou, Z. *et al.* The biological function and clinical significance of SF3B1 mutations in cancer. *Biomark. Res.* **8**, 1–14 (2020).
27. Kaida, D. *et al.* Spliceostatin A targets SF3b and inhibits both splicing and nuclear retention of pre-mRNA. *Nat. Chem. Biol.* **3**, 576–583 (2007).
28. Kotake, Y. *et al.* Splicing factor SF3b as a target of the antitumor natural product pladienolide. *Nat. Chem. Biol.* **3**, 570–575 (2007).

29. Neuvéglise, C., Marck, C. & Gaillardin, C. The intronome of budding yeasts. *Comptes Rendus - Biol.* **334**, 662–670 (2011).
30. Berget, S. M. Exon recognition in vertebrate splicing. *J. Biol. Chem.* **270**, 2411–2415 (1995).
31. Nilsen, T. W. & Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**, 457–463 (2010).
32. Matlin, A. J., Clark, F. & Smith, C. W. J. Understanding alternative splicing: Towards a cellular code. *Nat. Rev. Mol. Cell Biol.* **6**, 386–398 (2005).
33. Barash, Y. *et al.* Deciphering the splicing code. *Nature* **465**, 53–59 (2010).
34. De La Mata, M. *et al.* A slow RNA polymerase II affects alternative splicing in vivo. *Mol. Cell* **12**, 525–532 (2003).
35. Fong, N., Saldi, T., Sheridan, R. M., Cortazar, M. A. & Bentley, D. L. RNA Pol II Dynamics Modulate Co-transcriptional Chromatin Modification, CTD Phosphorylation, and Transcriptional Direction. *Mol. Cell* 1–12 (2017) doi:10.1016/j.molcel.2017.04.016.
36. Maslon, M. M. *et al.* A slow transcription rate causes embryonic lethality and perturbs kinetic coupling of neuronal genes. *EMBO J.* **38**, 1–18 (2019).
37. Kornblihtt, A. R. Transcriptional control of alternative splicing along time: Ideas change, experiments remain. *RNA* **21**, 670–672 (2015).
38. Tilgner, H. *et al.* Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* **22**, 1616–1625 (2012).
39. Vargas, D. Y. *et al.* Single-Molecule Imaging of Transcriptionally Coupled and Uncoupled Splicing. *Cell* **147**, 1054–1065 (2012).
40. Ameur, A. *et al.* Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat. Struct. Mol. Biol.* **18**, 1435–1440 (2011).
41. Saldi, T., Cortazar, M. A., Sheridan, R. M. & Bentley, D. L. Coupling of RNA Polymerase II Transcription Elongation with Pre-mRNA Splicing. *J. Mol. Biol.* **428**, 2623–2635 (2016).
42. Coté, A. *et al.* The spatial distributions of pre-mRNAs suggest post-transcriptional splicing of specific introns within endogenous genes. *bioRxiv* (2020) doi:10.1101/2020.04.06.028092.
43. Drexler, H. L., Choquet, K. & Churchman, L. S. Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores. *Mol. Cell* **77**, 985-998.e8 (2020).
44. Alpert, T., Herzel, L. & Neugebauer, K. M. Perfect timing: splicing and transcription rates in living cells. *Wiley Interdiscip. Rev. RNA* **8**, 1–12 (2017).
45. Singh, J. & Padgett, R. A. Rates of in situ transcription and splicing in large human genes. *Nat. Struct. Mol. Biol.* **16**, 1128–1133 (2009).
46. Rabani, M. *et al.* High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. *Cell* **159**, 1698–1710 (2014).
47. Lacadie, S. A., Tardiff, D. F., Kadener, S. & Rosbash, M. In vivo commitment to yeast cotranscriptional splicing is sensitive to transcription elongation mutants. *Genes Dev.* **20**, 2055–2066 (2006).
48. Barrass, J. D. *et al.* Transcriptome-wide RNA processing kinetics revealed using extremely short 4tU labeling. *Genome Biol.* **16**, 1–17 (2015).
49. Carrillo Oesterreich, F., Preibisch, S. & Neugebauer, K. M. Global analysis of nascent rna reveals transcriptional pausing in terminal exons. *Mol. Cell* **40**, 571–581 (2010).
50. Niwa, M. & Berget, S. M. Mutation of the AAUAAA polyadenylation signal depresses in vitro splicing of proximal but not distal introns. *Genes Dev.* **5**, 2086–2095 (1991).
51. Vagner, S., Vagner, C. & Mattaj, I. W. The carboxyl terminus of vertebrate poly(A) polymerase interacts with U2AF 65 to couple 3'-end processing and splicing. *Genes Dev.* **14**, 403–413 (2000).
52. Yap, K., Lim, Z. Q., Khandelia, P., Friedman, B. & Makeyev, E. V. Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention. *Genes Dev.* **26**, 1209–1223 (2012).
53. Rabani, M. *et al.* Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat. Biotechnol.* **29**, 436–442 (2011).
54. Boutz, P. L., Bhutkar, A. & Sharp, P. A. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes Dev.* **29**, 63–80 (2015).
55. Hansen, T. B. *et al.* Natural RNA circles function as efficient microRNA sponges. *Nature* **495**, 384–388 (2013).
56. Memczak, S., Jens, M., Elefsinioti, A. & Torti, F. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495**, 333–8 (2013).

57. Ashwal-Fluss, R. *et al.* CircRNA Biogenesis competes with Pre-mRNA splicing. *Mol. Cell* **56**, 55–66 (2014).
58. Stover, N. A., Kaye, M. S. & Cavalcanti, A. R. O. Spliced leader trans-splicing. *Curr. Biol.* **16**, 8–9 (2006).
59. Galganski, L., Urbanek, M. O. & Krzyzosiak, W. J. Nuclear speckles: Molecular organization, biological function and role in disease. *Nucleic Acids Res.* **45**, 10350–10368 (2017).
60. Girard, C. *et al.* Post-transcriptional spliceosomes are retained in nuclear speckles until splicing completion. *Nat. Commun.* **3**, (2012).
61. Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J. M. & Gautheret, D. Patterns of variant polyadenylation signal usage in human genes. *Genome Res.* **10**, 1001–1010 (2000).
62. Chan, S. L. *et al.* CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3′ processing. *Genes Dev.* **28**, 2370–2380 (2014).
63. Mandel, C. R. *et al.* Polyadenylation factor CPSF-73 is the pre-mRNA 3′-end-processing endonuclease. *Nature* **444**, 953–956 (2006).
64. Chan, S., Choi, E. A. & Shi, Y. Pre-mRNA 3′-end processing complex assembly and function. *Wiley Interdiscip. Rev. RNA* **2**, 321–335 (2011).
65. Schreieck, A. *et al.* RNA polymerase II termination involves C-terminal-domain tyrosine dephosphorylation by CPF subunit Glc7. *Nat. Struct. Mol. Biol.* **21**, 175–179 (2014).
66. Kumar, A., Clerici, M., Muckenfuss, L. M., Passmore, L. A. & Jinek, M. Mechanistic insights into mRNA 3′-end processing. *Curr. Opin. Struct. Biol.* **59**, 143–150 (2019).
67. Nandagopal, N. & Roux, P. P. Regulation of global and specific mRNA translation by the mTOR signaling pathway. *Translation* **3**, e983402 (2015).
68. Yang, Y. & Wang, Z. IRES-mediated cap-independent translation, a path leading to hidden proteome. *J. Mol. Cell Biol.* **11**, 911–919 (2019).
69. Mayr, C. What are 3′ UTRs doing? *Cold Spring Harb. Perspect. Biol.* **11**, (2019).
70. Yaffe, D., Nudel, U., Mayer, Y. & Neuman, S. Highly conserved sequencim in the 3′ untranslated region of mRNAs coding for homologous proteins in distantly related species. *Nucleic Acids Res.* **13**, 3723–3737 (1985).
71. Wang, W. *et al.* Evolutionary and functional implications of 3′ untranslated region length of mRNAs by comprehensive investigation among four taxonomically diverse metazoan species. *Genes and Genomics* **41**, 747–755 (2019).
72. Mayr, C. Regulation by 3'-Untranslated Regions. *Annu. Rev. of Genetics* **51**, 171–94 (2017).
73. Shyu, A. B., Greenberg, M. E. & Belasco, J. G. The c-fos transcript is targeted for rapid decay by two distinct mRNA degradation pathways. *Genes Dev.* **3**, 60–72 (1989).
74. Chen, C. Y. *et al.* AU binding proteins recruit the exosome to degrade ARE-containing mRNAs. *Cell* **107**, 451–464 (2001).
75. Lykke-Andersen, J. & Wagner, E. Recruitment and activation of mRNA decay enzymes by two ARE-mediated decay activation domains in the proteins TTP and BRF-1. *Genes Dev.* **19**, 351–361 (2005).
76. Hau, H. H. *et al.* Tristetraprolin recruits functional mRNA decay complexes to ARE sequences. *J. Cell. Biochem.* **100**, 1477–1492 (2007).
77. Mukherjee, D. *et al.* The mammalian exosome mediates the efficient degradation of mRNAs that contain AU-rich elements. *EMBO J.* **21**, 165–174 (2002).
78. Fabian, M. R. *et al.* Structural basis for the recruitment of the human CCR4-NOT deadenylase complex by tristetraprolin. *Nat. Struct. Mol. Biol.* **20**, 735–739 (2013).
79. Fan, X. C. & Steitz, J. A. Overexpression of HuR, a nuclear-cytoplasmic shuttling protein, increases the in vivo stability of ARE-containing mRNAs. *EMBO J.* **17**, 3448–3460 (1998).
80. Linker, K. *et al.* Involvement of KSRP in the post-transcriptional regulation of human iNOS expression-complex interplay of KSRP with TTP and HuR. *Nucleic Acids Res.* **33**, 4813–4827 (2005).
81. Gebert, L. F. R. & MacRae, I. J. Regulation of microRNA function in animals. *Nat. Rev. Mol. Cell Biol.* **20**, 21–37 (2019).
82. De La Mata, M. *et al.* Potent degradation of neuronal miRNAs induced by highly complementary targets. *EMBO Rep.* **16**, 500–511 (2015).
83. Kristjánsdóttir, K., Fogarty, E. A. & Grimson, A. Systematic analysis of the Hmga2 3' UTR identifies many independent regulatory sequences and a novel interaction between distal sites. *RNA* **21**, 1346–1360 (2015).
84. Vella, M. C., Choi, E. Y., Lin, S. Y., Reinert, K. & Slack, F. J. The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3′UTR. *Genes Dev.* **18**, 132–137 (2004).

85. Zappulo, A. *et al.* RNA localization is a key determinant of neurite-enriched proteome. *Nat. Commun.* **8**, 1–12 (2017).
86. Biever, A. *et al.* Monosomes actively translate synaptic mRNAs in neuronal processes. *Science.* **367**, (2020).
87. Irion, U. & St Johnston, D. bicoid RNA localization requires specific binding of an endosomal sorting complex. *Nature* **445**, 554–558 (2007).
88. Kislauskis, E. H., Zhu, X. & Singer, R. H. Sequences responsible for intracellular localization of β-actin messenger RNA also affect cell phenotype. *J. Cell Biol.* **127**, 441–451 (1994).
89. Jambhekar, A. & Derisi, J. L. Cis-acting determinants of asymmetric, cytoplasmic RNA transport. *RNA* **13**, 625–642 (2007).
90. Sandberg, R., Neilson, J., Sarma, A., Sharp, P. A. & Burge, C. B. Proliferating Cells Express mRNAs with Shortened 3′ Untranslated Regions and Fewer MicroRNA Target Sites. *Science.* **320**, 1643–1647 (2008).
91. Spies, N., Burge, C. B. & Bartel, D. P. 3′ UTR-Isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res.* **23**, 2078–2090 (2013).
92. Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S. & Mayr, C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.* **27**, 2380–2396 (2013).
93. Ulitsky, I. *et al.* Extensive alternative polyadenylation during zebrafish development. *Genome Res.* **22**, 2054–2066 (2012).
94. Tian, B., Hu, J., Zhang, H. & Lutz, C. S. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.* **33**, 201–212 (2005).
95. Ozsolak, F. *et al.* Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* **143**, 1018–1029 (2010).
96. Lee, S. H. *et al.* Widespread intronic polyadenylation inactivates tumour suppressor genes in leukaemia. *Nature* **561**, 127–131 (2018).
97. Luo, W. *et al.* The Conserved Intronic Cleavage and Polyadenylation Site of CstF-77 Gene Imparts Control of 3′ End Processing Activity through Feedback Autoregulation and by U1 snRNP. *PLoS Genet.* **9**, 1–14 (2013).
98. Mayr, C. & Bartel, D. P. Widespread Shortening of 3′UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. *Cell* **138**, 673–684 (2009).
99. Li, Z., Lee, J. Y., Pan, Z., Bingjun, J. & Tian, B. Progressive lengthening of 3'untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc. Natl. Acad. Sci.* **106**, 9535 (2009).
100. Miura, P., Shenker, S., Andreu-Agullo, C., Westholm, J. O. & Lai, E. C. Widespread and extensive lengthening of 3'UTRs in the mammalian brain. *Genome Res.* **23**, 812–825 (2013).
101. Hilgers, V. *et al.* Neural-specific elongation of 3′ UTRs during Drosophila development. *Proc. Natl. Acad. Sci.* **108**, 15864–15869 (2011).
102. Lackford, B. *et al.* Fip1 regulates mRNA alternative polyadenylation to promote stem cell self-renewal. *EMBO J.* **33**, 878–889 (2014).
103. Geisberg, J. V., Moqtaderi, Z. & Struhl, K. The transcriptional elongation rate regulates alternative polyadenylation in yeast. *Elife* **9**, 1–55 (2020).
104. Takagaki, Y., Seipelt, R. L., Peterson, M. L. & Manley, J. L. The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. *Cell* **87**, 941–952 (1996).
105. Berkovits, B. D. & Mayr, C. Alternative 3′ UTRs act as scaffolds to regulate membrane protein localization. *Nature* **522**, 363–367 (2015).
106. Hilgers, V., Lemke, S. B. & Levine, M. ELAV mediates 3′ UTR extension in the Drosophila nervous system. *Genes Dev.* **26**, 2259–2264 (2012).
107. Oktaba, K. *et al.* ELAV links paused pol II to alternative polyadenylation in the drosophila nervous system. *Mol. Cell* **57**, 341–348 (2015).
108. Misra, A., Ou, J., Zhu, L. J. & Green, M. R. Global Promotion of Alternative Internal Exon Usage by mRNA 3' End Formation Factors. *Mol. Cell* **58**, 819–831 (2014).
109. Gunderson, S. I., Polycarpou-Schwarz, M. & Mattaj, I. W. U1 snRNP inhibits pre-mRNA polyadenylation through a direct interaction between U1 70K and poly(A) polymerase. *Mol. Cell* **1**, 255–264 (1998).
110. Almada, A. E., Wu, X., Kriz, A. J., Burge, C. B. & Sharp, P. A. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**, 360–363 (2013).

111. Reed, R. & Cheng, H. TREX, SR proteins and export of mRNA. *Curr. Opin. Cell Biol.* **17**, 269–273 (2005).
112. Björk, P. & Wieslander, L. Integration of mRNP formation and export. *Cell. Mol. Life Sci.* **74**, 2875–2897 (2017).
113. Stewart, M. Ratcheting mRNA out of the Nucleus. *Mol. Cell* **25**, 327–330 (2007).
114. Vicens, Q., Kieft, J. S. & Rissland, O. S. Revisiting the Closed-Loop Model and the Nature of mRNA 5′–3′ Communication. *Mol. Cell* **72**, 805–812 (2018).
115. Adivarahan, S. *et al.* Spatial Organization of Single mRNPs at Different Stages of the Gene Expression Pathway. *Mol. Cell* **72**, 727-738.e5 (2018).
116. Maquat, L. E., Tarn, W. Y. & Isken, O. The pioneer round of translation: Features and functions. *Cell* **142**, 368–374 (2010).
117. Kurosaki, T., Popp, M. W. & Maquat, L. E. Quality and quantity control of gene expression by nonsense-mediated mRNA decay. *Nat. Rev. Mol. Cell Biol.* **20**, 406–420 (2019).
118. Schwannhäuser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
119. Tani, H. *et al.* Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Research.* **22**,947–956 (2012)
120. Herzog, V. A. *et al.* Thiol-linked alkylation of RNA to assess expression dynamics. *Nat. Methods* **14**, 1198–1204 (2017).
121. Wang, Y. *et al.* Precision and functional specificity in mRNA decay. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 5860–5865 (2002).
122. Stoecklin, G. & Anderson, P. Posttranscriptional Mechanisms Regulating the Inflammatory Response. *Adv. Immunol.* **89**, 1–37 (2006).
123. Bakheet, T., Hitti, E. & Khabar, K. S. A. ARED-Plus: An updated and expanded database of AU-rich element-containing mRNAs and pre-mRNAs. *Nucleic Acids Res.* **46**, D218–D220 (2018).
124. Wisdom, R. & Lee, W. The protein-coding region of c-myc mRNA contains a sequence that specifies rapid mRNA turnover and induction by protein synthesis inhibitors. *Genes Dev.* **5**, 232–243 (1991).
125. Tchen, C. R., Brook, M., Saklatvala, J. & Clark, A. R. The stability of tristetraprolin mRNA is regulated by mitogen-activated protein kinase p38 and by tristetraprolin itself. *J. Biol. Chem.* **279**, 32393–32400 (2004).
126. Brook, M. *et al.* Posttranslational Regulation of Tristetraprolin Subcellular Localization and Protein Stability by p38 Mitogen-Activated Protein Kinase and Extracellular Signal-Regulated Kinase Pathways. *Mol. Cell. Biol.* **26**, 2408–2418 (2006).
127. Shim, J. & Karin, M. The control of mRNA stability in response to extracellular stimuli. *Mol. Cells* **14**, 323–331 (2002).
128. Rabani, M., Pieper, L., Chew, G. L. & Schier, A. F. A Massively Parallel Reporter Assay of 3′ UTR Sequences Identifies In Vivo Rules for mRNA Degradation. *Mol. Cell* **68**, 1083-1094.e5 (2017).
129. Vejnar, C. E. *et al.* Genome wide analysis of 3′ UTR sequence elements and proteins regulating mRNA stability during maternal-to-zygotic transition in zebrafish. *Genome Res.* **29**, 1100–1114 (2019).
130. Conrad, N. K., Mili, S., Marshall, E. L., Shu, M. Di & Steitz, J. A. Identification of a Rapid Mammalian Deadenylation-Dependent Decay Pathway and Its Inhibition by a Viral RNA Element. *Mol. Cell* **24**, 943–953 (2006).
131. Beelman, C. A. *et al.* An essential component of the decapping enzyme required for normal rates of mRNA turnover. *Nature* **382**, 642–646 (1996).
132. Chowdhury, A., Mukhopadhyay, J. & Tharun, S. The decapping activator Lsm1p-7p-Pat1p complex has the intrinsic ability to distinguish between oligoadenylated and polyadenylated RNAs. *RNA* **13**, 998–1016 (2007).
133. Tharun, S. *et al.* Yeast Sm-like proteins function in mRNA decapping and decay. *Nature* **404**, 515–518 (2000).
134. Nagarajan, V. K., Jones, C., Newbury, S. & Green, P. J. XRN 5'→3' exoribonucleases: Structure, mechanisms and functions. *Biochim Biophys Acta*. **1829**, 590–603 (2013).
135. Houseley, J., LaCava, J. & Tollervey, D. RNA-quality control by the exosome. *Nat. Rev. Mol. Cell Biol.* **7**, 529–539 (2006).
136. Garneau, N. L., Wilusz, J. & Wilusz, C. J. The highways and byways of mRNA decay. *Nat. Rev. Mol. Cell Biol.* **8**, 113–126 (2007).
137. Badis, G., Saveanu, C., Fromont-Racine, M. & Jacquier, A. Targeted mRNA degradation by deadenylation-independent decapping. *Mol. Cell* **15**, 5–15 (2004).

138. Luo, Y., Na, Z. & Slavoff, S. A. P-Bodies: Composition, Properties, and Functions. *Biochemistry* **57**, 2424–2431 (2018).
139. Zhao, L. *et al.* NONCODEV6: An updated database dedicated to long non-coding RNA annotation in both animals and plants. *Nucleic Acids Res.* **49**, D165–D171 (2021).
140. Piovesan, A. *et al.* Human protein-coding genes and gene feature statistics in 2019. *BMC Res. Notes* **12**, 1–5 (2019).
141. Cabili, M. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).
142. Mukherjee, N. *et al.* Integrative classification of human coding and noncoding genes through RNA metabolism profiles. *Nat. Struct. Mol. Biol.* **24**, 86–96 (2017).
143. Palazzo, A. F. & Lee, E. S. Sequence determinants for nuclear retention and cytoplasmic export of mRNAs and lncRNAs. *Front. Genet.* **9**, 1–16 (2018).
144. Miyagawa, R. *et al.* Identification of cis- and trans-acting factors involved in the localization of MALAT-1 noncoding RNA to nuclear speckles. *RNA* **18**, 738–751 (2012).
145. Yin, Y. *et al.* U1 snRNP regulates chromatin retention of noncoding RNAs. *Nature* **580**, 147–150 (2020).
146. Edmonds, M. & Caramela, M. G. The isolation and characterization of adenosine monophosphate-rich polynucleotides synthesized by Ehrlich ascites cells. *J. Biol. Chem.* **244**, 1314–1324 (1969).
147. Edmonds, M. & Abrams, R. Polynucleotide biosynthesis: formation of a sequence of adenylate units from adenosine triphosphate by an enzyme from thymus nuclei. *J. Biol. Chem.* **235**, 1142–1149 (1960).
148. Kates, J. & Beeson, J. Ribonucleic acid synthesis in vaccinia virus. I. The mechanism of synthesis and release of RNA in vaccinia cores. *J. Mol. Biol.* **50**, 1–18 (1970).
149. Sheiness, D. & Darnell, J. E. Polyadenylic acid segment becomes shorter with age. *Nat. New Biol.* **241**, 265–268 (1973).
150. Dreyfus, M. & Régnier, P. The poly(A) tail of mRNAs: Bodyguard in eukaryotes, scavenger in bacteria. *Cell* **111**, 611–613 (2002).
151. Temperley, R. J., Wydro, M., Lightowlers, R. N. & Chrzanowska-Lightowlers, Z. M. Human mitochondrial mRNAs-like members of all families, similar but different. *Biochim. Biophys. Acta - Bioenerg.* **1797**, 1081–1085 (2010).
152. Rorbach, J., Nicholls, T. J. J. & Minczuk, M. PDE12 removes mitochondrial RNA poly(A) tails and controls translation in human mitochondria. *Nucleic Acids Res.* **39**, 7750–7763 (2011).
153. Osinga, K. A., De Vries, E., Van der Horst, G. & Tabak, H. F. Processing of yeast mitochondrial messenger RNAs at a conserved dodecamer sequence. *EMBO J.* **3**, 829–834 (1984).
154. Schäfer, B., Hansen, M. & Lang, B. F. Transcription and RNA-processing in fission yeast mitochondria. *RNA* **11**, 785–795 (2005).
155. Li, H. & Zassenhaus, H. P. Purification and characterization of an RNA dodecamer sequence binding protein from mitochondria of Saccharomyces cerevisiae. *Biochem. Biophys. Res. Commun.* **261**, 740–745 (1999).
156. Schuster, G. & Stern, D. RNA Polyadenylation and Decay in Mitochondria and Chloroplasts. *Progress in Molecular Biology and Translational Science* Vol 85 (2009).
157. Subtelny, A. O., Eichhorn, S. W., Chen, G. R., Sive, H. & Bartel, D. P. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature* **508**, 66–71 (2014).
158. Lima, S. A. *et al.* Short poly(A) tails are a conserved feature of highly expressed genes. *Nat. Struct. Mol. Biol.* **24**, 1057–1063 (2017).
159. Palatnik, C. M., Storti, R. V. & Jacobson, A. Fractionation and functional analysis of newly synthesized and decaying messenger RNAs from vegetative cells of Dictyostelium discoideum. *J. Mol. Biol.* **128**, 371–395 (1979).
160. Eichhorn, S. W. *et al.* mRNA poly(A)-tail changes specified by deadenylation broadly reshape translation in Drosophila oocytes and early embryos. *Elife* **5**, 1–24 (2016).
161. Kaufmann, I., Martin, G., Friedlein, A., Langen, H. & Keller, W. Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. *EMBO J.* **23**, 616–626 (2004).
162. Kerwitz, Y. *et al.* Stimulation of poly(A) polymerase through a direct interaction with the nuclear poly(A) binding protein allosterically regulated by RNA. *EMBO J.* **22**, 3705–3714 (2003).
163. Keller, R. W. *et al.* The nuclear poly(A) binding protein, PABP2, forms an oligomeric particle covering the length of the poly(A) tail. *J. Mol. Biol.* **297**, 569–583 (2000).
164. Wahle, E. Poly(A) tail length control is caused by termination of processive synthesis. *J. Biol. Chem.* **270**, 2800–2808 (1995).

165. Mendecki, J., Lee, S. Y. & Brawerman, G. Characteristics of the Polyadenylic Acid Segment Associated with Messenger Ribonucleic Acid in Mouse Sarcoma 180 Ascites Cells. *Biochemistry* **11**, 792–798 (1972).
166. Preker, P. J., Ohnacker, M., Minvielle-Sebastia, L. & Keller, W. A multisubunit 3' end processing factor from yeast containing poly(A) polymerase and homologues of the subunits of mammalian cleavage and polyadenylation specificity factor. *EMBO J.* **16**, 4727–4737 (1997).
167. Dheur, S., Nykamp, K. R., Viphakone, N., Swanson, M. S. & Minvielle-Sebastia, L. Yeast mRNA poly(A) tail length control can be reconstituted in vitro in the absence of Pab1p-dependent poly(A) nuclease activity. *J. Biol. Chem.* **280**, 24532–24538 (2005).
168. Kelly, S. M. *et al.* Recognition of polyadenosine RNA by the zinc finger domain of nuclear poly(A) RNA-binding protein 2 (Nab2) is required for correct mRNA 3′-end formation. *J. Biol. Chem.* **285**, 26022–26032 (2010).
169. Hector, R. E. *et al.* Dual requirement for yeast hnRNP Nab2p in mRNA poly(A) tail length control and nuclear export. *EMBO J.* **21**, 1800–1810 (2002).
170. Kelly, S. M. *et al.* A conserved role for the zinc finger polyadenosine RNA binding protein, ZC3H14, in control of poly(A) tail length. *Rna* **20**, 681–688 (2014).
171. Rha, J. *et al.* The RNA-binding protein, ZC3H14, is required for proper poly(A) tail length control, expression of synaptic proteins, and brain function in mice. *Hum. Mol. Genet.* **26**, 3663–3681 (2017).
172. Laishram, R. S. Poly(A) polymerase (PAP) diversity in gene expression - Star-PAP vs canonical PAP. *FEBS Lett.* **588**, 2185–2197 (2014).
173. Kashiwabara, S. I. *et al.* Identification of a novel isoform of poly(A) polymerase, TPAP, specifically present in the cytoplasm of spermatogenic cells. *Dev. Biol.* **228**, 106–115 (2000).
174. Mellman, D. L. *et al.* A PtdIns4,5P2-regulated nuclear poly(A) polymerase controls expression of select mRNAs. *Nature* **451**, 1013–1017 (2008).
175. Kühn, U., Buschmann, J. & Wahle, E. The nuclear poly(A) binding protein of mammals, but not of fission yeast, participates in mRNA polyadenylation. *RNA* **23**:473-482 (2017)
176. Sachs, A. B., Davis, R. W. & Kornberg, R. D. A single domain of yeast poly(A)-binding protein is necessary and sufficient for RNA binding and cell viability. *Mol. Cell. Biol.* **7**, 3268–3276 (1987).
177. Jenal, M. *et al.* The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites. *Cell* **149**, 538–553 (2012).
178. Muniz, L., Davidson, L. & West, S. Poly(A) Polymerase and the Nuclear Poly(A) Binding Protein, PABPN1, Coordinate the Splicing and Degradation of a Subset of Human Pre-mRNAs. *Mol. Cell. Biol.* **35**, 2218–2230 (2015).
179. Brais, B. *et al.* Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. *Nat. Genet.* **18**, 164–167 (1998).
180. Abbassi-Daloii, T. *et al.* An alanine expanded PABPN1 causes increased utilization of intronic polyadenylation sites. *npj Aging Mech. Dis.* **3**, 1–8 (2017).
181. Beaulieu, Y. B., Kleinman, C. L., Landry-Voyer, A. M., Majewski, J. & Bachand, F. Polyadenylation-Dependent Control of Long Noncoding RNA Expression by the Poly(A)-Binding Protein Nuclear 1. *PLoS Genet.* **8**, (2012).
182. Meola, N. *et al.* Identification of a Nuclear Exosome Decay Pathway for Processed Transcripts. *Mol. Cell* **64**, 520–533 (2016).
183. Benoit, B. *et al.* An essential cytoplasmic function for the nuclear poly(A) binding protein, PABP2, in poly(A) tail length control and early development in Drosophila. *Dev. Cell* **9**, 511–522 (2005).
184. Loh, B., Jonas, S. & Izaurralde, E. The SMG5-SMG7 heterodimer directly recruits the CCR4-NOT deadenylase complex to mRNAs containing nonsense codons via interaction with POP2. *Genes Dev.* **27**, 2125–2138 (2013).
185. Nagpal, N. & Agarwal, S. Telomerase RNA processing: Implications for human health and disease. *Stem Cells* **38**, 1532–1543 (2020).
186. Tseng, C. K. *et al.* Human Telomerase RNA Processing and Quality Control. *Cell Rep.* **13**, 2232–2243 (2015).
187. Nguyen, D. *et al.* A Polyadenylation-Dependent 3′ End Maturation Pathway Is Required for the Synthesis of the Human Telomerase RNA. *Cell Rep.* **13**, 2244–2257 (2015).
188. Marzluff, W. F., Wagner, E. J. & Duronio, R. J. Metabolism and regulation of canonical histone mRNAs: Life without a poly(A) tail. *Nat. Rev. Genet.* **9**, 843–854 (2008).
189. Pandey, N. B., Chodchoy, N., Liu, T. J. & Marzluff, W. F. Introns in histone genes alter the distribution of 3′ ends. *Nucleic Acids Res.* **18**, 3161–3170 (1990).

190. Wyers, F. *et al.* Cryptic Pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell* **121**, 725–737 (2005).
191. Preker, P. *et al.* RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters. *Science* 1851–1854 (2008).
192. Gudipati, R. K. *et al.* Extensive Degradation of RNA Precursors by the Exosome in Wild-Type Cells. *Mol. Cell* **48**, 409–421 (2012).
193. Bousquet-Antonelli, C., Presutti, C. & Tollervey, D. Identification of a regulated pathway for nuclear pre-mRNA turnover. *Cell* **102**, 765–775 (2000).
194. Ogami, K., Chen, Y. & Manley, J. L. RNA surveillance by the nuclear RNA exosome: Mechanisms and significance. *Non-coding RNA* **4**, (2018).
195. Schmid, M. *et al.* The Nuclear PolyA-Binding Protein Nab2p Is Essential for mRNA Production. *Cell Rep.* **12**, 128–139 (2015).
196. Milligan, L., Torchet, C., Allmang, C., Shipman, T. & Tollervey, D. A Nuclear Surveillance Pathway for mRNAs with Defective Polyadenylation. *Mol. Cell. Biol.* **25**, 9996–10004 (2005).
197. Holub, P. *et al.* Air2p is critical for the assembly and RNA-binding of the TRAMP complex and the KOW domain of Mtr4p is crucial for exosome activation. *Nucleic Acids Res.* **40**, 5679–5693 (2012).
198. Vaňáčová, Š. *et al.* A New Yeast Poly(A) Polymerase Complex Involved in RNA Quality Control. *PLoS Biol.* **3**, e189 (2005).
199. Delan-Forino, C., Spanos, C., Rappsilber, J. & Tollervey, D. Substrate specificity of the TRAMP nuclear surveillance complexes. *Nat. Commun.* **11**, (2020).
200. Jia, H. *et al.* The RNA helicase Mtr4p modulates polyadenylation in the TRAMP complex. *Cell* **145**, 890–901 (2011).
201. Schmidt, K. & Butler, J. S. Nuclear RNA surveillance: Role of TRAMP in controlling exosome specificity. *Wiley Interdiscip. Rev. RNA* **4**, 217–231 (2013).
202. Lubas, M. *et al.* Interaction Profiling Identifies the Human Nuclear Exosome Targeting Complex. *Mol. Cell* **43**, 624–637 (2011).
203. Shcherbik, N., Wang, M., Lapik, Y. R., Srivastava, L. & Pestov, D. G. Polyadenylation and degradation of incomplete RNA polymerase i transcripts in mammalian cells. *EMBO Rep.* **11**, 106–111 (2010).
204. Bresson, S. M. & Conrad, N. K. The Human Nuclear Poly(A)-Binding Protein Promotes RNA Hyperadenylation and Decay. *PLoS Genet.* **9**, (2013).
205. Bresson, S. M., Hunter, O. V., Hunter, A. C. & Conrad, N. K. Canonical Poly(A) Polymerase Activity Promotes the Decay of a Wide Variety of Mammalian Nuclear RNAs. *PLoS Genet.* **11**, 1–25 (2015).
206. Lubas, M. *et al.* The human nuclear exosome targeting complex is loaded onto newly synthesized RNA to direct early ribonucleolysis. *Cell Rep.* **10**, 178–192 (2015).
207. Dower, K., Kuperwasser, N., Merrikh, H. & Rosbash, M. A synthetic A tail rescues yeast nuclear accumulation of a ribozyme-terminated transcript. *RNA* **10**, 1888–1899 (2004).
208. Aibara, S., Gordon, J. M. B., Riesterer, A. S., McLaughlin, S. H. & Stewart, M. Structural basis for the dimerization of Nab2 generated by RNA binding provides insight into its contribution to both poly(A) tail length determination and transcript compaction in Saccharomyces cerevisiae. *Nucleic Acids Res.* **45**, 1529–1538 (2017).
209. Batisse, J., Batisse, C., Budd, A., Böttcher, B. & Hurt, E. Purification of nuclear poly(A)-binding protein Nab2 reveals association with the yeast transcriptome and a messenger ribonucleoprotein core structure. *J. Biol. Chem.* **284**, 34911–34917 (2009).
210. Galy, V. *et al.* Nuclear Retention of Unspliced mRNAs in Yeast Is Mediated by Perinuclear Mlp1. *Cell* **116**, 63–73 (2004).
211. Fasken, M. B., Stewart, M. & Corbett, A. H. Functional significance of the interaction between the mRNA-binding protein, Nab2, and the nuclear pore-associated protein, mlp1, in mRNA export. *J. Biol. Chem.* **283**, 27130–27143 (2008).
212. Qu, X. *et al.* Assembly of an Export-Competent mRNP Is Needed for Efficient Release of the 3′-End Processing Complex after Polyadenylation. *Mol. Cell. Biol.* **29**, 5327–5338 (2009).
213. Silla, T., Karadoulama, E., Mąkosa, D., Lubas, M. & Jensen, T. H. The RNA Exosome Adaptor ZFC3H1 Functionally Competes with Nuclear Export Activity to Retain Target Transcripts. *Cell Rep.* **23**, 2199–2210 (2018).
214. Dunn, E. F., Hammell, C. M., Hodge, C. A. & Cole, C. N. Yeast poly(A)-binding protein, Pab1, and PAN, a poly(A) nuclease complex recruited by Pab1, connect mRNA biogenesis to export. *Genes Dev.* **19**, 90–103 (2005).

215. Jacobson, A. & Favreau, M. Possible Involvement of poly(A) in protein synthesis. *Nucleic Acids Res.* **11**, 6353–6368 (1983).
216. Borman, A. M., Michel, Y. M. & Kean, K. M. Biochemical characterisation of cap-poly(A) synergy in rabbit reticulocyte lysates: The elF4G-PABP interaction increases the functional affinity of elF4E for the capped mRNA 5'-end. *Nucleic Acids Res.* **28**, 4068–4075 (2000).
217. Munroe, D. & Jacobson, A. mRNA poly(A) tail, a 3' enhancer of translational initiation. *Mol. Cell. Biol.* **10**, 3441–3455 (1990).
218. Preiss, T., Muckenthaler, M. & Hentze, M. W. Poly(A)-tail-promoted translation in yeast : implications for translational control. *RNA* **4**, 1321–1331 (1998).
219. Tarun, S. Z. & Sachs, A. B. A common function for mRNA 5' and 3'ends in translation initiation in yeast. *Genes Dev.* **9**, 2997–3007 (1995).
220. Hentze, M. W. & Preiss, T. Dual function of the messenger RNA cap structure in poly(A)-tail-promoted translation in yeast. *Nature* **392**, 516–20 (1998).
221. Gallie, D. R. The cap and poly(A) tail function synergistically to regulate mRNA translational Efficiency. *Genes Dev.* **5**, 2108–2116 (1991).
222. McGrew, L. L., Dworkin-Rastl, E., Dworkin, M. B. & Richter, J. D. Poly(A) elongation during Xenopus oocyte maturation is required for translational recruitment and is mediated by a short sequence element. *Genes Dev.* **3**, 803–815 (1989).
223. Hyman, L. E. & Wormington, W. M. Translational inactivation of ribosomal protein mRNAs during Xenopus oocyte maturation. *Genes Dev.* **2**, 598–605 (1988).
224. Palatnik, C. M., Wilkins, C. & Jacobson, A. Translational control during early dictyostelium Development: Possible involvement of poly(A) sequences. *Cell* **36**, 1017–1025 (1984).
225. Huarte, J., Belin, D., Vassalli, A., Strickland, S. & Vassalli, J. D. Meiotic maturation of mouse oocytes triggers the translation and polyadenylation of dormant tissue-type plasminogen activator mRNA. *Genes Dev.* **1**, 1201–1211 (1987).
226. Restifo, L. L. & Guild, G. Poly(A) Shortening of Coregulated Transcripts in Drosophila. *Dev. Biol.* **115**, 507–510 (1986).
227. Chang, H., Lim, J., Ha, M. & Kim, V. N. TAIL-seq: Genome-wide Determination of Poly(A) Tail Length and 3' End Modifications. *Mol. Cell* **53**, 1044–1052 (2014).
228. Xiang, K. & Bartel, D. P. The molecular basis of coupling between poly(A)-tail length and translational efficiency. *bioRxiv* 2021.01.18.427055 (2021).
229. Paynton, B. V., Rempel, R. & Bachvarova, R. Changes in state of adenylation and time course of degradation of maternal mRNAs during oocyte maturation and early embryonic development in the mouse. *Dev. Biol.* **129**, 304–314 (1988).
230. Stebbins-Boaz, B., Cao, Q., De Moor, C. H., Mendez, R. & Richter, J. D. Maskin is a CPEB-associated factor that transiently interacts with eIF-4E. *Mol. Cell* **4**, 1017–1027 (1999).
231. Belloc, E. & Méndez, R. A deadenylation negative feedback mechanism governs meiotic metaphase arrest. *Nature* **452**, 1017–1021 (2008).
232. Piqué, M., López, J. M., Foissac, S., Guigó, R. & Méndez, R. A Combinatorial Code for CPE-Mediated Translational Control. *Cell* **132**, 434–448 (2008).
233. Rouhana, L. *et al.* Vertebrate GLD2 poly(A) polymerases in the germline and the brain. *RNA* **11,**1117–1130 (2005)
234. Novoa, I., Gallego, J., Ferreira, P. G. & Mendez, R. Mitotic cell-cycle progression is regulated by CPEB1 and CPEB4-dependent translational control. *Nat. Cell Biol.* **12**, 447–456 (2010).
235. Huang, Y. S., Jung, M. Y., Sarkissian, M. & Richter, J. D. N-methyl-D-aspartate receptor signaling results in Aurora kinase-catalyzed CPEB phosphorylation and αCaMKII mRNA polyadenylation at synapses. *EMBO J.* **21**, 2139–2148 (2002).
236. Atkins, C. M., Nozaki, N., Shigeri, Y. & Soderling, T. R. Cytoplasmic polyadenylation element binding protein-dependent protein synthesis is regulated by calcium/calmodulin-dependent protein kinase II. *J. Neurosci.* **24**, 5193–5201 (2004).
237. Decker, C. J. & Parker, R. A turnover pathway for both stable and unstable mRNAs in yeast: Evidence for a requirement for deadenylation. *Genes Dev.* **7**, 1632–1643 (1993).
238. Jonas, S. *et al.* An asymmetric PAN3 dimer recruits a single PAN2 exonuclease to mediate mRNA deadenylation and decay. *Nat. Struct. Mol. Biol.* **21**, 599–608 (2014).
239. Schäfer, I. B. *et al.* Molecular Basis for poly(A) RNP Architecture and Recognition by the Pan2-Pan3 Deadenylase. *Cell* **177**, 1619-1631.e21 (2019).

240. Boeck, R. *et al.* The Yeast Pan2 Protein Is Required for Poly (A) -binding Protein-stimulated Poly (A) -nuclease Activity *. **271**, 432–438 (1996).
241. Wolf, J. & Passmore, L. A. mRNA deadenylation by Pan2–Pan3. *Biochem. Soc. Trans.* **42**, 184–187 (2014).
242. Siddiqui, N. *et al.* Poly (A) Nuclease Interacts with the C-terminal Domain of Polyadenylate-binding Protein Domain from Poly (A) -binding Protein. *J. Biol. Chem.* **282**, 25067–25075 (2007).
243. Sachs, A. B. & Deardorff, J. A. Translation initiation requires the PAB-dependent poly(A) ribonuclease in yeast. *Cell* **70**, 961–973 (1992).
244. Huang, K. L., Chadee, A. B., Chen, C. Y. A., Zhang, Y. & Shyu, A. Bin. Phosphorylation at intrinsically disordered regions of PAM2 motif-containing proteins modulates their interactions with PABPC1 and influences mRNA fate. *RNA* **19**, 295–305 (2013).
245. Yi, H. *et al.* PABP Cooperates with the CCR4-NOT Complex to Promote mRNA Deadenylation and Block Precocious Decay. *Mol. Cell* **70**, 1081-1088.e5 (2018).
246. Brown, C. E. & Sachs, A. B. Poly(A) tail length control in Saccharomyces cerevisiae occurs by message-specific deadenylation. *Mol. Cell. Biol.* **18**, 6548–59 (1998).
247. Webster, M. W. *et al.* mRNA Deadenylation Is Coupled to Translation Rates by the Differential Activities of Ccr4-Not Nucleases. *Mol. Cell* **70**, 1089-1100.e8 (2018).
248. Singhania, R. *et al.* Nuclear poly(A) tail size is regulated by Cnot1 during the serum response. *bioRxiv* (2019) doi:10.1101/773432.
249. Tucker, M. *et al.* The transcription factor associated Ccr4 and Caf1 proteins are components of the major cytoplasmic mRNA deadenylase in Saccharomyces cerevisiae. *Cell* **104**, 377–386 (2001).
250. Yamashita, A. *et al.* Concerted action of poly(A) nucleases and decapping enzyme in mammalian mRNA turnover. *Nat. Struct. Mol. Biol.* **12**, 1054–1063 (2005).
251. Morris, J. Z., Hong, A., Lilly, M. A. & Lehmann, R. Twin, a CCR4 homolog, regulates cyclin poly(A) tail length to permit Drosophila oogenesis. *Development* **132**, 1165–1174 (2005).
252. Björklund, M. *et al.* Identification of pathways regulating cell size and cell-cycle progression by RNAi. *Nature* **439**, 1009–1013 (2006).
253. Washio-Oikawa, K. *et al.* Cnot7-null mice exhibit high bone mass phenotype and modulation of BMP actions. *J. Bone Miner. Res.* **22**, 1217–1223 (2007).
254. Goldstrohm, A. C. & Wickens, M. Multifunctional deadenylase complexes diversify mRNA control. *Nat. Rev. Mol. Cell Biol.* **9**, 337–344 (2008).
255. Körner, C. G. *et al.* The deadenylating nuclease (DAN) is involved in poly(A) tail removal during the meiotic maturation of Xenopus oocytes. *EMBO J.* **17**, 5427–5437 (1998).
256. Kim, J. H. & Richter, J. D. Opposing Polymerase-Deadenylase Activities Regulate Cytoplasmic Polyadenylation. *Mol. Cell* **174**, 173–183 (2006).
257. Moon, D. H. *et al.* Poly(A)-specific ribonuclease (PARN) mediates 3′-end maturation of the telomerase RNA component. *Nat. Genet.* **47**, 1482–1488 (2015).
258. Cevher, M. A. *et al.* Nuclear deadenylation/polyadenylation factors regulate 3′ processing in response to DNA damage. *EMBO J.* **29**, 1674–1687 (2010).
259. Lejeune, F., Li, X. & Maquat, L. E. Nonsense-mediated mRNA decay in mammalian cells involves decapping, deadenylating, and exonucleolytic activities. *Mol Cell* **12**, 675–687 (2003).
260. Kubota, K. *et al.* Identification of 2 ˜-Phosphodiesterase , Which Plays a Role in the 2 – 5A System Regulated by Interferon. *J. Biol. Chem.* **279**, 37832–37841 (2004).
261. Green, C. B. & Besharse, J. C. Identification of a novel vertebrate circadian clock-regulated gene encoding the protein nocturnin. *Proc. Natl. Acad. Sci.* **93**, 14884–14888 (1996).
262. Baggs, J. E. & Green, C. B. Nocturnin, a deadenylase in Xenopus laevis retina: A mechanism for posttranscriptional control of circadian-related mRNA. *Curr. Biol.* **13**, 189–198 (2003).
263. Green, C. B. *et al.* Loss of Nocturnin, a circadian deadenylase, confers resistance to hepatic steatosis and diet-induced obesity. *Proc. Natl. Acad. Sci.* **104**, 9888–9893 (2007).
264. Labno, A., Tomecki, R. & Dziembowski, A. Cytoplasmic RNA decay pathways - Enzymes and mechanisms. *Biochimica et Biophysica Acta* **1863**, 3125–3147 (2016).
265. Eisen, T. J. *et al.* The Dynamics of Cytoplasmic mRNA Metabolism. *Mol Cell* **77**, 1-14 (2020)
266. Sachs, A. B. & Davis, R. W. The poly(A) binding protein is required for poly(A) shortening and 60S ribosomal subunit-dependent translation initiation. *Cell* **58**, 857–867 (1989).
267. Caponigro, G. & Parker, R. Multiple functions for the poly(A)binding protein in mRNA decapping and deadenylation in yeast. *Genes Dev.* **9**, 2421–2432 (1995).

268. Rissland, O. S. *et al.* The influence of microRNAs and poly(A) tail length on endogenous mRNA-protein complexes. *Genome Biol.* **18**, 1–18 (2017).
269. Mauxion, F., Faux, C. & Séraphin, B. The BTG2 protein is a general activator of mRNA deadenylation. *EMBO J.* **27**, 1039–1048 (2008).
270. Webster, M. W., Stowell, J. A. & Passmore, L. A. RNA-binding proteins distinguish between similar sequence motifs to promote targeted deadenylation by Ccr4-Not. *Elife* **8**, (2019).
271. Lim, J. *et al.* Mixed tailing by TENT4A and TENT4B shields mRNA from rapid deadenylation. *Science* **361**, 701–704 (2018).
272. Tang, T. T. L., Stowell, J. A. W., Hill, C. H. & Passmore, L. A. The intrinsic structure of poly(A) RNA determines the specificity of Pan2 and Caf1 deadenylases. *Nat. Struct. Mol. Biol.* **26**, 433–442 (2019).
273. Wahle, E. Purification and characterization of a mammalian polyadenylate polymerase involved in the 3' end processing of messenger RNA precursors. *J. Biol. Chem.* **266**, 3131–3139 (1991).
274. Shen, B. & Goodman, H. M. Uridine addition after microRNA-directed cleavage. *Science* **306**, 997 (2004).
275. Trippe, R., Sandrock, B. & Benecke, B. J. A highly specific terminal uridylyl transferase modifies the 3'-end of U6 small nuclear RNA. *Nucleic Acids Res.* **26**, 3119–3126 (1998).
276. Rissland, O. S. & Norbury, C. J. Decapping is preceded by 3′ uridylation in a novel pathway of bulk mRNA turnover. *Nat. Struct. Mol. Biol.* **16**, 616–623 (2009).
277. Lim, J. *et al.* Uridylation by TUT4 and TUT7 marks mRNA for degradation. *Cell* **159**, 1365–76 (2014).
278. Thomas, M. P. *et al.* Apoptosis Triggers Specific, Rapid, and Global mRNA Decay with 3' Uridylated Intermediates Degraded by DIS3L2. *Cell Rep.* **11**, 1079–1089 (2015).
279. Lubas, M. *et al.* Exonuclease hDIS3L2 specifies an exosome-independent 3′-5′ degradation pathway of human cytoplasmic mRNA. *EMBO J.* **32**, 1855–1868 (2013).
280. Morgan, M. *et al.* mRNA 3′ uridylation and poly(A) tail length sculpt the mammalian maternal transcriptome. *Nature* **548**, 347–351 (2017).
281. Mullen, T. E. & Marzluff, W. F. Degradation of histone mRNA requires oligouridylation followed by decapping and simultaneous degradation of the mRNA both 5′ to 3′ and 3′ to 5′. *Genes Dev.* **22**, 50–65 (2008).
282. Hoefig, K. P. *et al.* Eri1 degrades the stem-loop of oligouridylated histone mRNAs to induce replication-dependent decay. *Nat. Struct. Mol. Biol.* **20**, 73–81 (2013).
283. Liu, Y., Nie, H., Liu, H. & Lu, F. Poly(A) inclusive RNA isoform sequencing (PAIso−seq) reveals wide-spread non-adenosine residues within RNA poly(A) tails. *Nat. Commun.* **10**, (2019).
284. Hubstenberger, A. *et al.* P-Body Purification Reveals the Condensation of Repressed mRNA Regulons. *Mol. Cell* **68**, 144-157 (2017).
285. Kamenska, A. *et al.* The DDX6-4E-T interaction mediates translational repression and P-body assembly. *Nucleic Acids Res.* **44**, 6318–6334 (2016).
286. Wilczynska, A., Aigueperse, C., Kress, M., Dautry, F. & Weil, D. The translational regulator CPEB1 provides a link between dcp1 bodies and stress granules. *J. Cell Sci.* **118**, 981–992 (2005).
287. Teixeira, D., Sheth, U., Valencia-Sanchez, M. A., Brengues, M. & Parker, R. Processing bodies require RNA for assembly and contain nontranslating mRNAs. *RNA* **11**, 371–382 (2005).
288. Eulalio, A., Behm-Ansmant, I., Schweizer, D. & Izaurralde, E. P-Body Formation Is a Consequence, Not the Cause, of RNA-Mediated Gene Silencing. *Mol. Cell. Biol.* **27**, 3970–3981 (2007).
289. Wang, R., Nambiar, R., Zheng, D. & Tian, B. PolyA-DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Res.* **46**, D315–D319 (2018).
290. Herrmann, C. J. *et al.* PolyASite 2.0 : a consolidated atlas of polyadenylation sites from 3 end sequencing. *Nucleic Acids Research* **48**, 174–179 (2020).
291. Tian, B. & Manley, J. L. Alternative Polyadenlyation of mRNA Precursors. *Nat Rev Mol Cell Biol.* **165**, 255–269 (2017).
292. He, S. L. & Green, R. Northern blotting. *Methods in Enzymology* vol. 530 (2013).
293. Frohman, M. A., Dush, M. K. & Martin, G. R. Rapdi production of full-length cDNAs from rare transcripts: Amplification using a single gene-specific oligonucleotide primer. *Proc. Nati. Acad. Sci.* **85**, 8998–9002 (1988).
294. Erben, L., He, M., Laeremans, A., Park, E. & Buonanno, A. A Novel Ultrasensitive In Situ Hybridization Approach to Detect Short Sequences and Splice Variants with Cellular Resolution. *Molecular Neurobiology* **55,** 6169–6181 (2018).
295. Sallés, F. J. & Strickland, S. Rapid and sensitive analysis of mRNA polyadenylation states by PCR. *Genome Res.* **4**, 317–321 (1995).

296. Bazzini, A. A., Lee, M. T. & Giraldez, A. J. Ribosome Profiling Shows That miR-430 Reduces Translation Before Causing mRNA Decay in Zebrafish. *Science.* **336**, 233–237 (2012).
297. Kusov, Y. Y., Shatirishvili, G., Dzagurov, G. & Gauss-Müller, V. A new G-tailing method for the determination of the poly(A) tail length applied to hepatitis A virus RNA. *Nucleic Acids Res.* **29**, 10–15 (2001).
298. Binder, R. *et al.* Evidence that the pathway of transferrin receptor mRNA degradation involves an endonucleolytic cleavage within the 3' UTR and does not involve poly(A) tail shortening. *EMBO J.* **13**, 1969–1980 (1994).
299. Beilharz, T. H. & Preiss, T. Widespread use of poly(A) tail length control to accentuate expression of the yeast transcriptome. *RNA* **13**, 982–997 (2007).
300. Metzker, M. L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).
301. Illumina. NovaSeq 6000 Specification. https://www.illumina.com/systems/sequencing-platforms/novaseq/specifications.html.
302. Cacho, A., Smirnova, E., Huzurbazar, S. & Cui, X. A comparison of base-calling algorithms for illumina sequencing technology. *Brief. Bioinform.* **17**, 786–795 (2016).
303. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
304. Adiconis, X. *et al.* Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat. Methods* **10**, 623–629 (2013).
305. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
306. Smith, T. & Waterman, S. Identification of Common Molecular Subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).
307. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
308. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, (2009).
309. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
310. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics Bioinforma.* **3**, 1–9 (2021).
311. Altshuler, D. L. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
312. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science.* **372**, (2021).
313. Ziemann, M., Kaspi, A. & El-Osta, A. Evaluation of microRNA alignment techniques. *RNA* **22**, 1120–1138 (2016).
314. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 1–19 (2016).
315. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
316. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2009).
317. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).
318. Zhang, C., Zhang, B., Lin, L. L. & Zhao, S. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics* **18**, 1–11 (2017).
319. Hoque, M. *et al.* Analysis of alternative cleavage and polyadenylation by 3′ region extraction and deep sequencing. *Nat. Methods* **10**, 133–139 (2013).
320. Lim, J., Lee, M., Son, A., Chang, H. & Kim, V. N. MTAIL-seq reveals dynamic poly(A) tail regulation in oocyte-to-embryo development. *Genes Dev.* **30**, 1671–1682 (2016).
321. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Scienc.* **323**, 133–138 (2009).
322. Korlach, J. Understanding Accuracy in SMRT® Sequencing. 1–9 (2015). https://www.pacb.com/wp-content/uploads/2015/09/Perspective_UnderstandingAccuracySMRTSequencing1.pdf
323. SMRT Sequencing. https://www.pacb.com/smrt-science/smrt-sequencing/.
324. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics, Proteomics Bioinforma.* **13**, 278–289 (2015).

325. Zhang, X. *et al.* Improving genome assemblies by sequencing PCR products with PacBio. *Biotechniques* **53**, 61–62 (2012).
326. He, Y. *et al.* Long-read assembly of the Chinese rhesus macaque genome and identification of ape-specific structural variants. *Nat. Commun.* **10**, 1–14 (2019).
327. Patel, A., Schwab, R., Liu, Y. T. & Bafna, V. Amplification and thrifty single-molecule sequencing of recurrent somatic structural variations. *Genome Res.* **24**, 318–328 (2014).
328. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* **31**, 1009–1014 (2013).
329. SMRT RNA Fractionation Experiments. https://www.pacb.com/wp-content/uploads/2015/09/User-Bulletin-Guidelines-for-Preparing-cDNA-Libraries-for-Isoform-Sequencing-Iso-Seq.pdf.
330. Kono, N. & Arakawa, K. Nanopore sequencing: Review of potential applications in functional genomics. *Dev. Growth Differ.* **61**, 316–326 (2019).
331. Payne, A., Holmes, N., Rakyan, V. & Loose, M. Bulkvis: A graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics* **35**, 2193–2198 (2019).
332. Nanopore Sequencing High Molecular Weight. https://www.protocols.io/view/ultra-long-read-sequencing-protocol-for-rad004-mrxc57n?step=7.
333. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–232 (2016).
334. Rang, F. J., Kloosterman, W. P. & de Ridder, J. From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* **19**, 1–11 (2018).
335. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**, 201–206 (2018).
336. Roundtree, I. A., Evans, M. E., Pan, T. & He, C. Dynamic RNA Modifications in Gene Expression Regulation. *Cell* **169**, 1187–1200 (2017).
337. Maier, K. C., Gressel, S., Cramer, P. & Schwalb, B. Native molecule sequencing by nano-ID reveals synthesis and stability of RNA isoforms. *Genome Res.* **30**, 1332–1344 (2020).
338. Byrne, A. *et al.* Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* **8**, 1–11 (2017).
339. Tang, A. D. *et al.* Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat. Commun.* **11**, 1–12 (2020).
340. Workman, R. E. *et al.* Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* **16**, 1297–1305 (2019).
341. Herzel, L., Straube, K. & Neugebauer, K. M. Long-read sequencing of nascent RNA reveals coupling among RNA processing events. *Genome Res.* **28**, 1008–1019 (2018).
342. Collart, M. A. The Ccr4-Not complex is a key regulator of eukaryotic gene expression. *Wiley Interdiscip. Rev. RNA* **7**, 438–454 (2016).
343. Weill, L., Belloc, E., Bava, F. A. & Méndez, R. Translational control by changes in poly(A) tail length: Recycling mRNAs. *Nat. Struct. Mol. Biol.* **19**, 577–585 (2012).
344. Schmid, M. & Jensen, T. H. Controlling nuclear RNA levels. *Nat. Rev. Genet.* **19**, 518–529 (2018).
345. Chomczynski, P. & Sacchi, N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem.* **162**, 156–159 (1987).
346. Ampure Size Selection. https://www.beckman.de/reagents/genomic/cleanup-and-size-selection/size-selection/performance (2021).
347. QuBit 3 Specification. (2021).
348. Laemmli, U. K. Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4. *Nature* **227**, 680–685 (1970).
349. Smargon, A. A. *et al.* Cas13b Is a Type VI-B CRISPR-Associated RNA-Guided RNase Differentially Regulated by Accessory Proteins Csx27 and Csx28. *Mol. Cell* **65**, 618–630 (2017).
350. Konermann, S. Transcriptome engineering with RNA-targeting Type VI-D CRISPR effectors. *Cell* **173(3)**, 665–676 (2019).
351. Lizio, M. *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.* **16**, 1–14 (2015).
352. Saito, T. L. *et al.* The transcription start site landscape of C. elegans. *Genome Res.* **23**, 1348–1361 (2013).
353. Alexa, A. & Rahnenführer, J. Gene set enrichment analysis with topGO. (2013). http://compdiag.molgen.mpg.de/ngfn/docs/2007/sep/topGO_Exercises.pdf

354. Schueler, M. *et al.* Differential protein occupancy profiling of the mRNA transcriptome. *Genome Biol.* **15**, 1–17 (2014).
355. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
356. Jürges, C., Dölken, L. & Erhard, F. Dissecting newly transcribed and old RNA using GRAND-SLAM. *Bioinformatics* **34**, i218–i226 (2018).
357. Durinck, S. *et al.* BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).
358. Legnini, I., Alles, J., Karaiskos, N., Ayoub, S. & Rajewsky, N. FLAM-seq: full-length mRNA sequencing reveals principles of poly(A) tail length control. *Nat. Methods* **16**, 879–886 (2019).
359. Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R. & Siebert, P. D. Reverse transcriptase template switching: A SMART approach for full-length cDNA library construction. *Biotechniques* **30**, 892–897 (2001).
360. Kapteyn, J., He, R., McDowell, E. T. & Gang, D. R. Incorporation of non-natural nucleotides into template-switching oligonucleotides reduces background and improves cDNA synthesis from very small RNA samples. *BMC Genomics* **11**, (2010).
361. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
362. Nam, D. K. *et al.* Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *PNAS* **99**, 6152–6156 (2002).
363. Balázs, Z. *et al.* Template-switching artifacts resemble alternative polyadenylation. *BMC Genomics* **20**, 1–10 (2019).
364. Novoradovskaya, N. *et al.* Universal Reference RNA as a standard for microarray experiments. *BMC Genomics* **5**:20 1–13 (2004)
365. Roach, N. P. *et al.* The full-length transcriptome of C. elegans using direct RNA sequencing. *bioRxiv* 299–312 (2019) doi:10.1101/598763.
366. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci.***100**, 15776–15781 (2003).
367. Allen, M. A., Hillier, L. D. W., Waterston, R. H. & Blumenthal, T. A global analysis of C. elegans trans-splicing. *Genome Res.* **21**, 255–264 (2011).
368. Krause, M. & Hirsh, D. A trans-spliced leader sequence on actin mRNA in C. elegans. *Cell* **49**, 753–761 (1987).
369. Li, X. Q. & Du, D. RNA polyadenylation sites on the genomes of microorganisms, animals, and plants. *PLoS One* **8**, (2013).
370. Hu, J., Lutz, C. S., Wilusz, J. & Tian, B. Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. *RNA* **11**, 1485–1493 (2005).
371. Yu, S. & Kim, V. N. A tale of non-canonical tails: gene regulation by post-transcriptional RNA tailing. *Nat. Rev. Mol. Cell Biol.* **21**, 542–556 (2020).
372. Sheiness, D., Puckett, L. & Darnell, J. E. Possible relationship of poly(A) shortening to mRNA turnover. *Proc. Natl. Acad. Sci.* **72**, 1077–1081 (1975).
373. Brawerman, G. & Diez, J. Metabolism of the polyadenylate sequence of nuclear RNA and messenger RNA in mammalian cells. *Cell* **5**, 271–280 (1975).
374. Wachutka, L., Caizzi, L., Gagneur, J. & Cramer, P. Global donor and acceptor splicing site kinetics in human cells. *Elife* **8**, 1–52 (2019).
375. Dölken, L. *et al.* High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA* **14**, 1959–1972 (2008).
376. Mor, A. *et al.* Dynamics of single mRNP nucleocytoplasmic transport and export through the nuclear pore in living cells. *Nat. Cell Biol.* **12**, 543–552 (2010).
377. Chen, T. & van Steensel, B. Comprehensive analysis of nucleocytoplasmic dynamics of mRNA in Drosophila cells. *PLoS Genet.* **13**, 1–25 (2017).
378. Jadot, M. *et al.* Accounting for protein subcellular localization: A compartmental map of the rat liver proteome. *Mol. Cell. Proteomics* **16**, 194–212 (2017).
379. Namba, T. BAP31 regulates mitochondrial function via interaction with Tom40 within ER-mitochondria contact sites. *Sci. Adv.* **5**, 1–13 (2019).
380. Wakana, Y. *et al.* Bap31 Is an Itinerant Protein That Moves between the Peripheral Endoplasmic Reticulum (ER) and a Juxtanuclear Compartment Related to ER-associated Degradation. *Mol. Biol. Cell* **18**, 3250–3263 (2007).
381. Ayala, Y. M. *et al.* Structural determinants of the cellular localization and shuttling of TDP-43. *J. Cell Sci.* **121**, 3778–3785 (2008).

382. Lubelsky, Y. & Ulitsky, I. Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. *Nature* **555**, 107–111 (2018).
383. Cas13b is a Type VI-B CRISPR-associated RNA-Guided RNAse differentially regulated by accessory proteins Csx27 and Csx28. *Mol Cell* **65**, 618–630 (2017).
384. Cox, D. B. T. *et al.* RNA editing with CRISPR-Cas13. *Science* **358**, 1019–1027 (2017).
385. Geiss, G. K. *et al.* Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat. Biotechnol.* **26**, 317–325 (2008).
386. Geisberg, J. V., Moqtaderi, Z., Fan, X., Ozsolak, F. & Struhl, K. Global analysis of mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast. *Cell* **156**, 812–824 (2014).
387. Taliaferro, J. M. *et al.* Distal Alternative Last Exons Localize mRNAs to Neural Projections. *Mol. Cell* **61**, 821–833 (2016).
388. Volden, R. *et al.* Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 9726–9731 (2018).
389. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–8 (2013).
390. Liu, Y. *et al.* Multi-omic measurements of heterogeneity in HeLa cells across laboratories. *Nat. Biotechnol.* **37**, 314–322 (2019).
391. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).
392. Baer, B. W. & Kornberg, R. D. The protein responsible for the repeating structure of cytoplasmic poly(A)-ribonucleoprotein. *J. Cell Biol.* **96**, 717–721 (1983).
393. Lawson, N. D. *et al.* An improved zebrafish transcriptome annotation for sensitive and comprehensive detection of cell type-specific genes. *Elife* **9**, 1–76 (2020).
394. Mangone, M. *et al.* The landscape of C. elegans 3′UTRs. *Science* **329**, 432–435 (2010).
395. Shen, Y. *et al.* Genome level analysis of rice mRNA 3 ’-end processing signals and alternative polyadenylation. **36**, 3150–3161 (2008).
396. Jan, C. H., Friedman, R. C., Ruby, J. G. & Bartel, D. P. Formation, regulation and evolution of Caenorhabditis elegans 3’UTRs. *Nature* **469**,97-101 (2011)
397. Steber, H., Gallante, C., Brien, O. S., Chiu, P. & Mangone, M. The C . elegans 3'UTRome V2 : an updated genomic resource to study 3’UTR biology. *bioRxiv* (2019).
398. Diag, A., Schilling, M., Klironomos, F., Ayoub, S. & Rajewsky, N. Regulation in the C . elegans Germline Resource Spatiotemporal m(i)RNA Architecture and 3'UTR Regulation in the C . elegans Germline. *Dev. Cell* **47**, 785-80 (2018).
399. Preston, M. A. *et al.* Unbiased screen of RNA tailing activities reveals a poly(UG) polymerase. *Nat. Methods* **16,** 437-445, (2019)
400. Dvinge, H. & Bradley, R. K. Widespread intron retention diversifies most cancer transcriptomes. *Genome Med.* **7**, 1–13 (2015).
401. García-Martínez, J. *et al.* The cellular growth rate controls overall mRNA turnover, and modulates either transcription or degradation rates of particular gene regulons. *Nucleic Acids Res.* **44**, 3643–3658 (2016).
402. Hilgers, V., Teixeira, D. & Parker, R. Translation-independent inhibition of mRNA deadenylation during stress in Saccharomyces cerevisiae. *RNA* **12**, 1835–1845 (2006).
403. Du, L. & Richter, J. D. Activity-dependent polyadenylation in neurons. *RNA* **11**, 1340–1347 (2005).
404. Winkler, G. S., Mulder, K. W., Bardwell, V. J., Kalkhoven, E. & Timmers, H. T. M. Human Ccr4-Not complex is a ligand-dependent repressor of nuclear receptor-mediated transcription. *EMBO J.* **25**, 3089–3099 (2006).
405. Cooke, A., Prigge, A. & Wickens, M. Translational repression by deadenylases. *J. Biol. Chem.* **285**, 28506–28513 (2010).
406. Brown, C. E., Tarun, S. Z., Boeck, R. & Sachs, A. B. PAN3 encodes a subunit of the Pab1p-dependent poly(A) nuclease in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **16**, 5744–5753 (1996).
407. Yesbolatova, A. *et al.* The auxin-inducible degron 2 technology provides sharp degradation control in yeast, mammalian cells, and mice. *Nat. Commun.* **11**, (2020).
408. Webster, M. W., Stowell, J. A. & Passmore, L. A. RNA-binding proteins distinguish between similar sequence motifs to promote targeted deadenylation by Ccr4-Not. *Elife* **8**, (2019).