The spirit of the woods
Sandro Del Prete, 1981

Sequencing Methods and Systems Virology

Lars Dölken

Systems Virology / Systems Biology

What is systems virology?

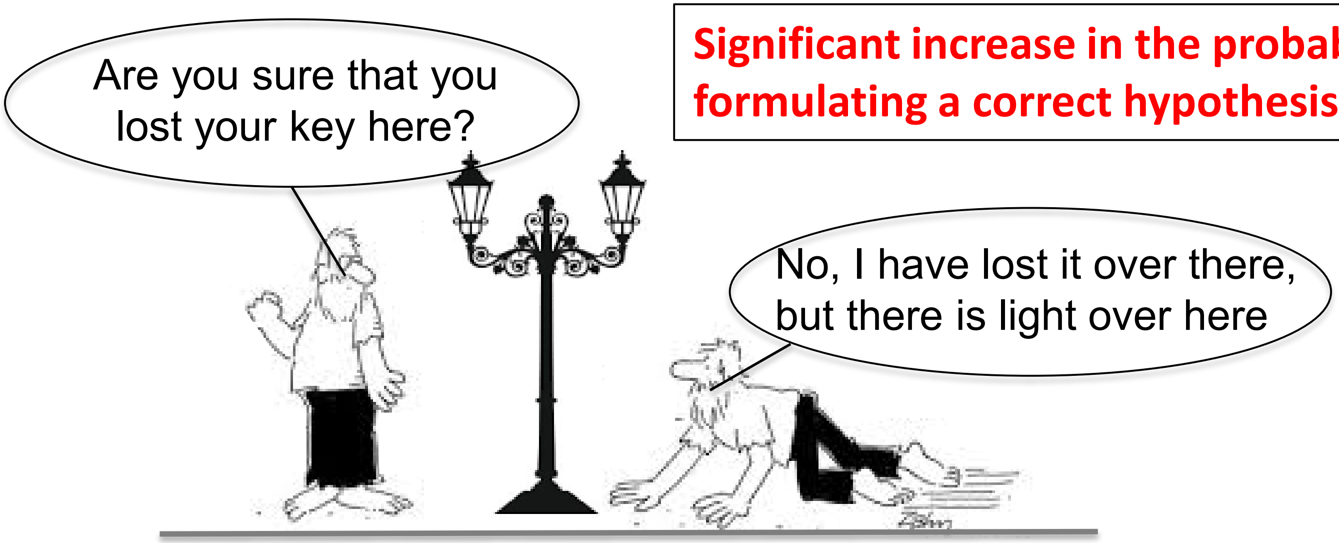
Wiki: ?

What is systems virology?

Wiki: = The attempt to understand virus – host interactions in their full complexity

= Opportunity to look outside the box (unknown unknown)

Significant increase in the probability of formulating a correct hypothesis


Are you sure that you lost your key here?
No, I have lost it over there, but there is light over here

Content

1. Overview on sequencing technologies
2. Metabolic labelling of newly transcribed RNA
3. Ribosome Profiling
4. Single cell sequencing (scRNA-seq) and its combination with metabolic labeling (scSLAM-seq)

Sequencing approaches

1st Generation ("Chain-termination" sequencing)

⇒ Sequencing by electrophoretical separation of amplified DNA

e.g.: Didesoxy method by Sanger

2nd Generation („Shotgun" sequencing)

⇒ Sequencing of millions of small DNA fragments

⇒ Monitoring DNA Polymerase/Ligase in action

e.g.: Illumina™ Sequencing (50-300nt)

3rd Generation („Single-molecule" sequencing)

⇒ Direct observation of single molecule synthesis

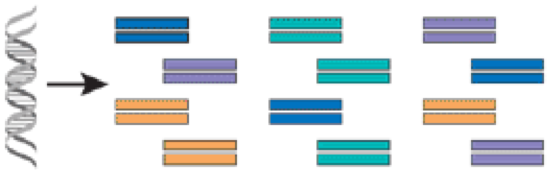
⇒ Very long sequences (>10kb)

e.g.: via changes in membrane potentials (Oxford Nanopores)
Fluorescence based (Pacific Biosciences)

Principle of 1st generation sequencing (Sanger)

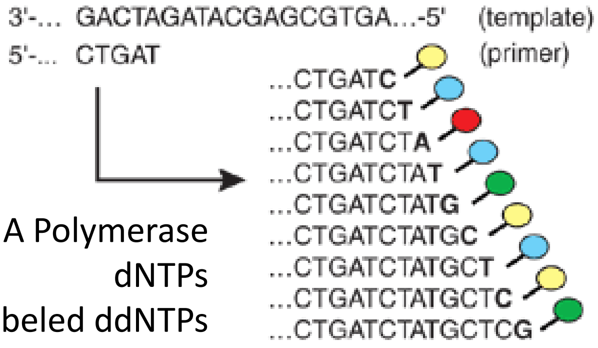
DNA Fragmentation

dsDNA of ≈1000bp



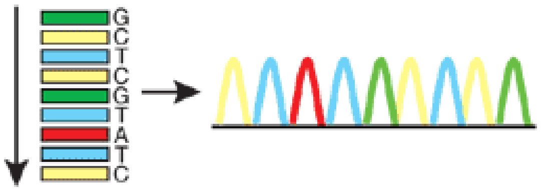
Amplification (Subcloning or PCR)

Sequencing in cycles

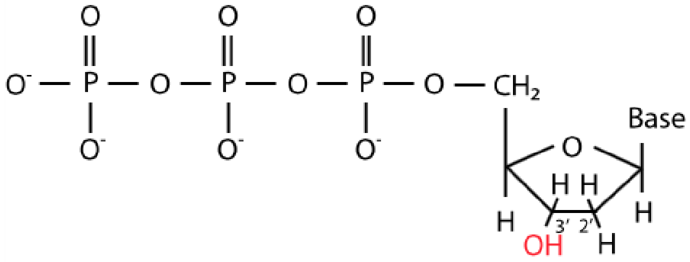


DNA Polymerase
dNTPs
Fluorophor-labeled ddNTPs

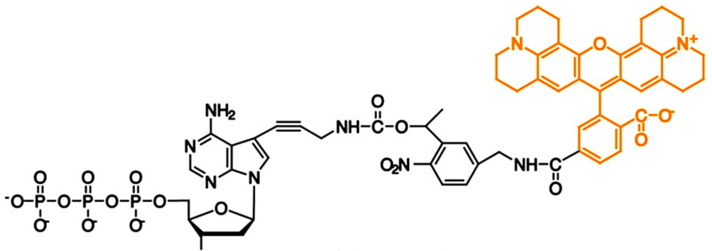
Electrophoretical size separation



Desoxynucleotide (dNTP)



Didesoxynucleotide (ddNTP)



dATP-PC-ROX

⇒ Chain termination

Principle of 2nd (next-) generation sequencing (NGS)

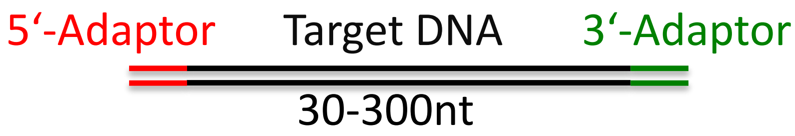
e.g. llumina Sequencing

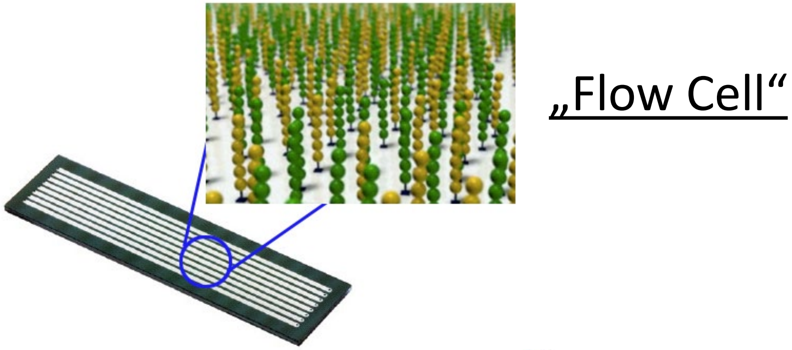
1. **Generation of cDNA libraries (>10⁸)**

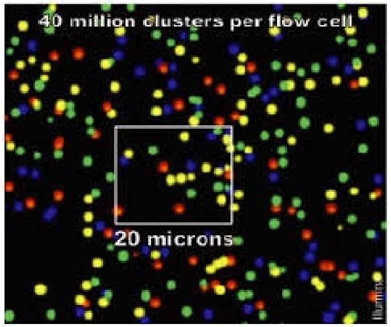
2. **Generation of clusters by DNA amplification**

3. **Sequencing by real-time monitoring of DNA-synthesis**

https://www.youtube.com/watch?v=womKfikWlxM







>100 millionen clusters (=Reads) of 35-300nt

⇒ Bioinformatics (mapping, assembly, quantification)

Illumina Sequencing

https://www.youtube.com/watch?v=womKfikWlxM

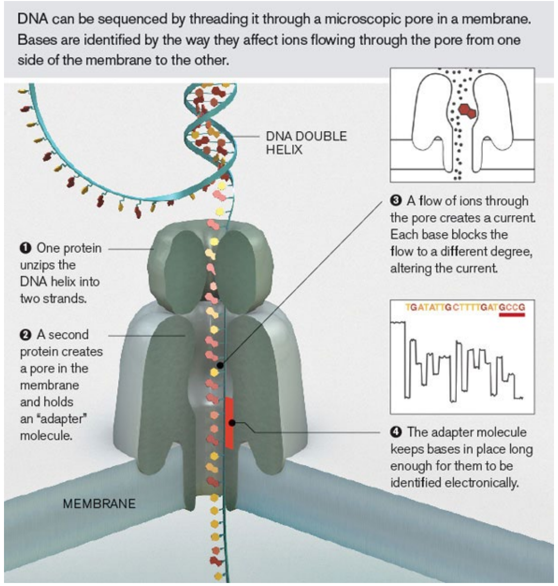
Principle of 3rd generation sequencing

= Single-molecule sequencing in real-time

⇒ Long DNA reads (>10kb)

Nanopore Sequencing (Oxford Nanopores)

https://www.youtube.com/watch?v=3UHw22hBpAk



DNA can be sequenced by threading it through a microscopic pore in a membrane. Bases are identified by the way they affect ions flowing through the pore from one side of the membrane to the other.

❶ One protein unzips the DNA helix into two strands.

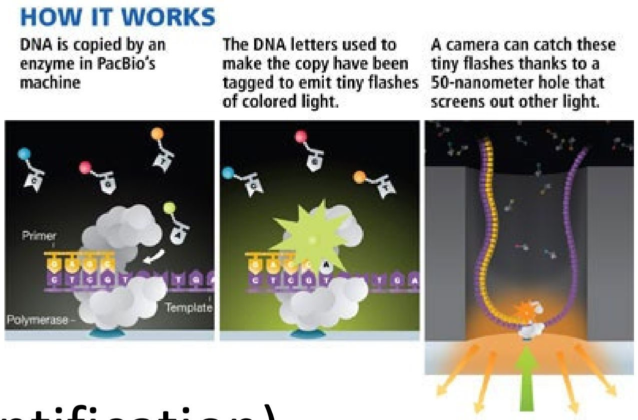
❷ A second protein creates a pore in the membrane and holds an "adapter" molecule.

❸ A flow of ions through the pore creates a current. Each base blocks the flow to a different degree, altering the current.

❹ The adapter molecule keeps bases in place long enough for them to be identified electronically.

Single molecule sequencing (Pacific Biosciences)

https://www.youtube.com/watch?v=v8p4ph2MAvI



HOW IT WORKS

DNA is copied by an enzyme in PacBio's machine

The DNA letters used to make the copy have been tagged to emit tiny flashes of colored light.

A camera can catch these tiny flashes thanks to a 50-nanometer hole that screens out other light.

⇒ Bioinformatics (mapping, assembly, quantification)

Oxford Nanopore Sequencing

https://www.youtube.com/watch?v=3UHw22hBpAk

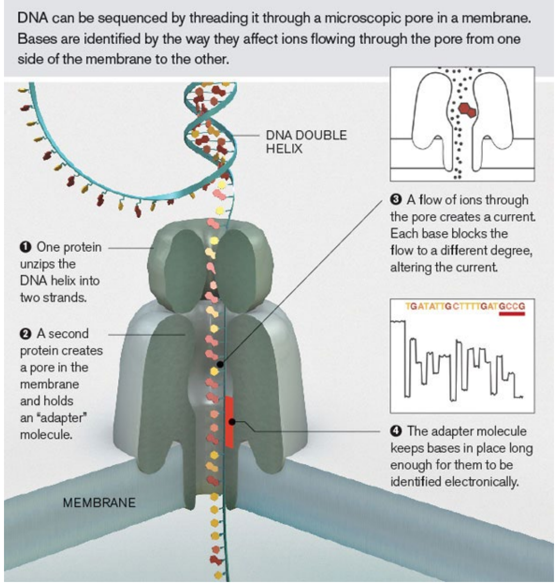
Principle of 3rd generation sequencing

= Single-molecule sequencing in real-time

⇒ Long DNA reads (>10kb)

Nanopore Sequencing (Oxford Nanopores)

https://www.youtube.com/watch?v=3UHw22hBpAk



DNA can be sequenced by threading it through a microscopic pore in a membrane. Bases are identified by the way they affect ions flowing through the pore from one side of the membrane to the other.

❶ One protein unzips the DNA helix into two strands.

❷ A second protein creates a pore in the membrane and holds an "adapter" molecule.

❸ A flow of ions through the pore creates a current. Each base blocks the flow to a different degree, altering the current.

❹ The adapter molecule keeps bases in place long enough for them to be identified electronically.

Single molecule sequencing (Pacific Biosciences)

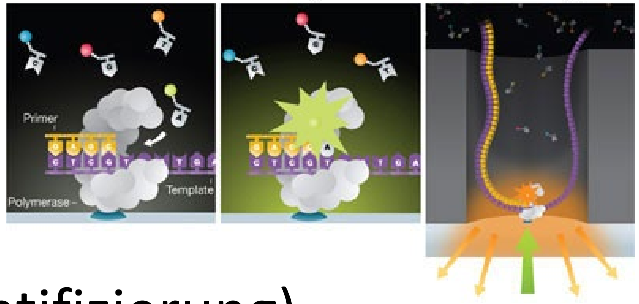
https://www.youtube.com/watch?v=v8p4ph2MAvI

HOW IT WORKS

DNA is copied by an enzyme in PacBio's machine

The DNA letters used to make the copy have been tagged to emit tiny flashes of colored light.

A camera can catch these tiny flashes thanks to a 50-nanometer hole that screens out other light.



⇒ Bioinformatics (mapping, assembly, quantifizierung)

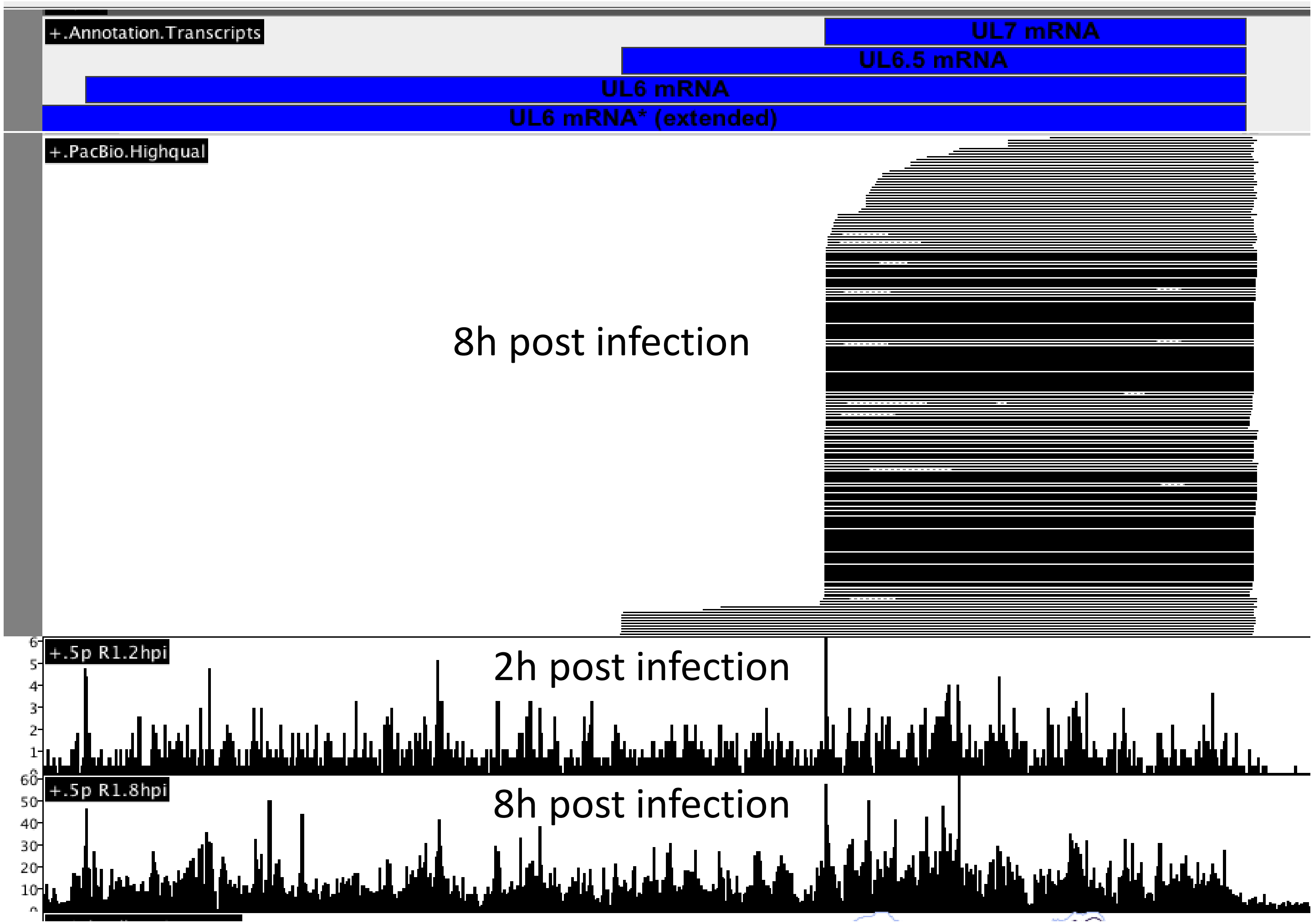
PacBio single-molecule sequencing

https://www.youtube.com/watch?v=v8p4ph2MAvI

Advantages and disadvantages of 2nd and 3rd generation sequencing

Approach	Read length	Advantage	Disadvantage
2nd generation (Illumina)	35 - 150nt	Ultra high throughput (>1 billion reads in 24h) Very low error rates (≈1:1000 nt) Main read-out for numerous pre-processing approaches, e.g. Ribo-seq, ChIP-seq, PAR-CLIP	Problems with repeat regions Assembly of full genomes without gaps is impossible Transcript isoforms not differentiated
3rd generation (PacBio)	up to >10,000	Identification of full length transcripts including alternative - transcription start sites - splicing isoforms - poly(A) sites No problem with repetitive regions allowing the correct assembly of complete genomes	High error rates of up to 10% Relative low throughput at present 10^4-10^5 reads

Mapping of HSV-1 transcripts by 2nd and 3rd generation sequencing



8h post infection

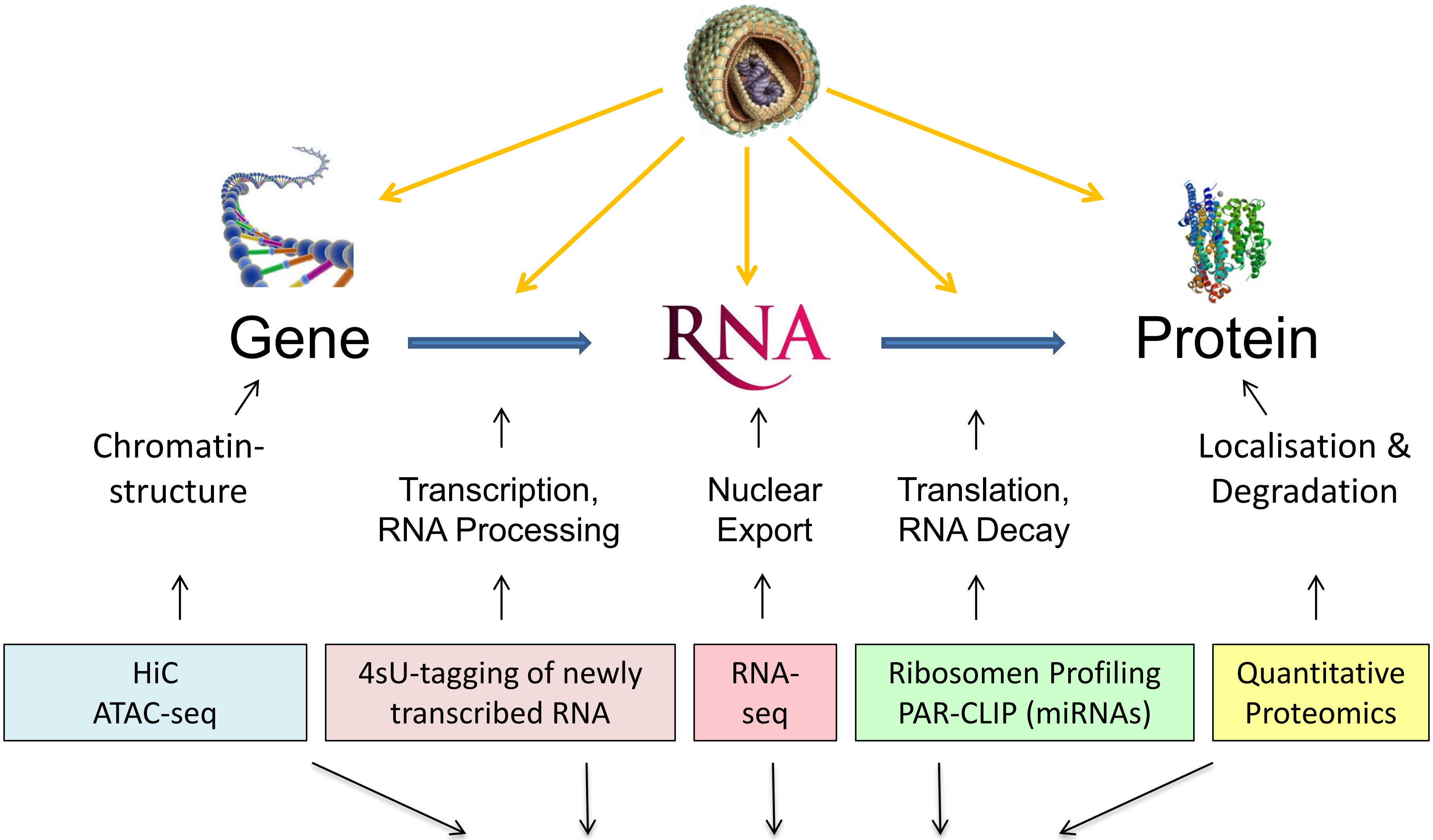
2h post infection

8h post infection

PacBio
3rd Generation
sequencing data

Illumina
2nd generation
Sequencing data

Regulation of cellular gene expression



Gene → RNA → Protein

Chromatin-structure

Transcription, RNA Processing

Nuclear Export

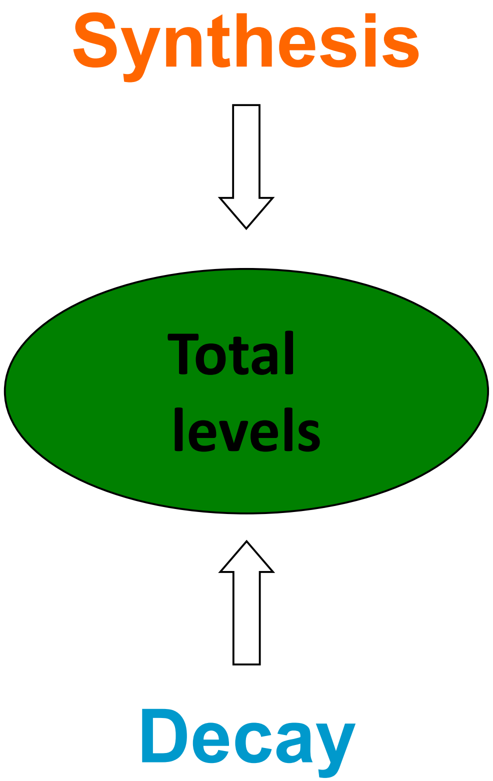
Translation, RNA Decay

Localisation & Degradation

HiC ATAC-seq	4sU-tagging of newly transcribed RNA	RNA-seq	Ribosomen Profiling PAR-CLIP (miRNAs)	Quantitative Proteomics

Activity, modulation and relevance of cellular pathways ?

Problems of standard gene expression profiling (RNA and proteins)



Δ total levels ≠ Δ synthesis rates

primary ⇏ secondary effects

low temporal resolution

Decay measurements

= imprecise & invasive

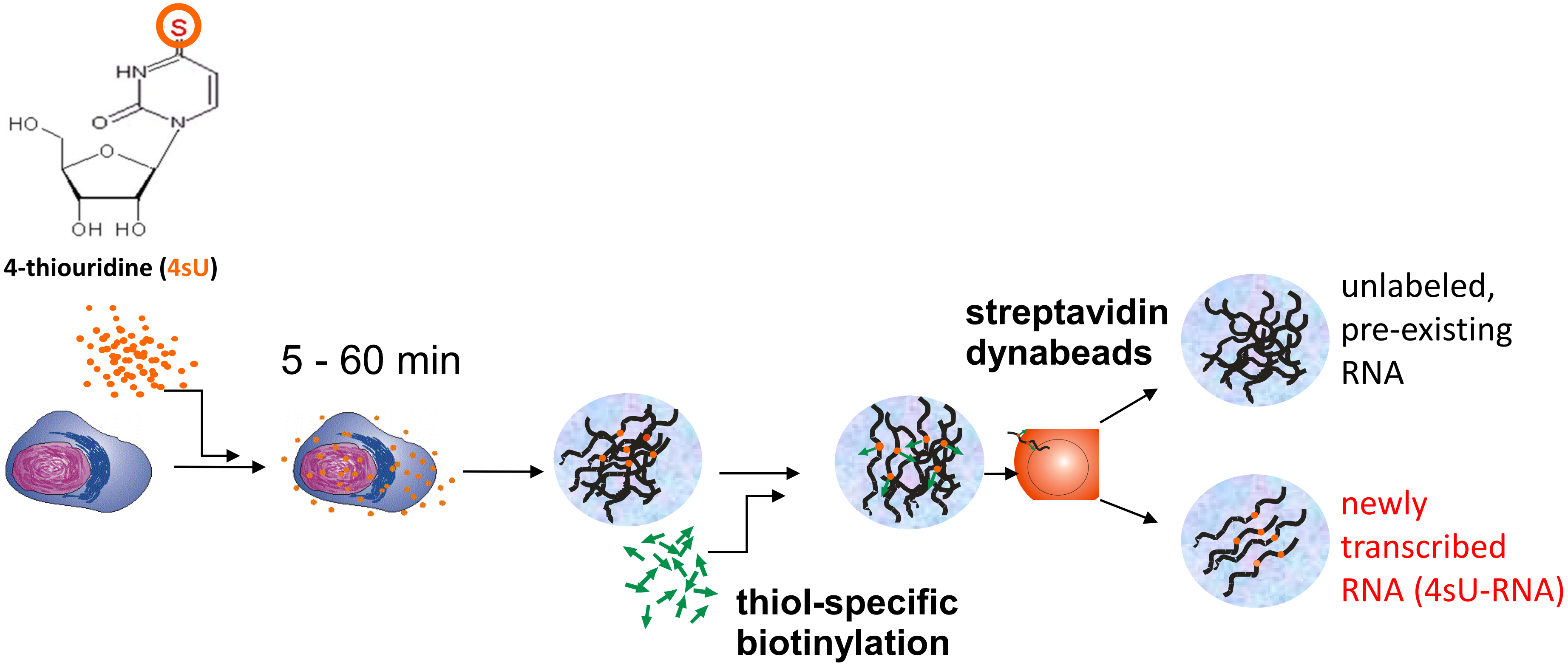
medium mRNA half-life (mammals): 5-10 h

(Yang et al., Genome Res 2003)
(Dölken et al., NAR 2009)

medium protein half-life (mammals): >20 h

(Schwanhäusser et al., Nature 2011)

Metabolic labeling and purification of newly transcribed RNA by 4sU-tagging



4-thiouridine (4sU)

5 - 60 min

thiol-specific biotinylation

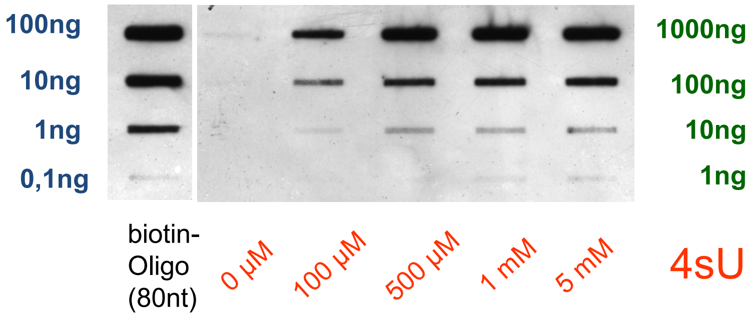
streptavidin dynabeads

unlabeled,
pre-existing
RNA

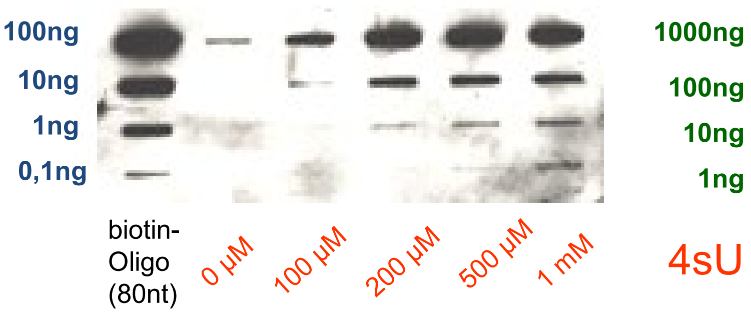
newly
transcribed
RNA (4sU-RNA)

Monitoring 4sU-incorporation into newly transcribed RNA

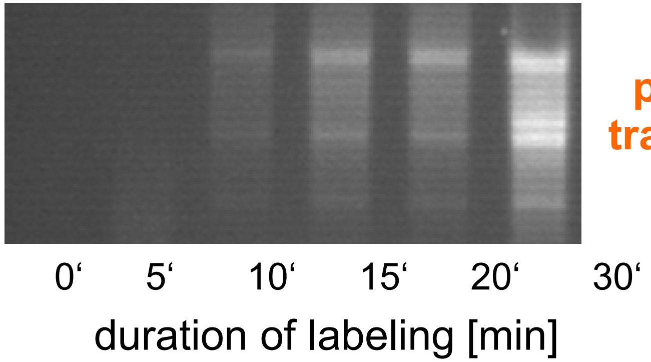
murine fibroblasts (NIH-3T3)



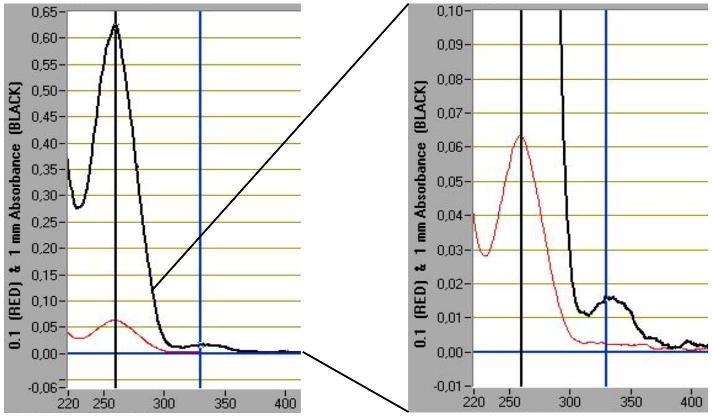
human B-cells (DG75)



input = 50 µg total RNA

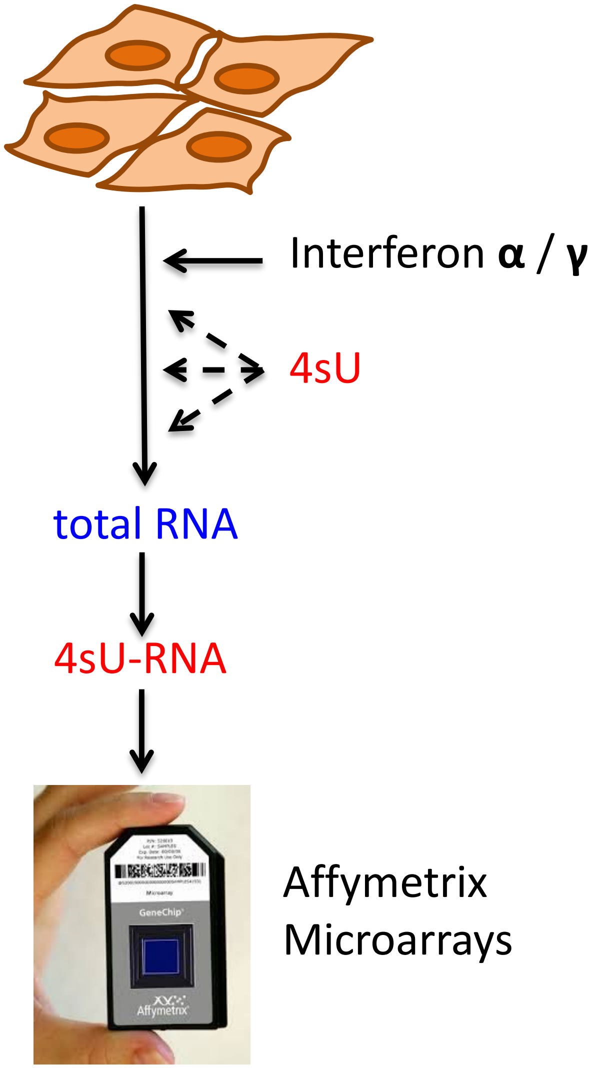


purified newly transcribed RNA



4sU incorporation: 1 : 50 nt

Validation of 4sU-tagging by analyzing the interferon response of murine fibroblasts



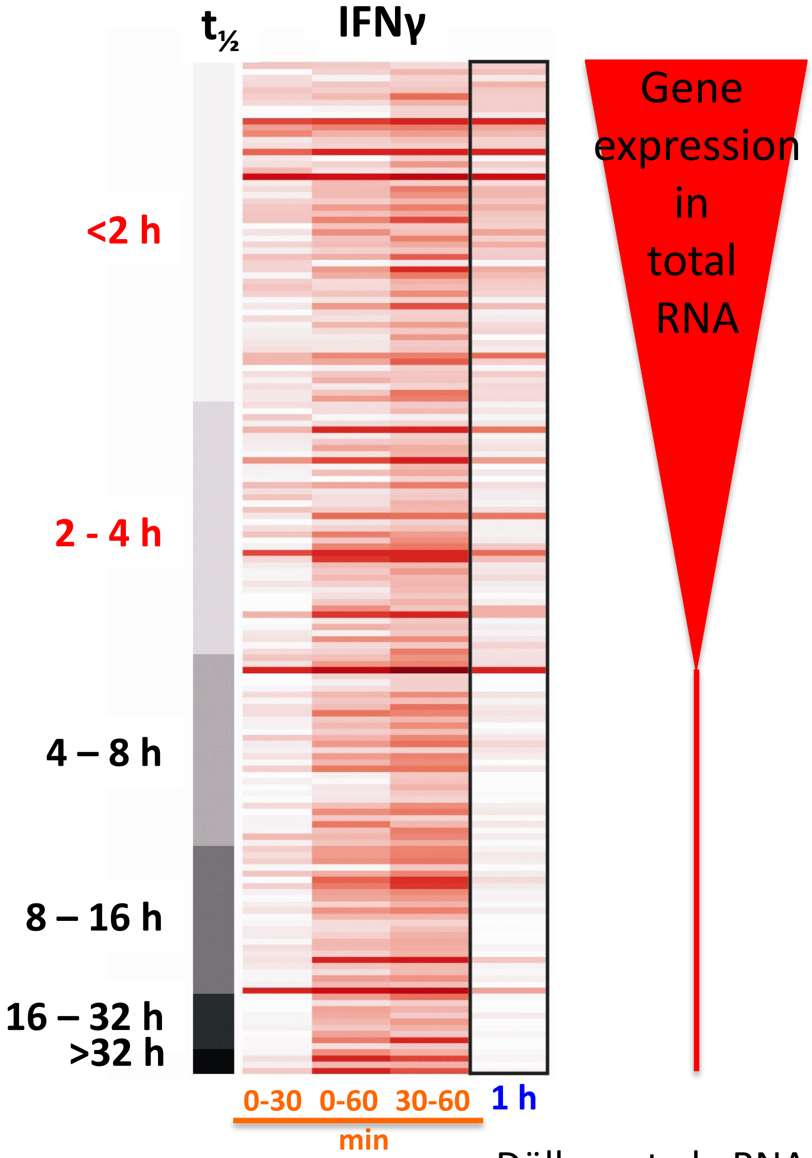
Interferon α / γ

4sU

total RNA

4sU-RNA

Affymetrix Microarrays



Dölken et al., RNA 2008

Validation of 4sU-tagging by analyzing the interferon response of murine fibroblasts

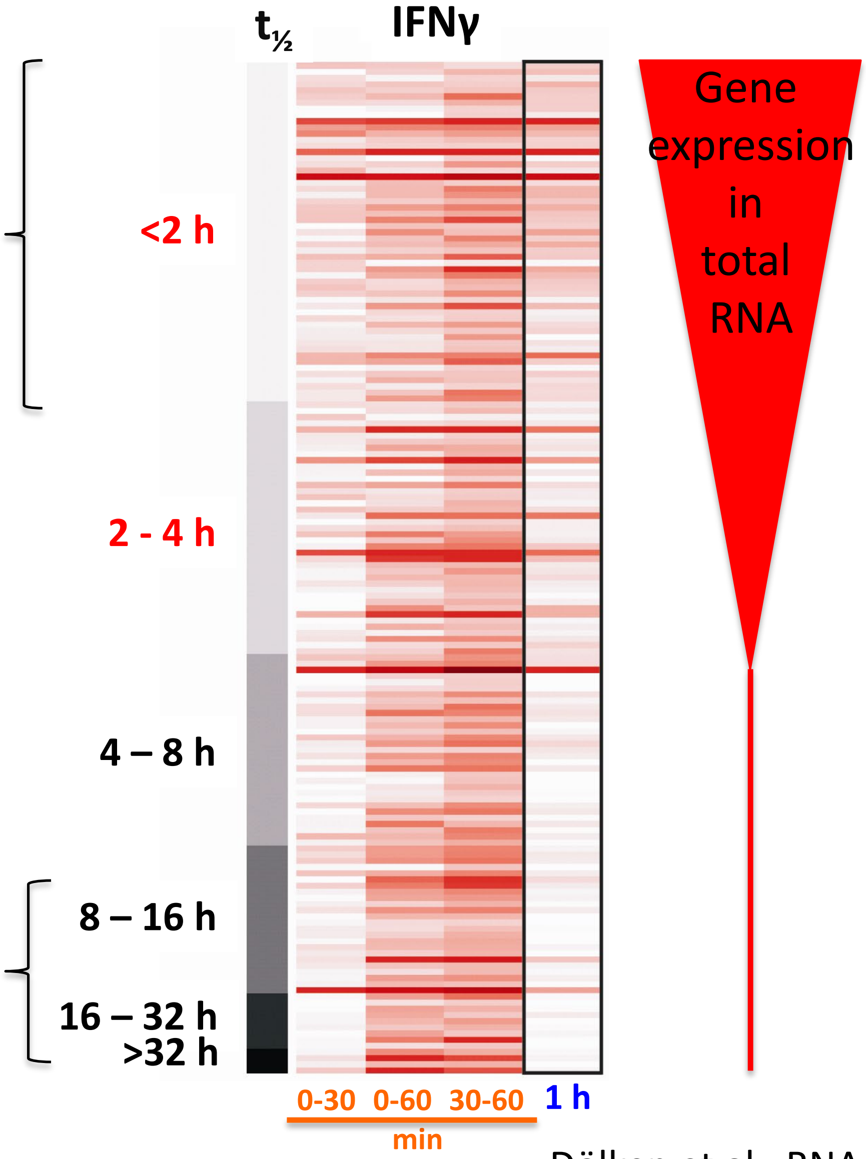
Transcription factors & regulatory genes

⇩

Systematic error

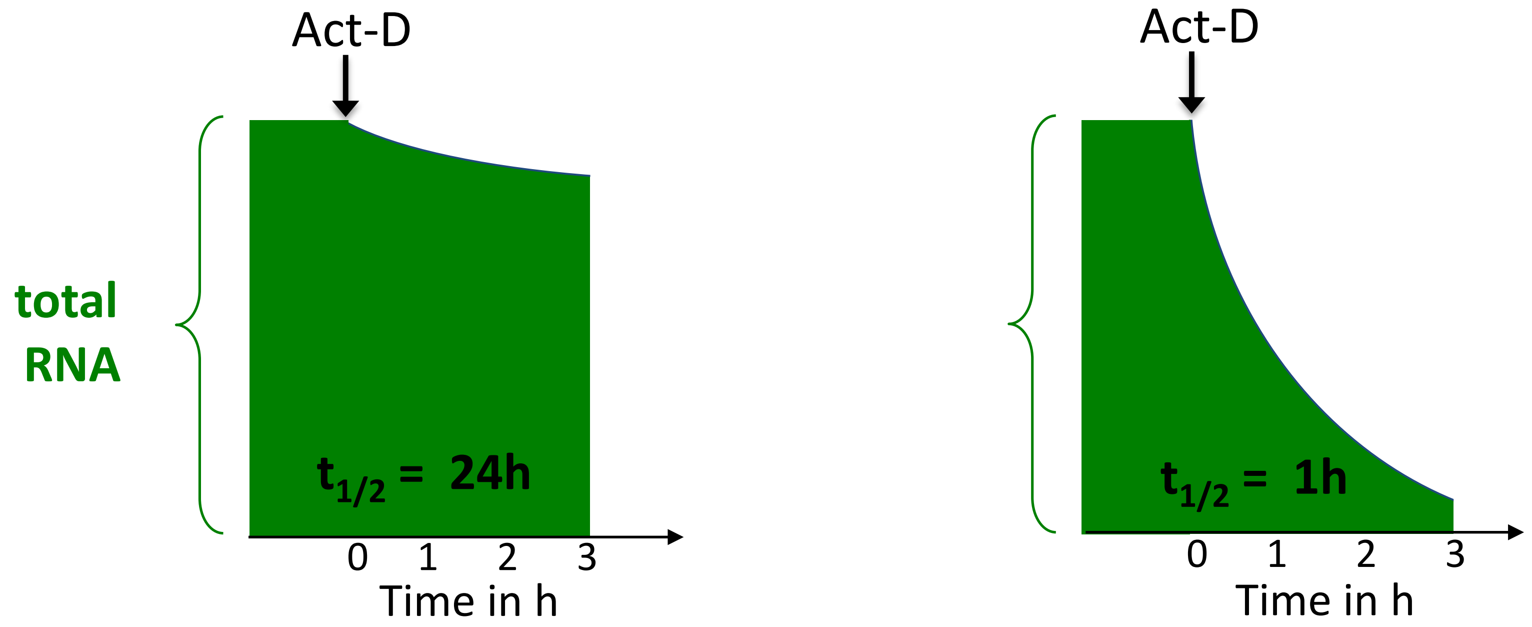
⇧

Metabolismus & house-keeping genes



Dölken et al., RNA 2008

Measuring RNA half-lives based on transcriptional arrest using Actinomycin D (Act-D)

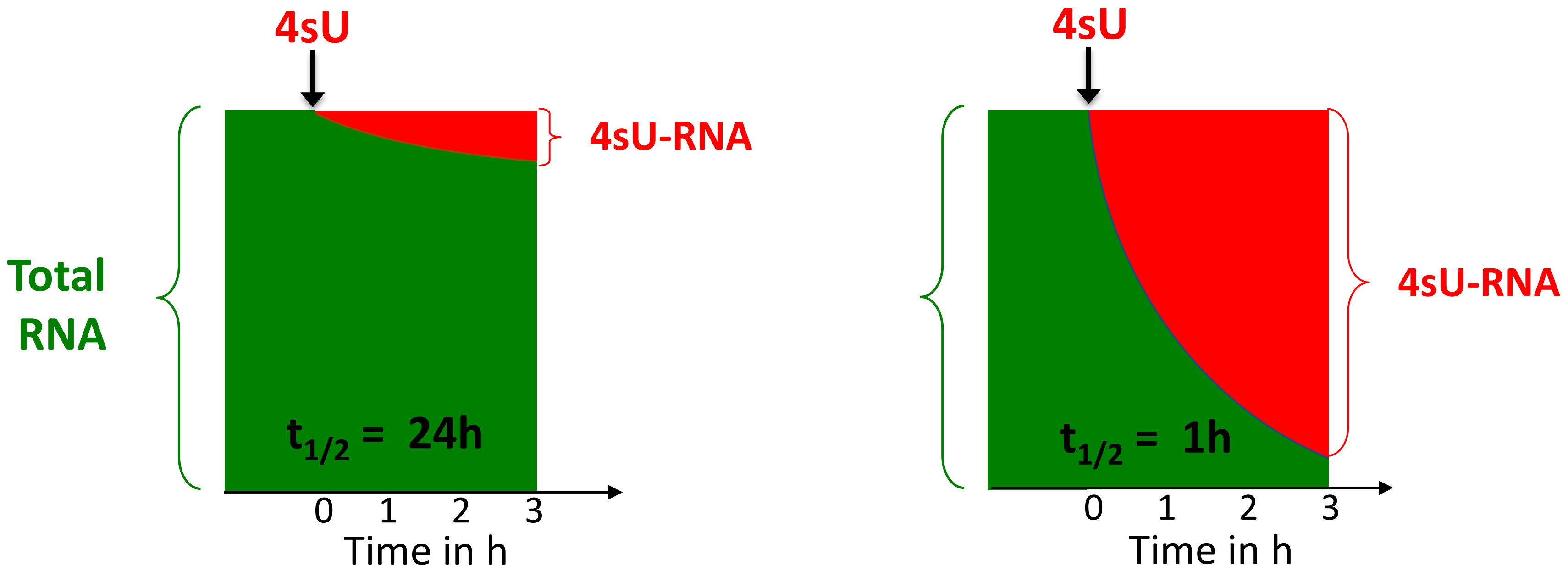


Problems:

- Act-D distorts normal RNA decay pathways
- Differences in total RNA levels are small even following prolonged Act-D exposure

⇒ Half-life measurements imprecise for medium- to long-lived RNAs

Measuring RNA half-lives based on transcriptional arrest using 4sU-RNA/total RNA ratios

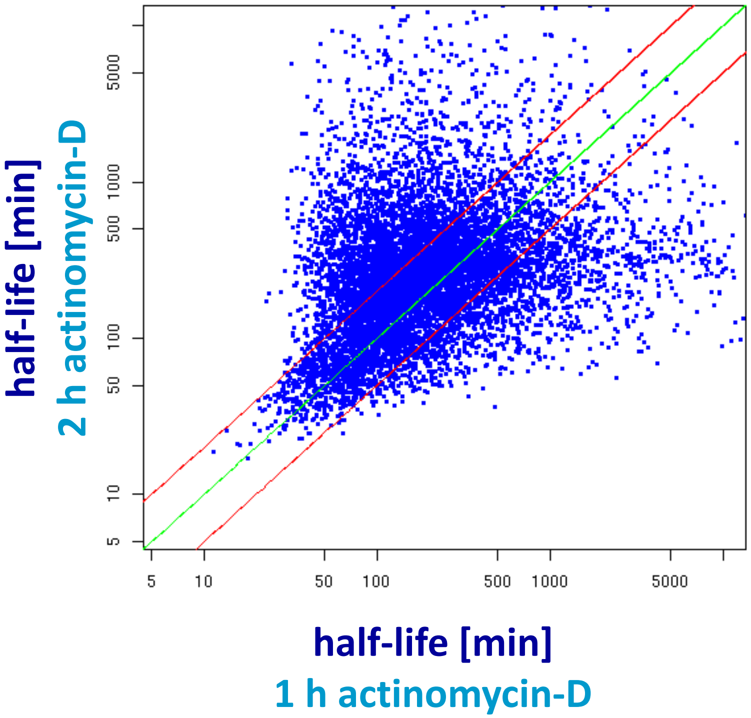


Advantages:

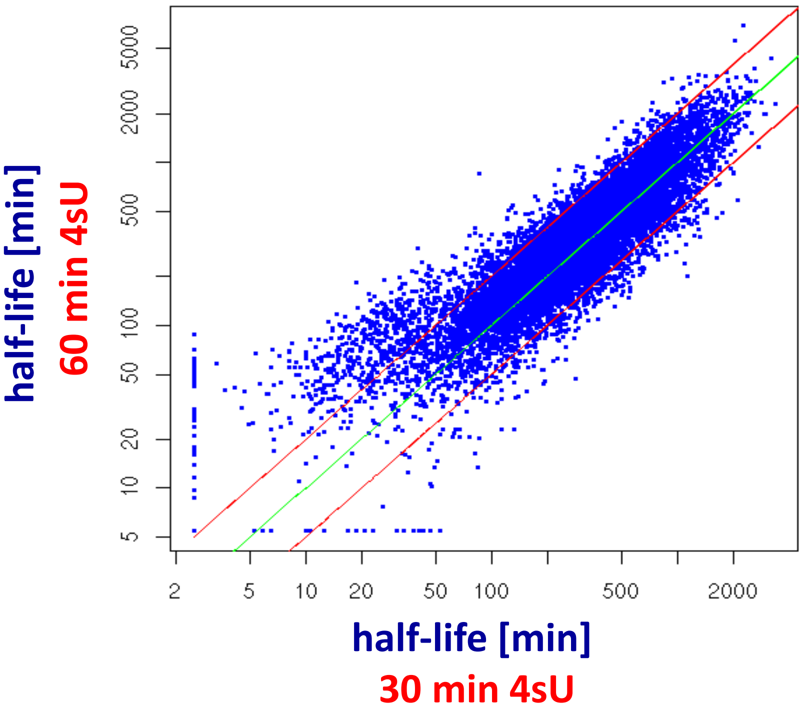
- No inhibition of RNA synthesis required
- Precise measurements of RNA half-life even for long-lived RNAs

RNA half-life measurements
- Actinomycin D vs 4sU-tagging -

RNA decay rates measured by
blocking transcription with actinomycin-D



RNA half-life measured based on
4sU-RNA / total RNA ratios

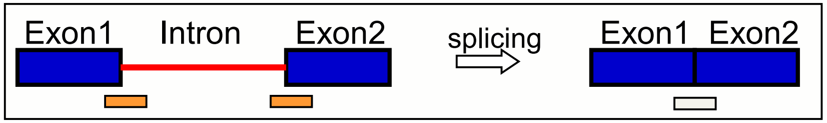


RNA half-lives [min] of >10.000 genes

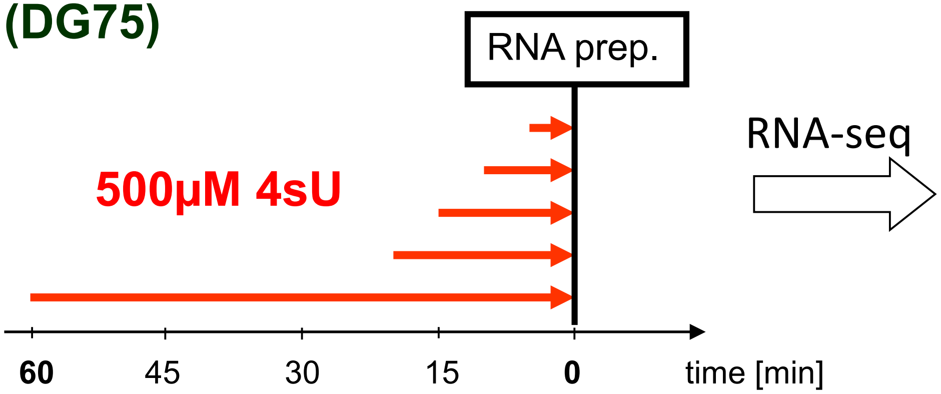
3 replicates Affymetrix MG430 2.0 arrays / condition

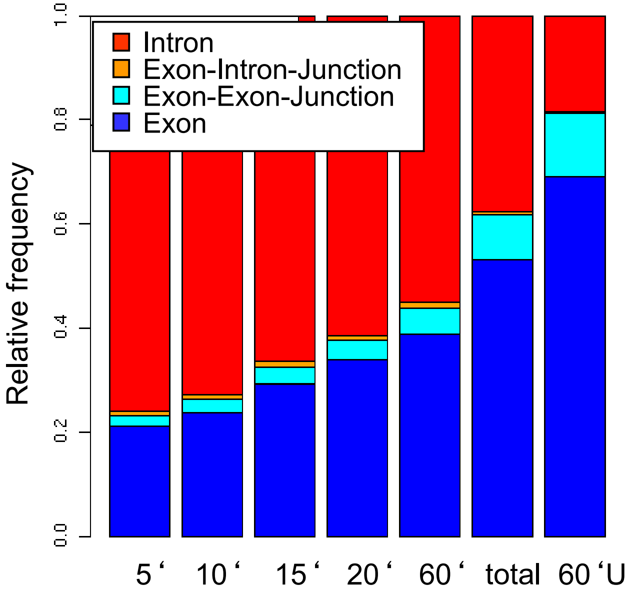
Friedel & Dölken et al., NAR 2009

Ultra-short and progressive 4sU-tagging reveals the kinetics of RNA processing at nucleotide resolution



Human B cells (DG75)





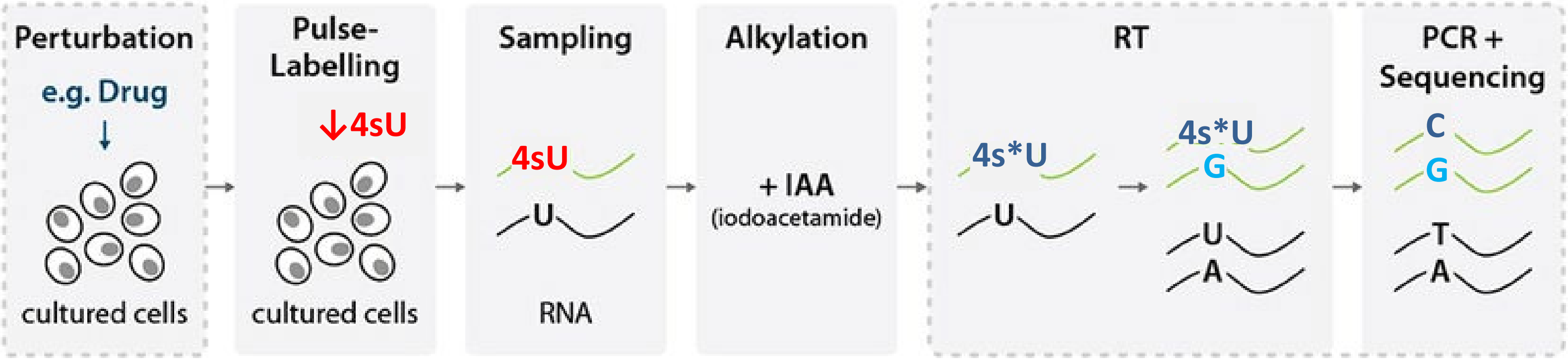
80% of introns already removed from 5‘ old 4sU-RNA

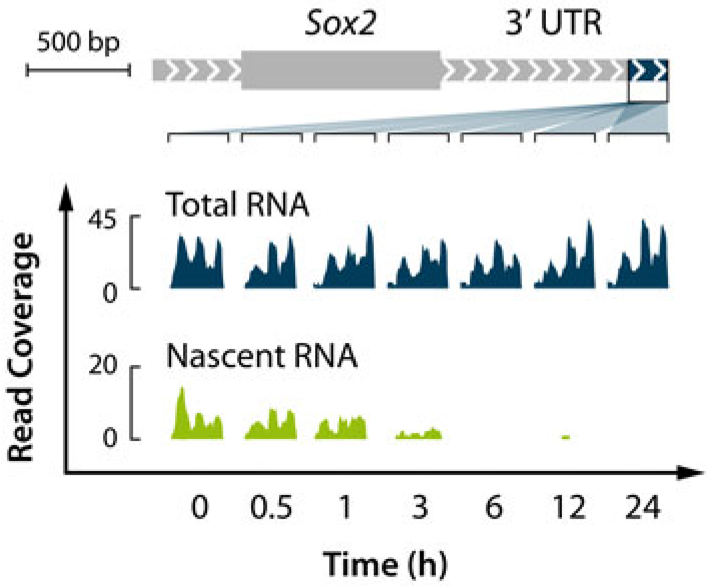
⇒ Splicing occurs co-transcriptional

Windhager et al., Genome Research 2012

Metabolic RNA labeling combined with nucleotide-conversion sequencing

SLAM-seq = Thiol (SH)-Linked Alkylation for the Metabolic sequencing of RNA.

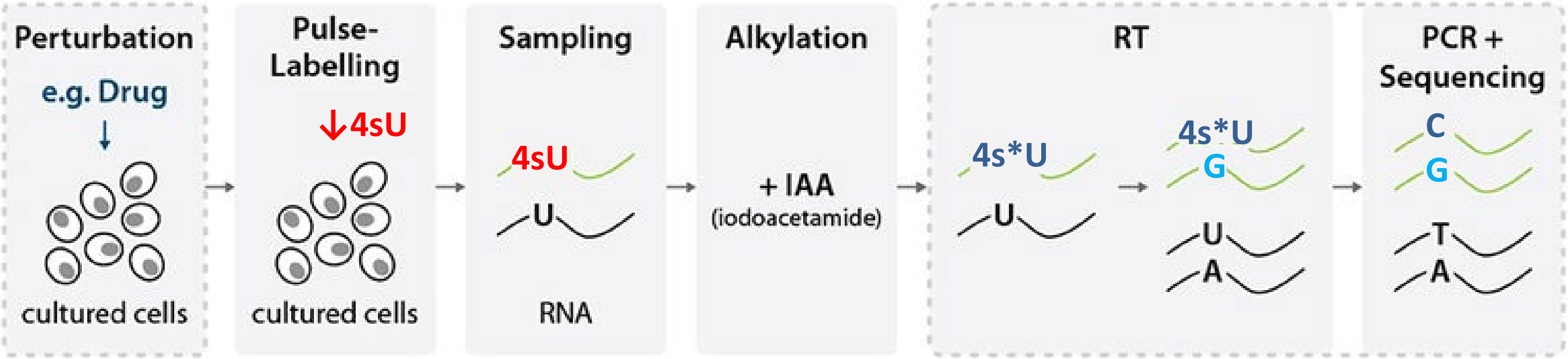


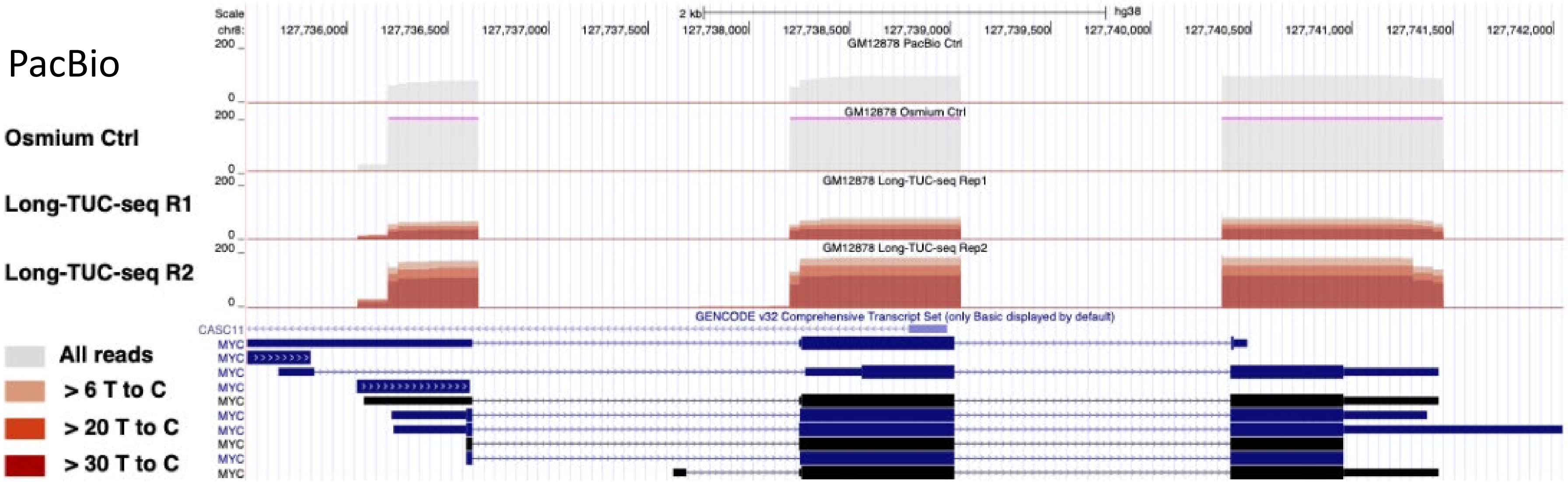


SLAM-seq:	Herzog et al., Nature Methods 2017
TimeLapse-seq:	Schofield et al. Nature Methods 2018
TUC-seq:	Riml et al., Angewandte Chemie 2017

Metabolic RNA labeling combined with nucleotide-conversion sequencing

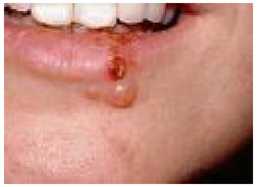
SLAM-seq = Thiol (SH)-Linked Alkylation for the Metabolic sequencing of RNA.



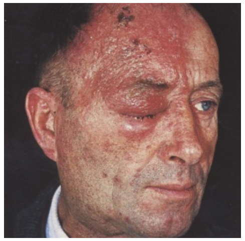


Rahmanian et al., bioRxiv 2020

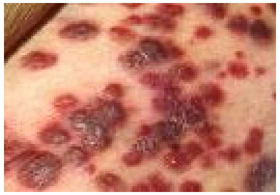
Herpesviruses



Herpes labialis (HSV-1)



Herpes zoster
⇒ VZV reactivation



Kaposi‘s sarkoma (KSHV) of an HIV patient

Large DNA Viruses (110-230 kb)

Primary infection

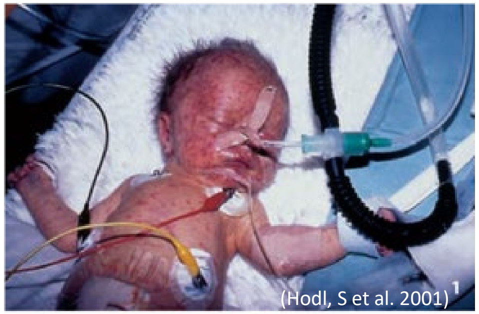
↓

Latent infection

↓ Immunosuppression

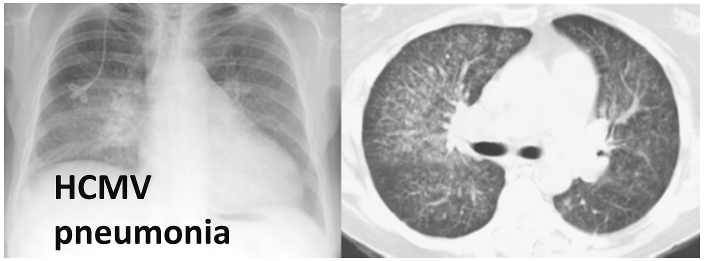
Reactivation

Human cytomegalovirus



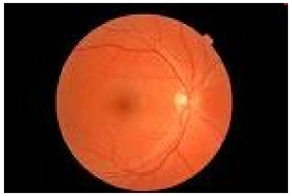
(Hodl, S et al. 2001)[1]

Blueberry Muffin Baby

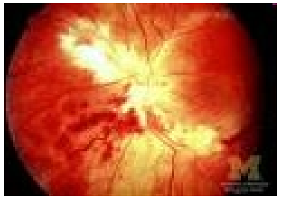


HCMV pneumonia

Healthy retina



HCMV retinitis

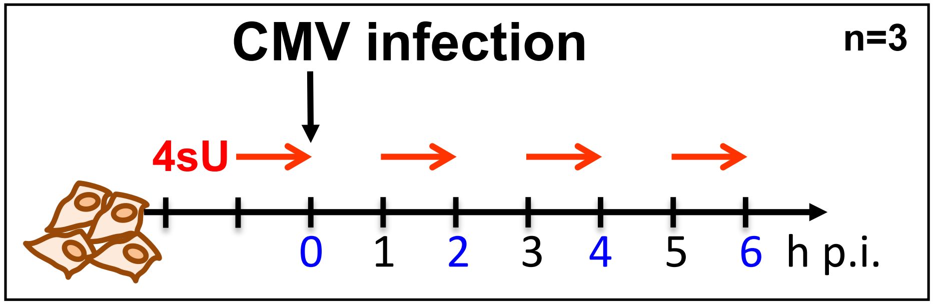


Key events during a productive virus infection

https://www.youtube.com/watch?v=Rpj0emEGShQ

Analysis of the transcriptional response to lytic CMV infection using 4sU-tagging

Experimental approach

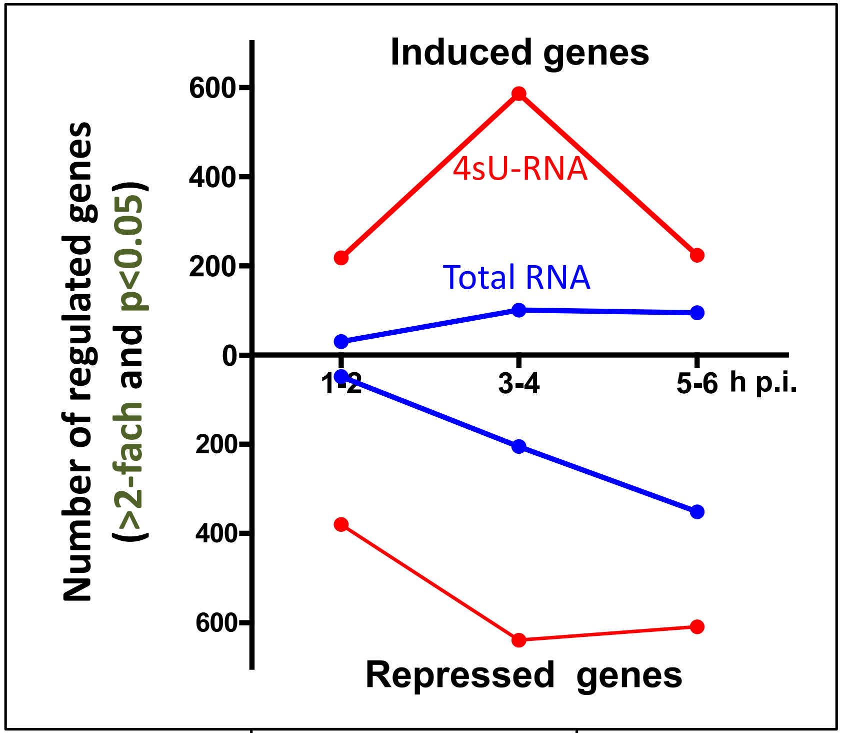


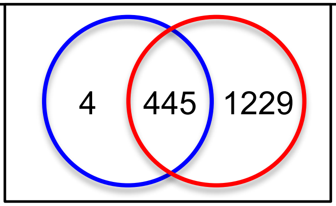
total RNA

4sU-RNA

Microarrays

Observed regulation

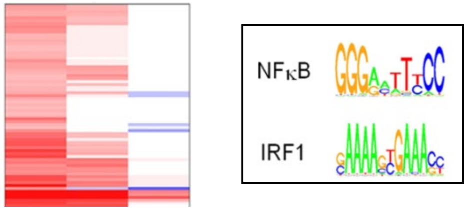




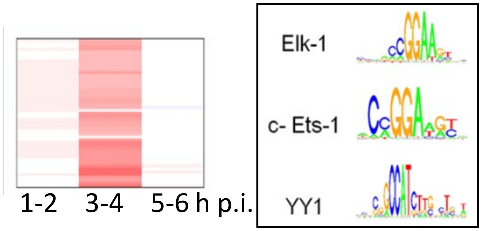
Marcinowski et al, PLoS Pathog. 2010

Transcriptionally regulated gene clusters during early cytomegalovirus infection

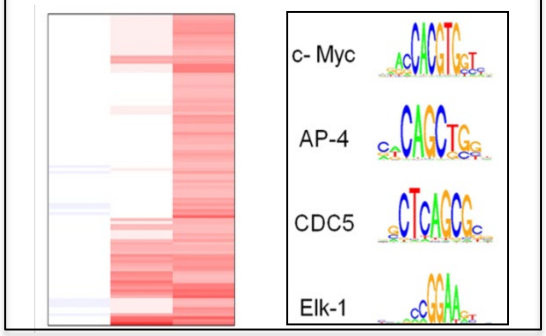
Interferon response
Inflammation (NF-κB)



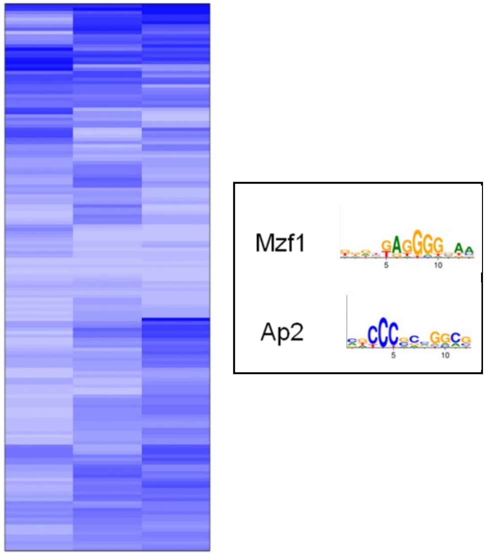
DNA Damage Response
Cell cycle



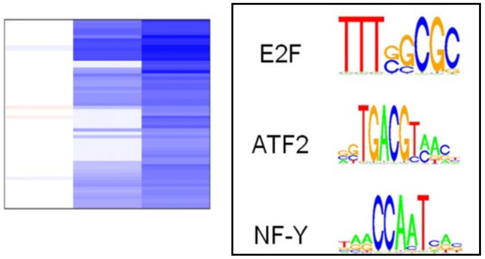
ER stress response



Cell adhesion / cytoskeleton
Regulation of transcription
Cell migration / Motility

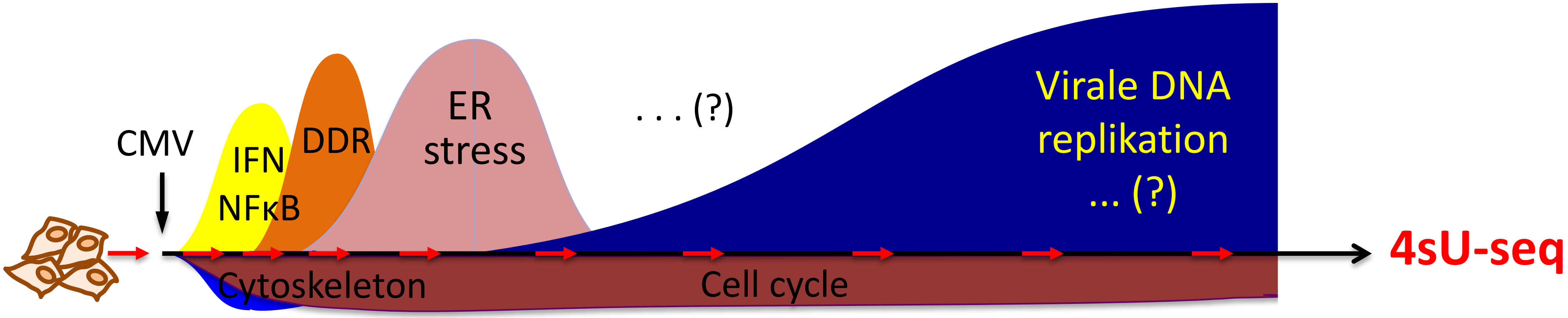


Chromatin organisation
Cell cycle



Marcinowski et al, PLoS Pathog. 2010

Characterisation of host cell modulation during lytic herpesvirus infection



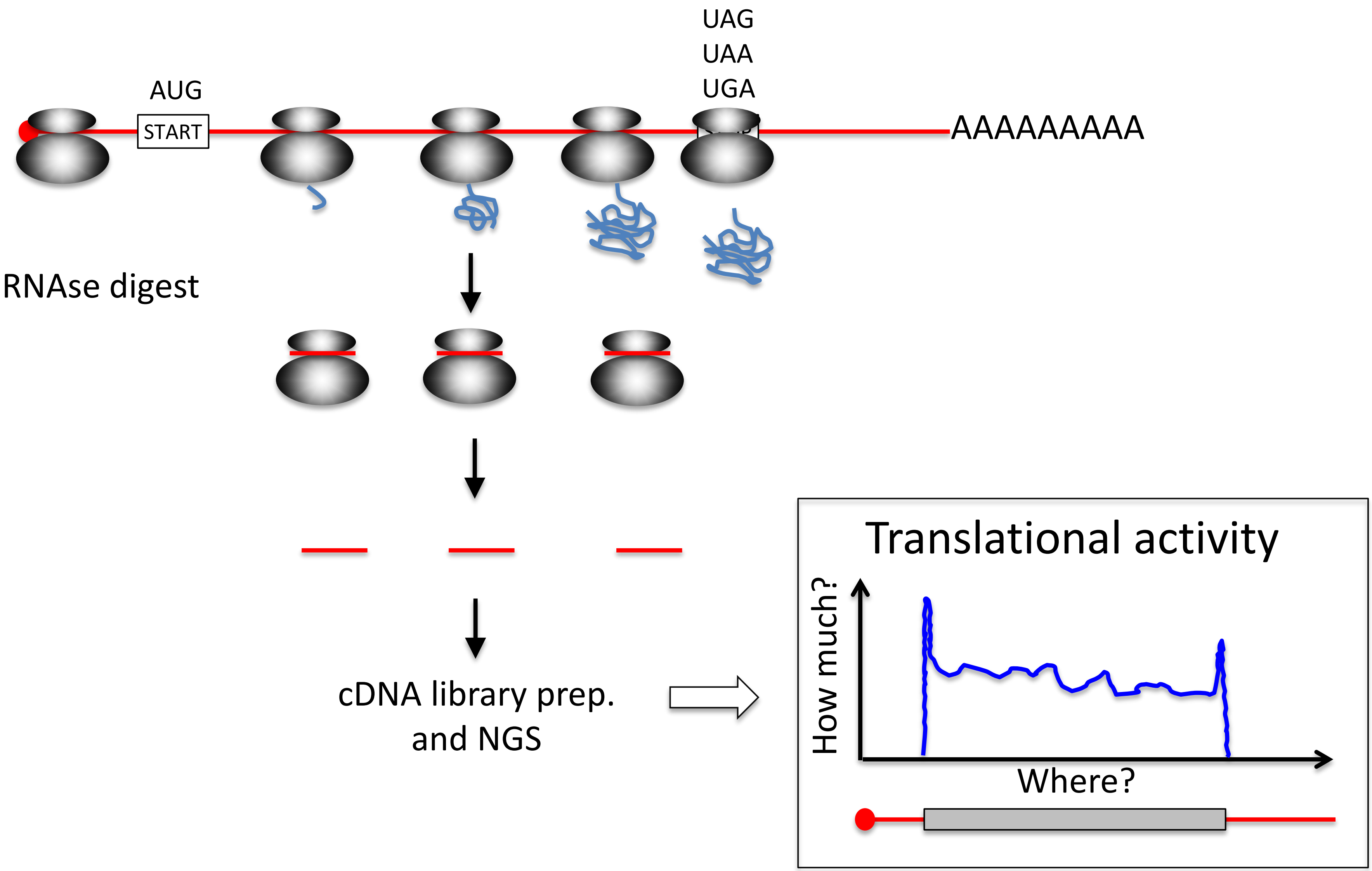
in silico Promoteranalysen

Chromatin structure
ATAC-seq / HiC

Targeted validation
(ChIP-seq/Reporter assays)

How does CMV reprogram its host cell?

Globale characterisation of translation using ribosome profiling



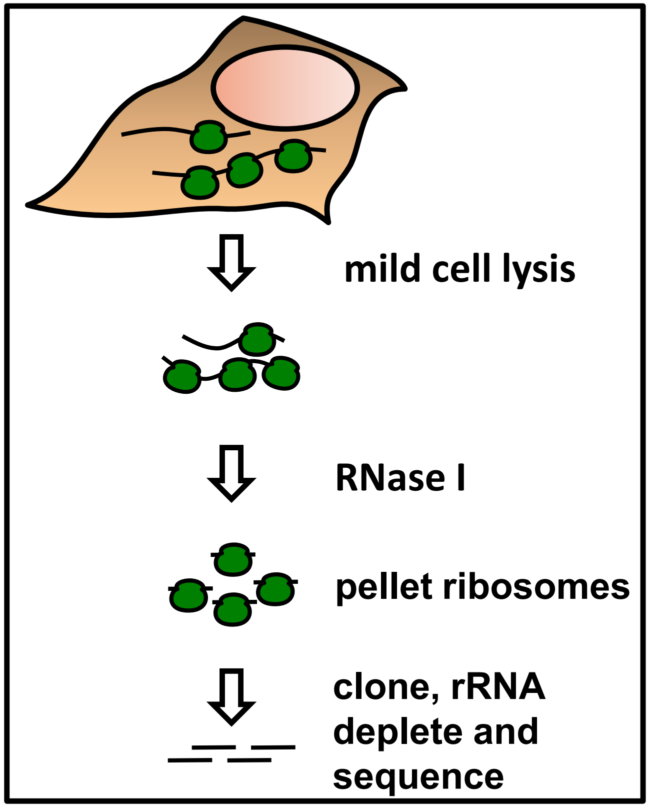
RNAse digest

cDNA library prep.
and NGS

Translational activity

Ingolia et al., Science 2009

Globale characterisation of translation using ribosome profiling



mild cell lysis

RNase I

pellet ribosomes

clone, rRNA deplete and sequence

Ingolia et al., Science 2009

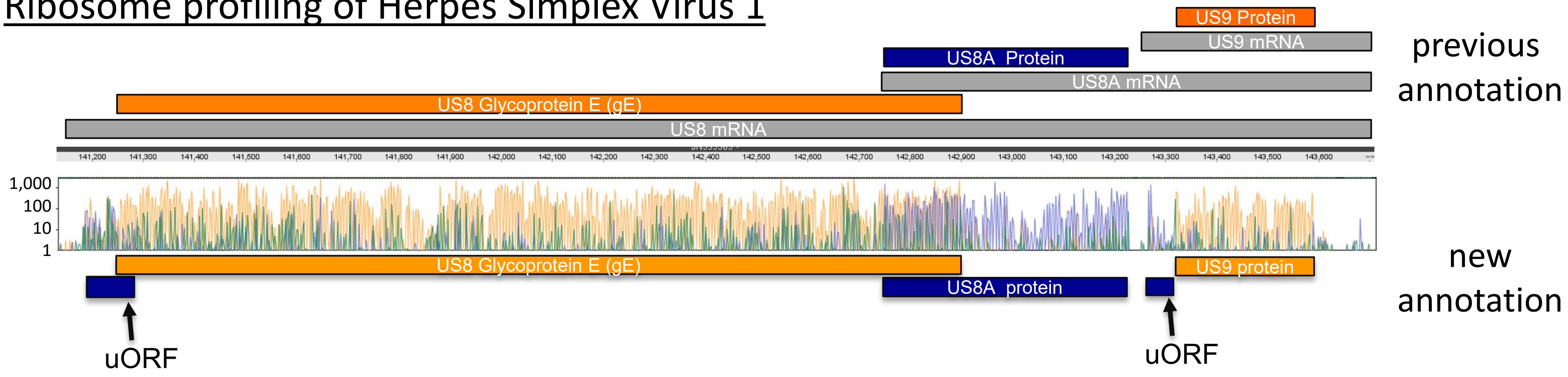
Real-time quantitative analysis of translational activity

Complete translatome
⇒ ORFs / uORFs
⇒ alternative translation start sites

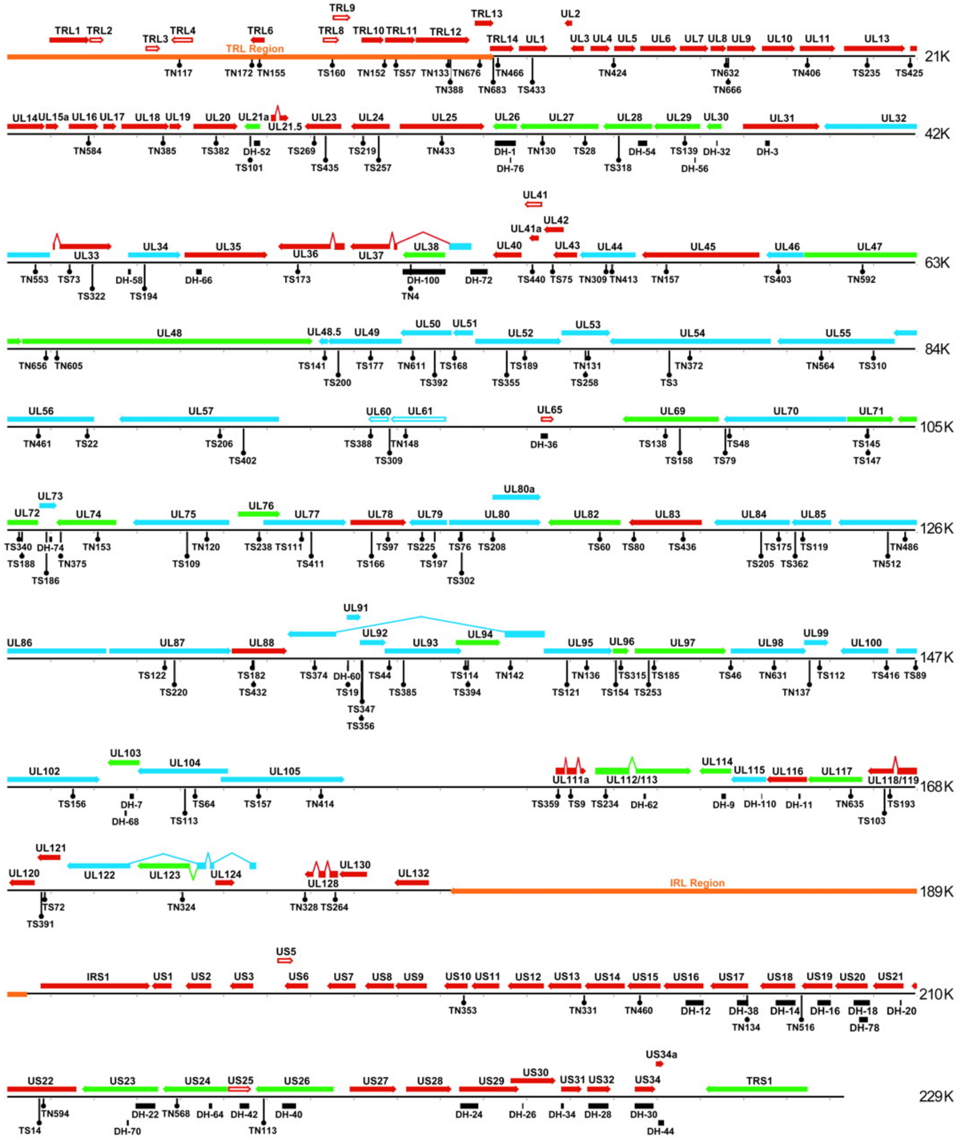
Pre-treatment with chemical inhibitors (Harringtonin, Lactimidomycin) allows translation start site profiling

Characteristics of ribosome profiling data

Ribosome profiling of Herpes Simplex Virus 1



Previous annotation of the human cytomegalovirus (HCMV) genome

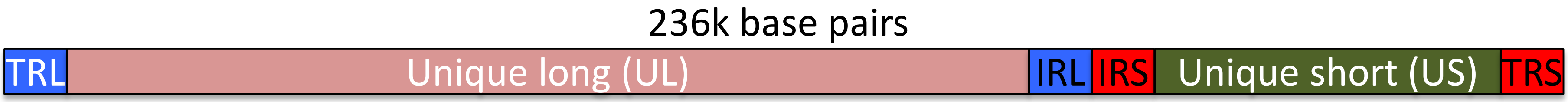


HCMV

- 236kb dsDNA genome
- 170-200 genes encoding for proteins >100aa (bioinformatic predictions)
- 11 pre-miRNAs
- 4 large non-coding RNAs

Dong Yu et al., PNAS 2003

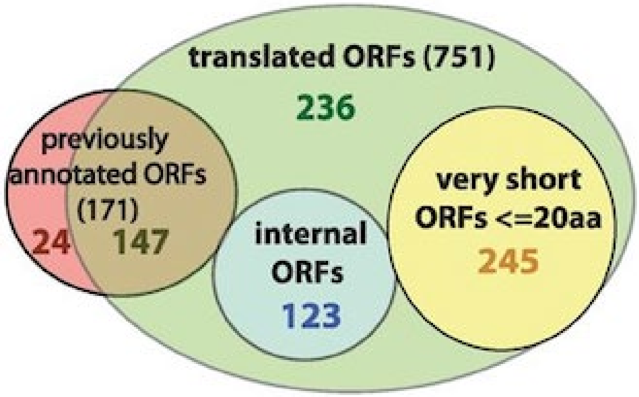
Re-annotation of the HCMV translatom using ribosome profiling

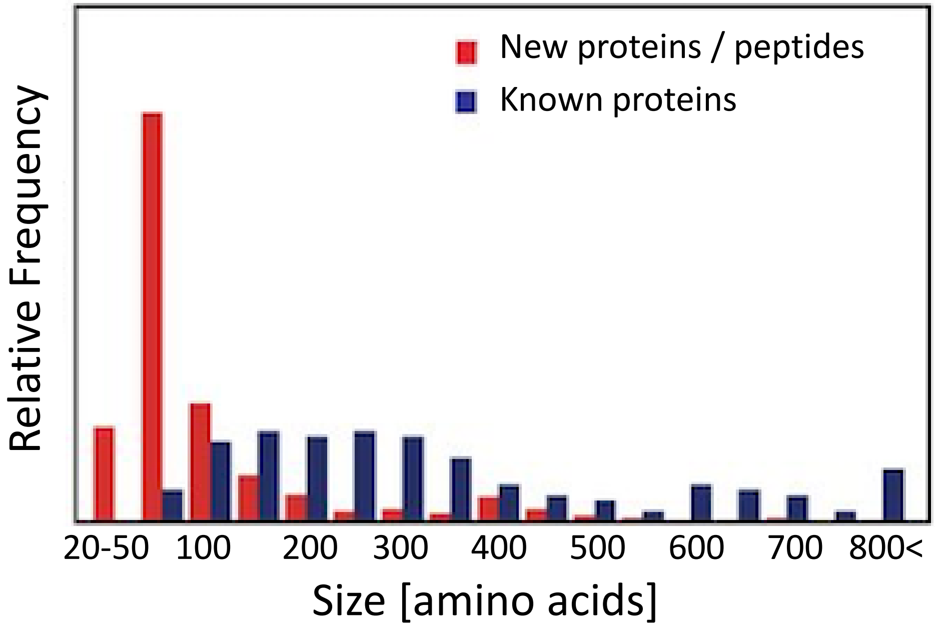


= **t**erminal **r**epeats (**l**ong)

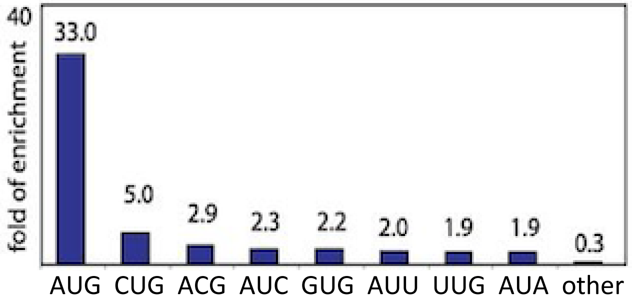
= **i**nternal **r**epeats (**s**hort)

171 proteins => >750 viral proteins / peptides



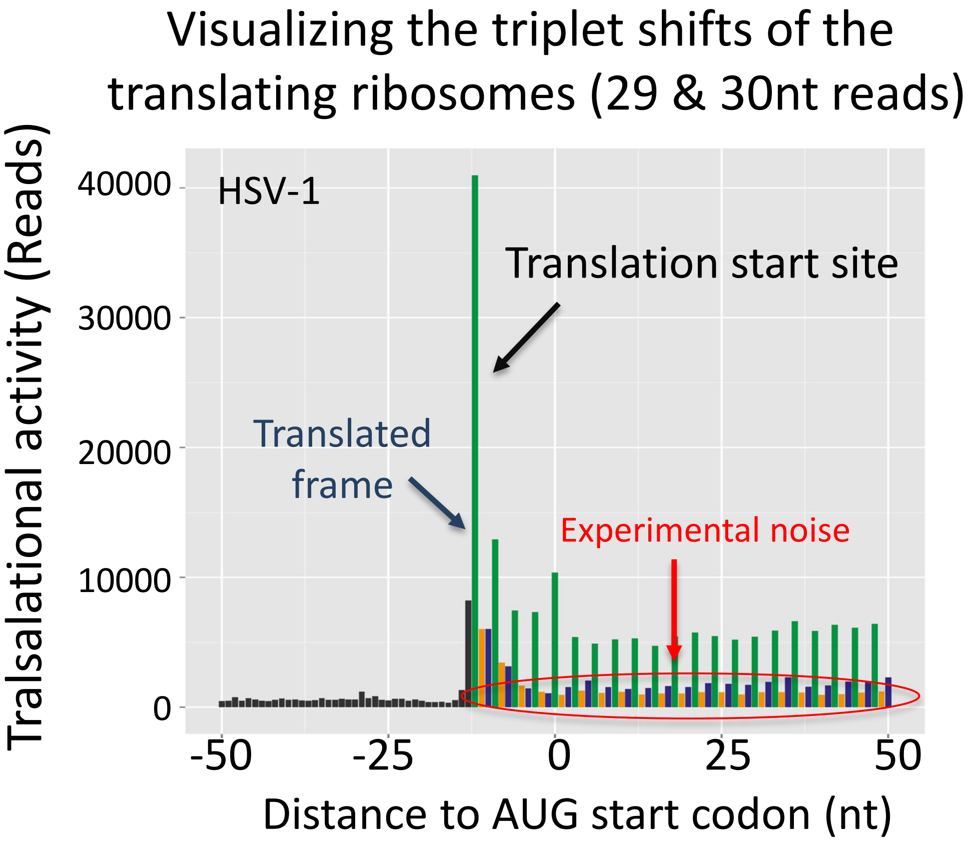


Usage of different start codons

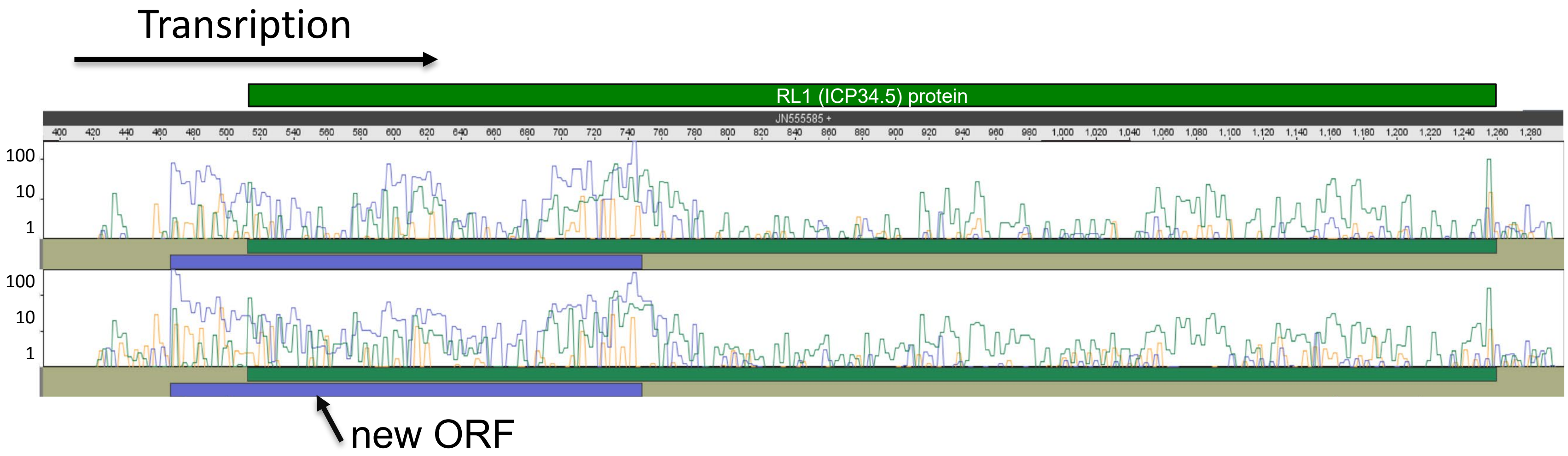


Decoding Cytomegalovirus
Stern-Ginossar et al., Science 2012

Ribosome profiling visualizes triplet shifts of translating ribosomes

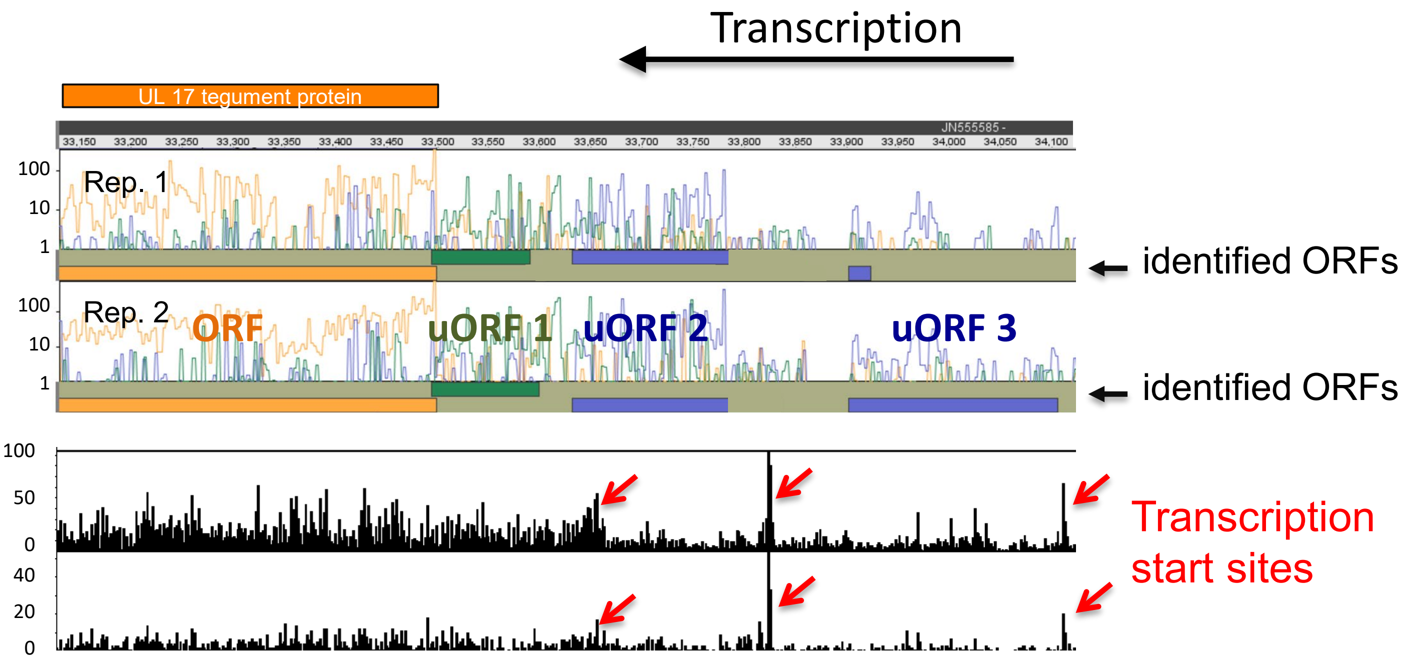


Examples of new viral proteins (HSV-1)



Transription

new ORF



Transcription

previous annotation

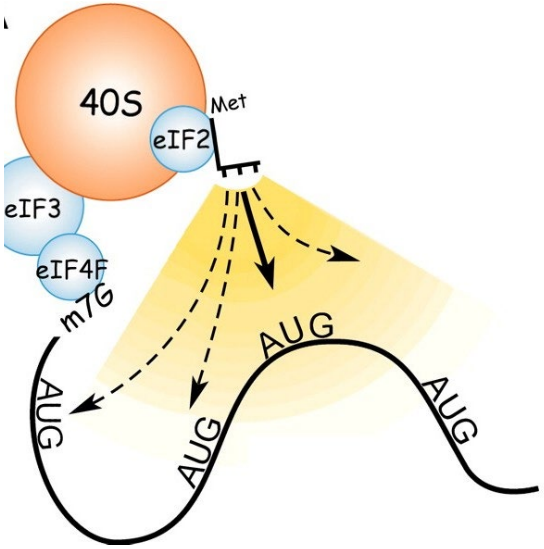
identified ORFs

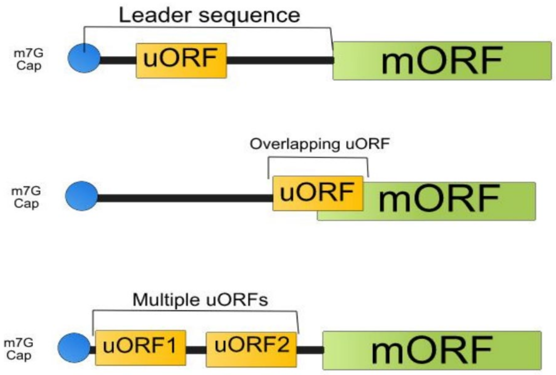
identified ORFs

Transcription start site profiling

Transcription start sites

uORFs regulate gene expression at the level of translation





Features of uORFs:

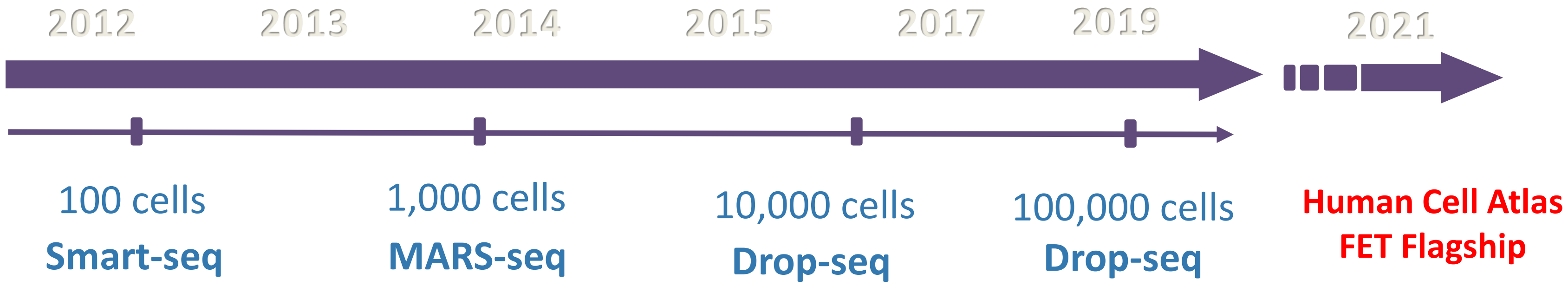
- present in >40% of our genes
- 20-30% initiate from non-canonical start codons (CUG, GUG, ACG)
- generally <100 aa in size
- vast majority of uORF-encoded polypeptides are inherently unstable

  ⇒ undetectable by whole proteome mass spec
- some uORFs encode functional polypeptides
- regulate translation of downstream ORFs by impairing translation initiation

How does a cell know how much mRNA to express for a given gene?

How does heterogeneity in the fulfillment of this task affect cell function

Timeline of single cell RNA-seq



2012

2013

2014

2015

2017

2019

2021

100 cells
Smart-seq

1,000 cells
MARS-seq

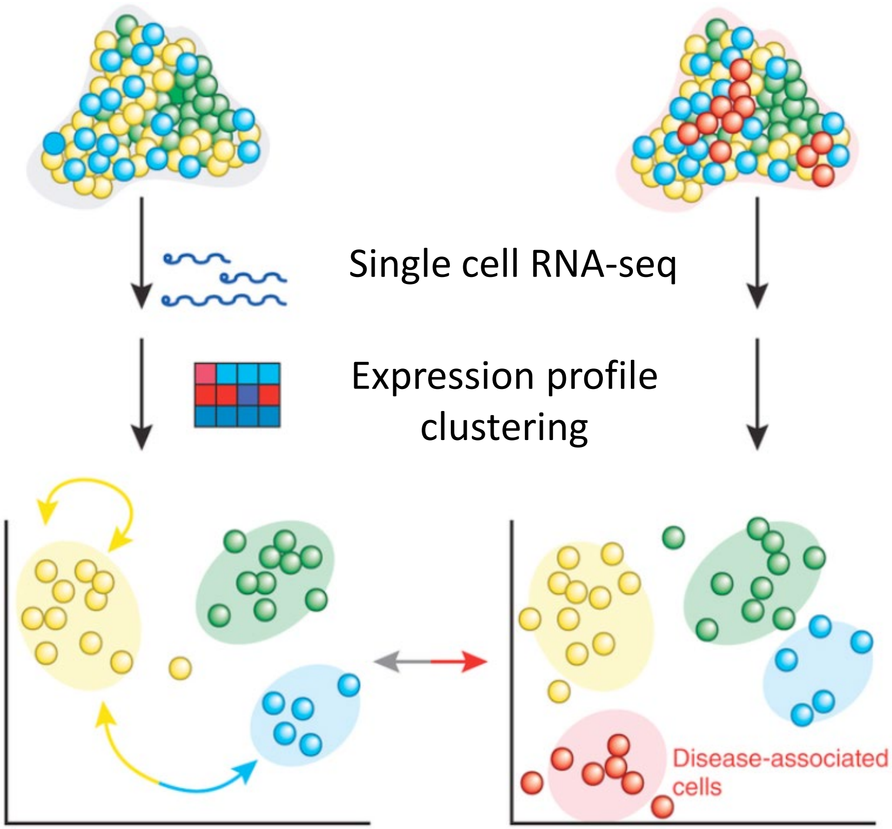
10,000 cells
Drop-seq

100,000 cells
Drop-seq

Human Cell Atlas
FET Flagship

MARS-seq = Massively parallel single cell RNA-seq

Dissection of tissue composition in health and disease



Healthy

Pathological

Tissues

Single cell RNA-seq

Expression profile clustering

Cell-type maps

Within cell type

- Stochasticity, burst
- Regulatory networks
- Allelic expression patterns

Between cell types

- Identify biomarkers
- (post)-transcriptional differences

Between tissues

- Cell type composition
- Disease-associated cells
- Altered transcription in matched cell types

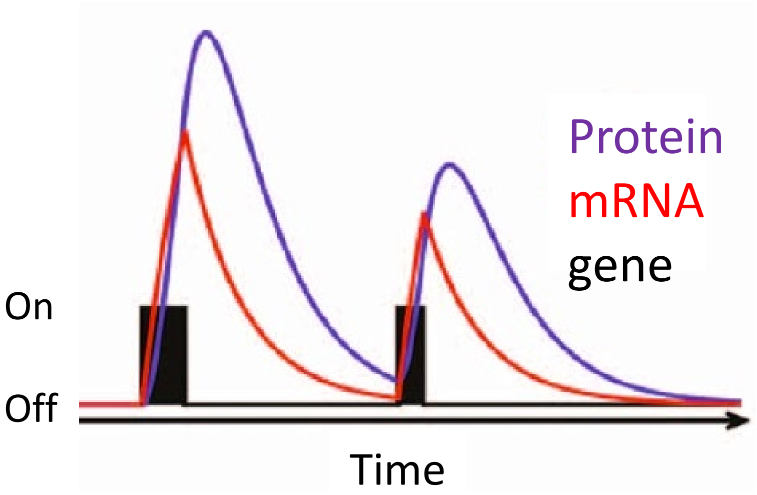
Why single cell RNA sequencing?

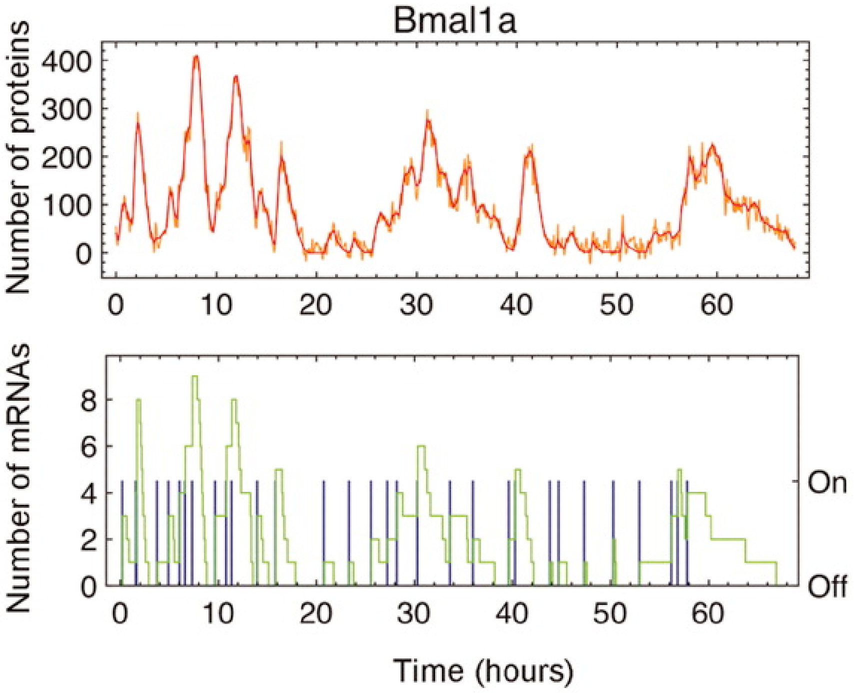
- Understanding heterogeneous tissues
- Identification and analysis of rare cell types
- Changes in cellular composition
- Transcriptional changes in subpopulations of cells

Examples of applications:

- Differentiation paths
- Cancer heterogeneity
- Neural cell classification
- Embyronic development
- Drug treatment responses

Transcriptional bursting

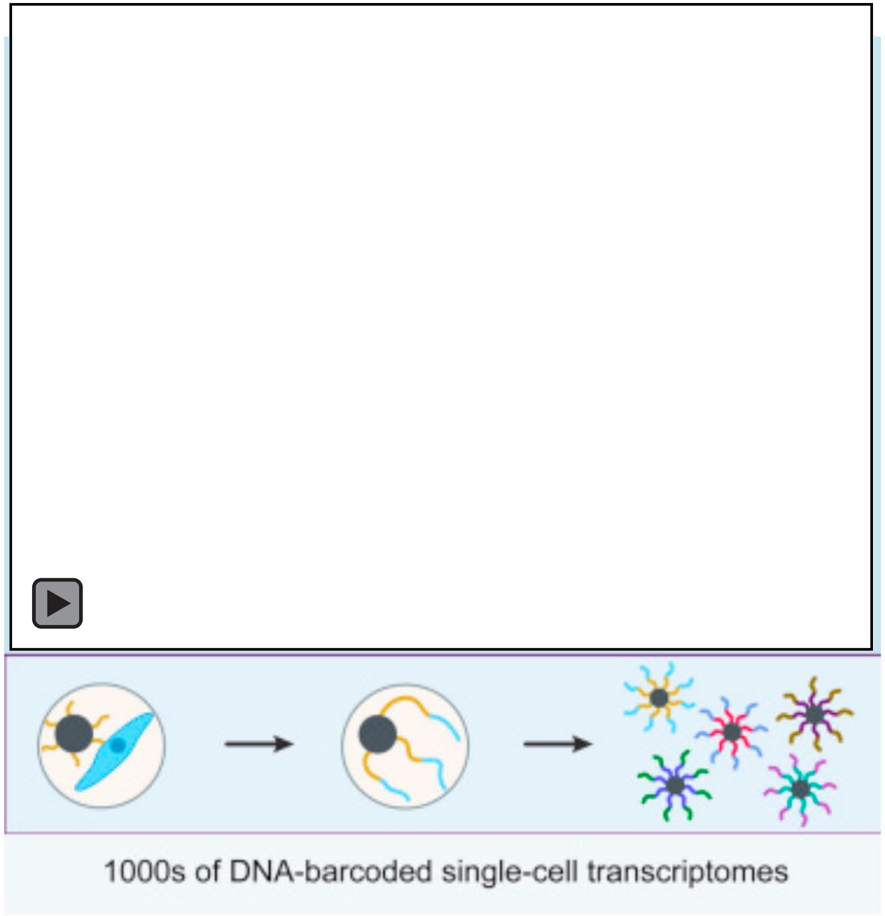




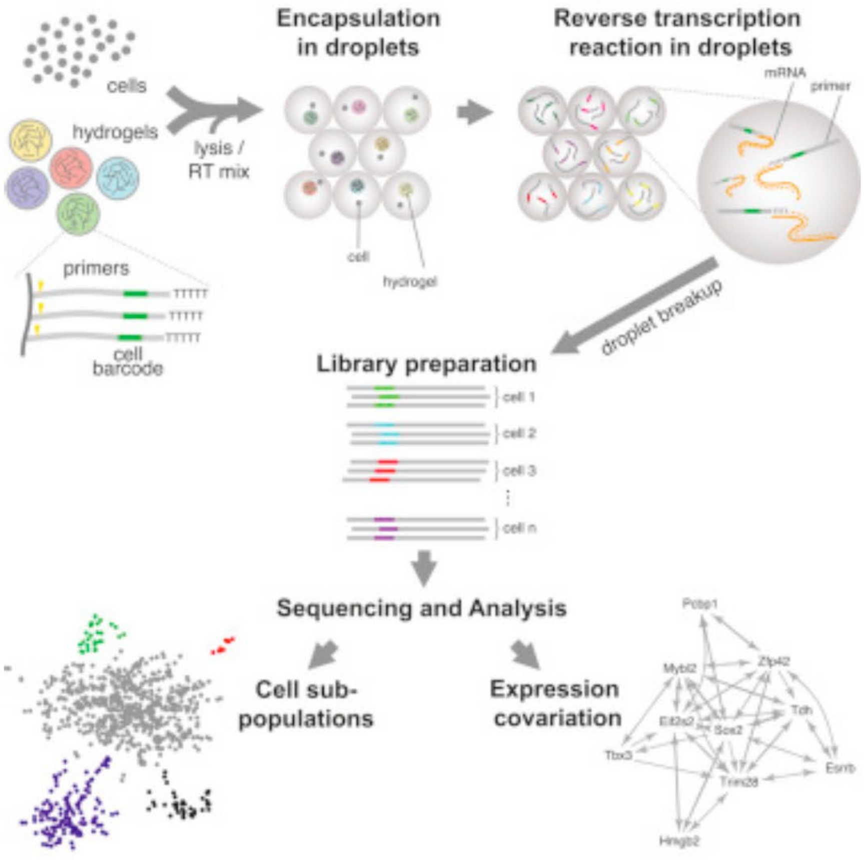
- Burst frequency and size is correlated with mRNA abundance
- Many TFs have low mean expression (and low burst frequency) and will only be detected in a fraction of the cells

Suter et al., Science 2011

Droplet-based microfluids approaches



Macosko et al. *Cell* 2015
McCarrol, Regev etc. Broad/Harvard



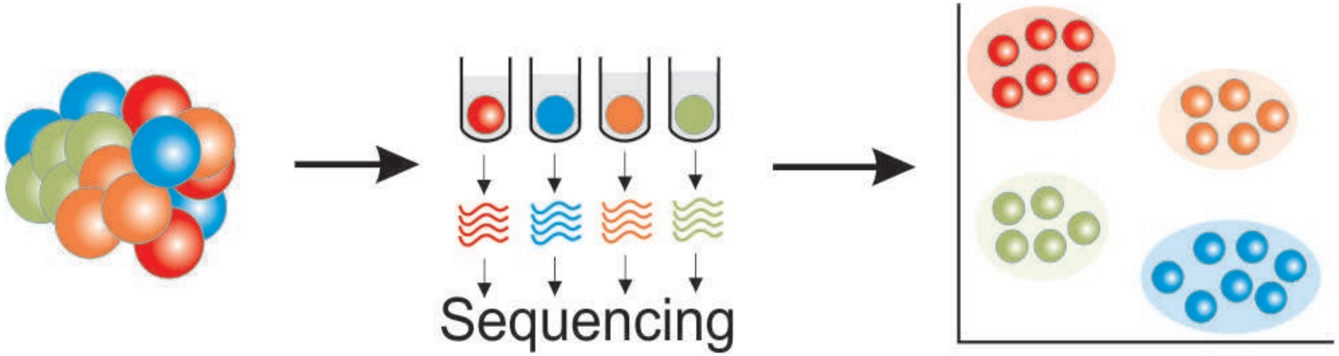
Klein et al. *Cell* 2015
Kirschner, Weitz etc. Harvard

Problems of single cell RNA-seq

- Amplification bias
- Drop-out rates (~4,000 vs. >10,000 genes per cell)
- Stochastic gene expression
- Sampling bias
- Bias due to cell-cycle, cell size and other factors
- Mainly for poly(A) transcripts so far

Current limitations of single cell RNA-seq

- Each cell can only by sequenced once
  - ⇒ scRNA-seq only allows to indirectly analyze cellular responses

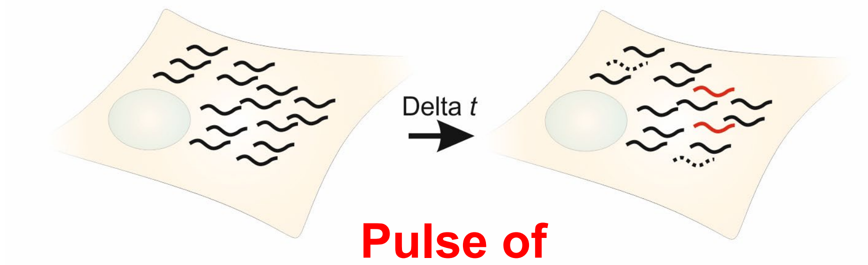


- Poor temporal resolution for short-term changes in transcriptional activity

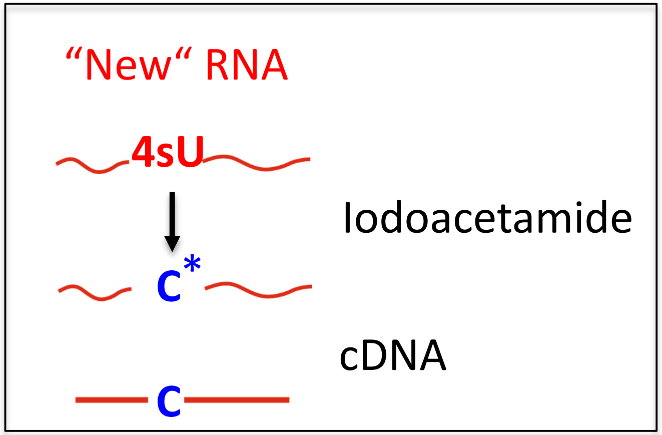
- No differentiation of changes in RNA synthesis processing and decay

SLAM-seq:
S-Linked Alkylation for the Metabolic labeling of RNA

Perturbation



Pulse of 4sU labeling



"New" RNA

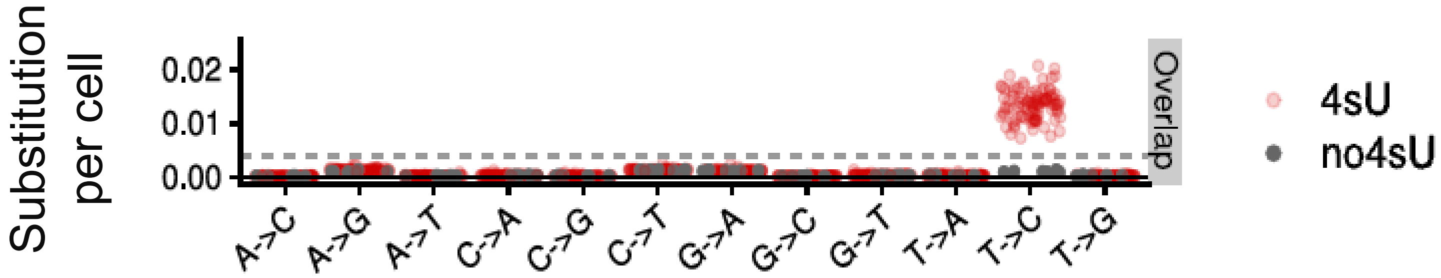
4sU

Iodoacetamide

C*

cDNA

C

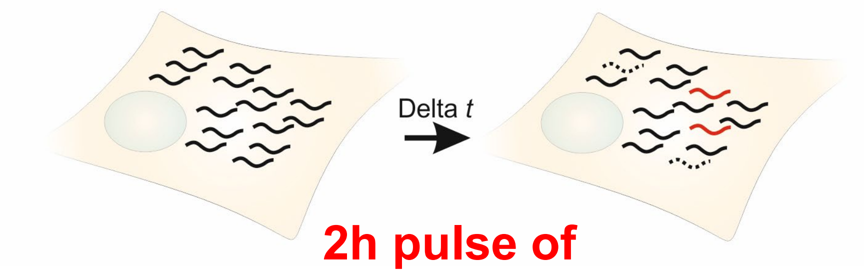


Alternatives: TimeLapse-seq, TUC-seq

Herzog et al., Nature Methods 2017

Development of single cell SLAM-seq (scSLAM-seq)

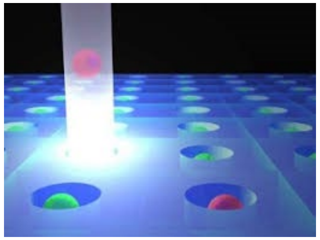
Perturbation



2h pulse of 4sU labeling

MCMV infection (MOI=10)

Single cell sorting (FACS)



96-well plate

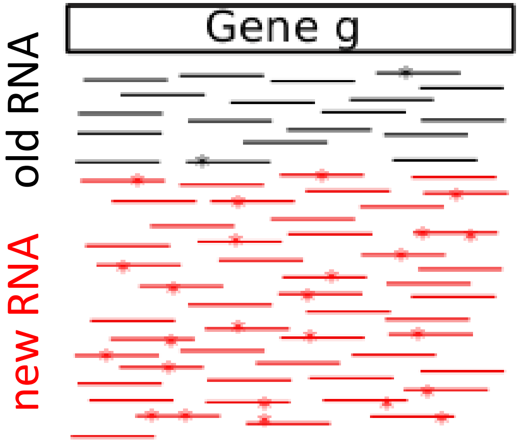
Lysis and 4sU alkylation
5 min 50°C IAA

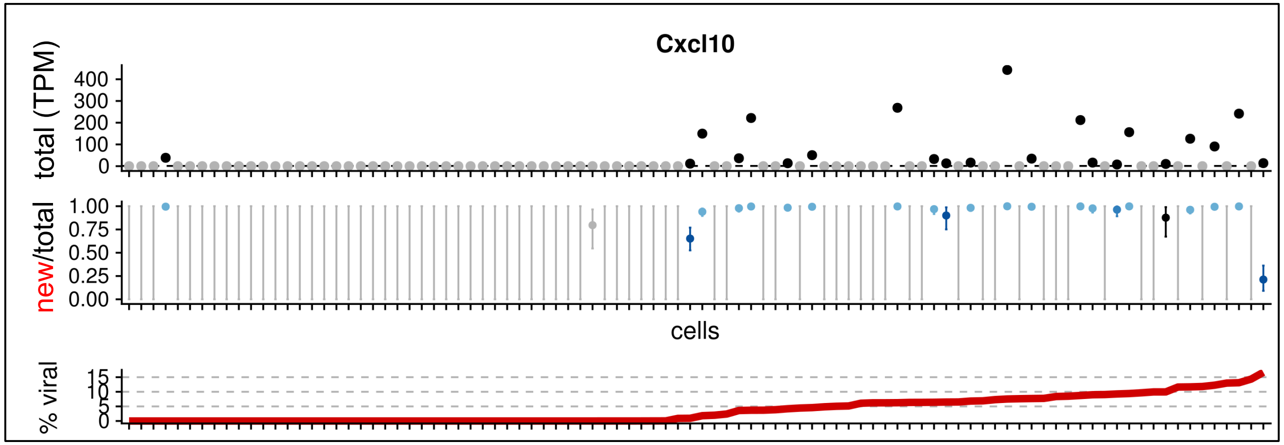
Library preparation (SMART-Seq) and RNA-seq

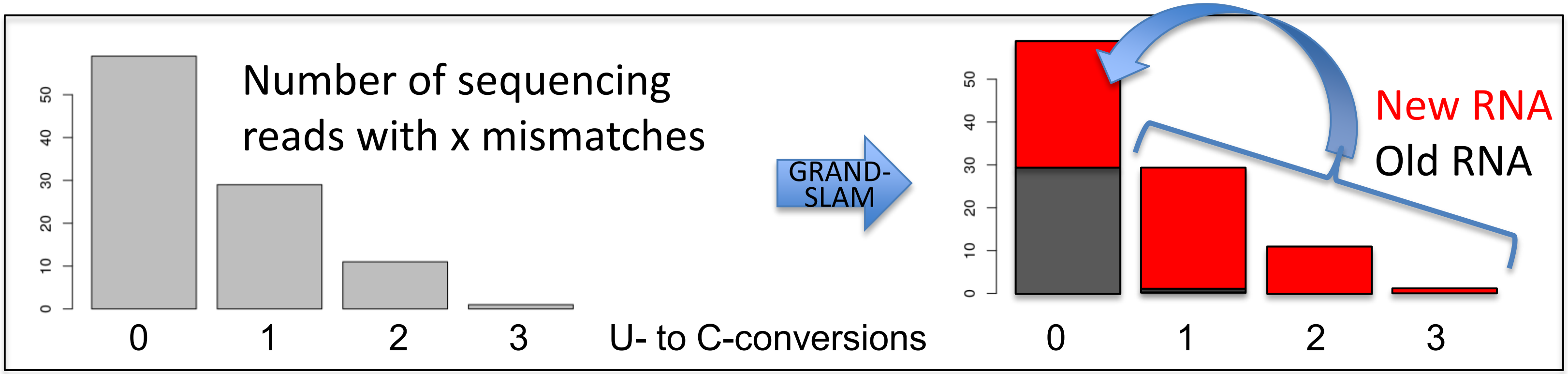
Erhard et al., Nature 2019

GRAND-SLAM

Globally Refined Analysis of Newly transcribed RNA and Decay rates using SLAM-seq

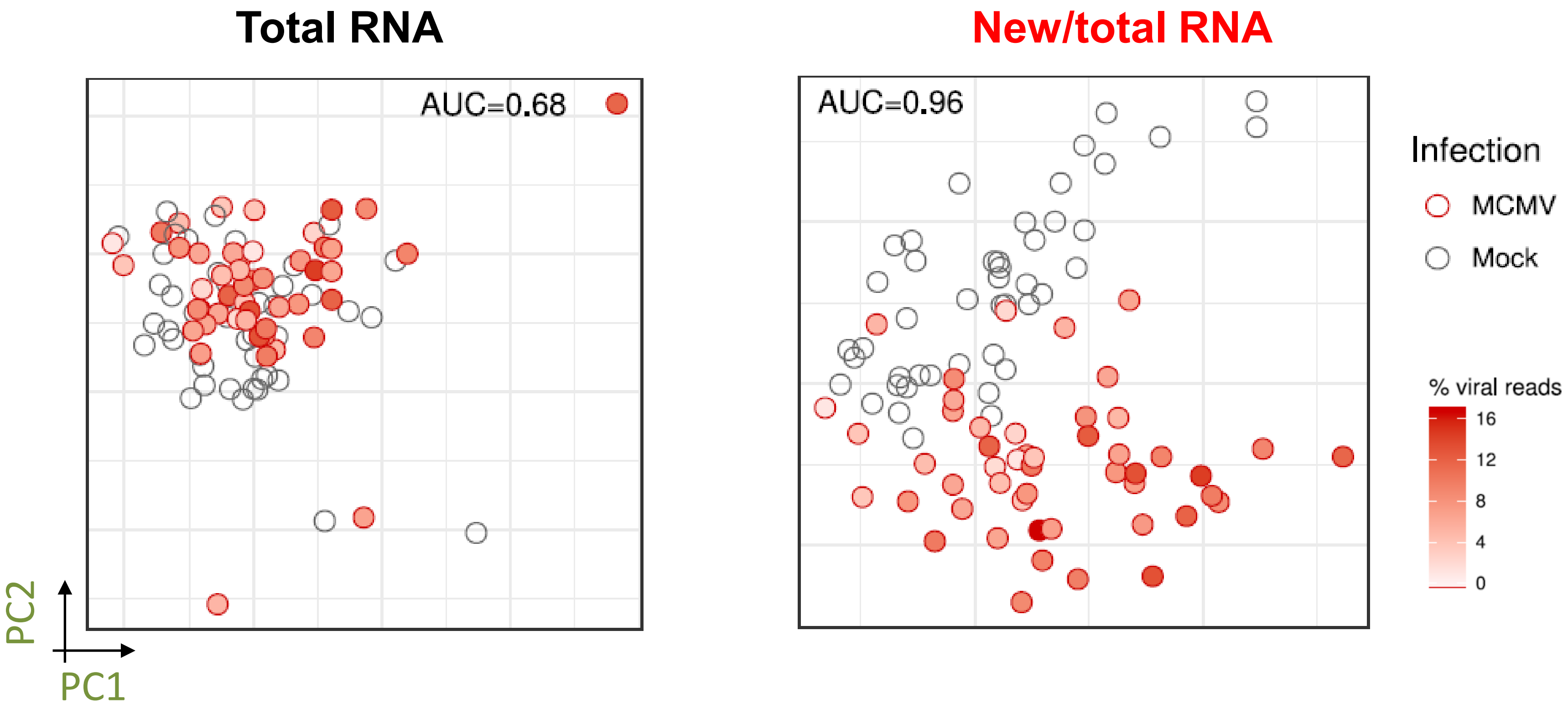





Number of sequencing reads with x mismatches

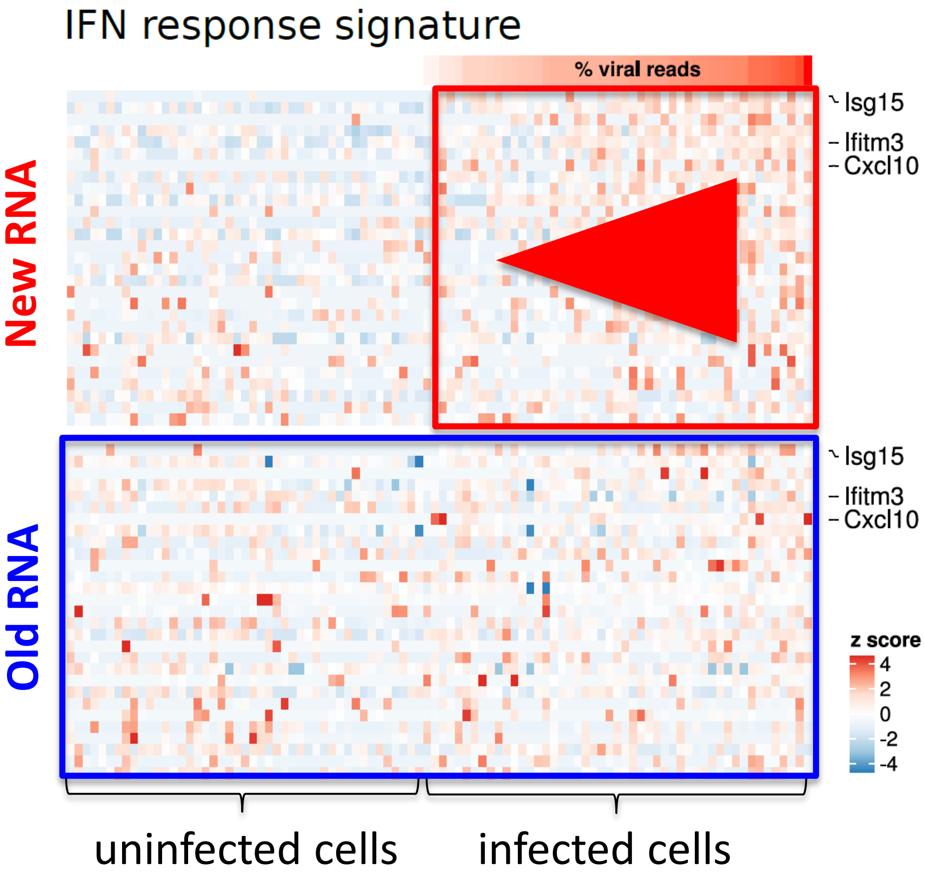
⇒ Credible intervals

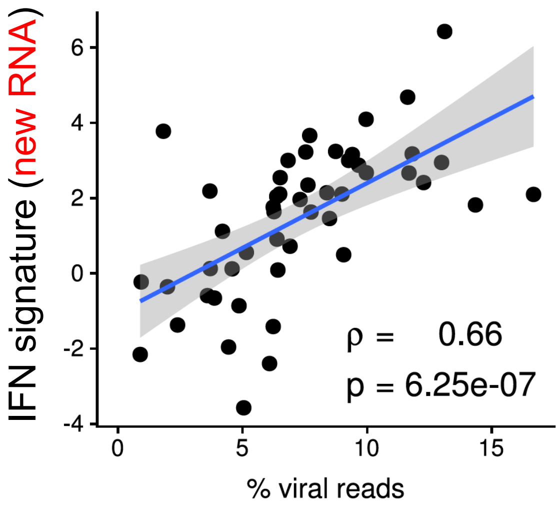
scSLAM-seq increases the temporal resolution for detecting rapid alterations in gene expression

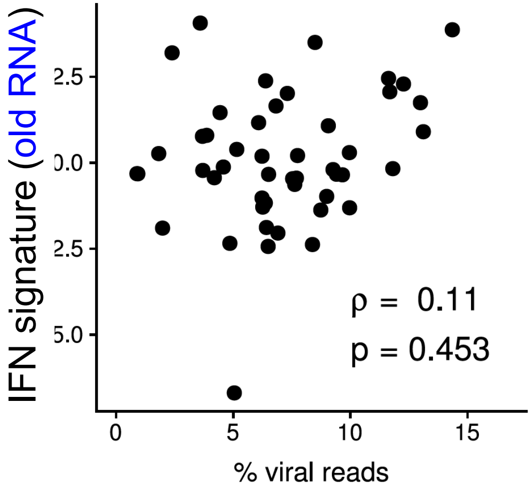


Total RNA:
Intercellular heterogeneity >>> virus-induced changes at 2h p.i.

The most highly infected cells activate the strongest interferon response

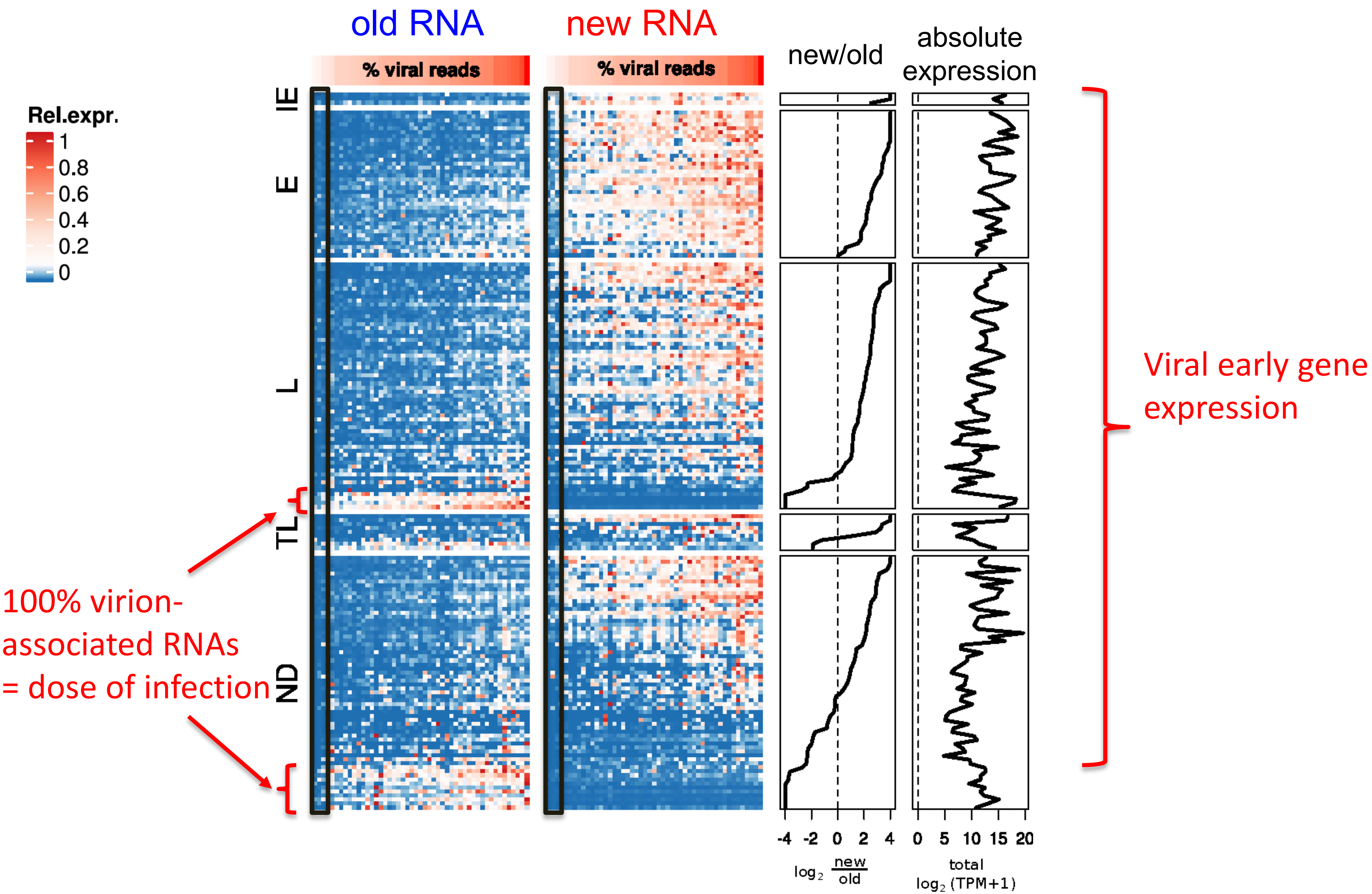






⇒ GRAND-SLAM works!

scSLAM-seq depicts the infection dose for each cell thereby enabling dose-response analysis

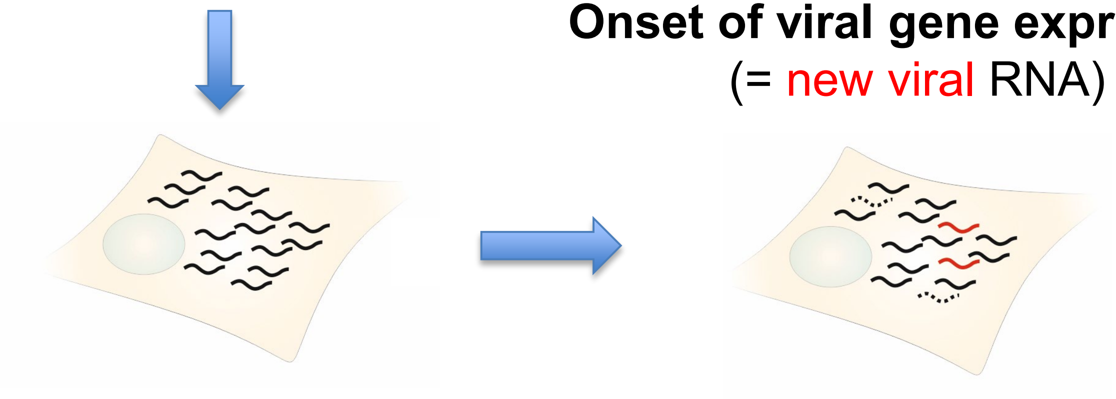


Viral early gene expression

100% virion-associated RNAs = dose of infection

scSLAM-seq adds a temporal dimension to single cell sequencing

Dose of infection per cell
= virion-associated RNA
(= old viral RNA)



Onset of viral gene expression
(= new viral RNA)

Outcome
(= total RNA)

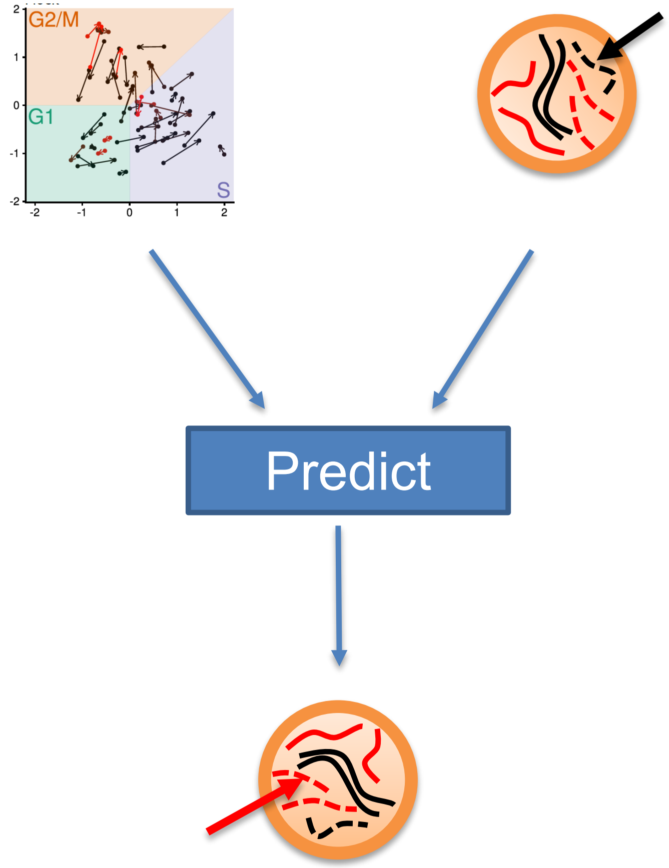
Cellular state prior to infection
(= old cellular RNA)

Cellular response
(= new cellular RNA)

(1) Cell cycle and (2) virus dose reliably predict infection efficiency

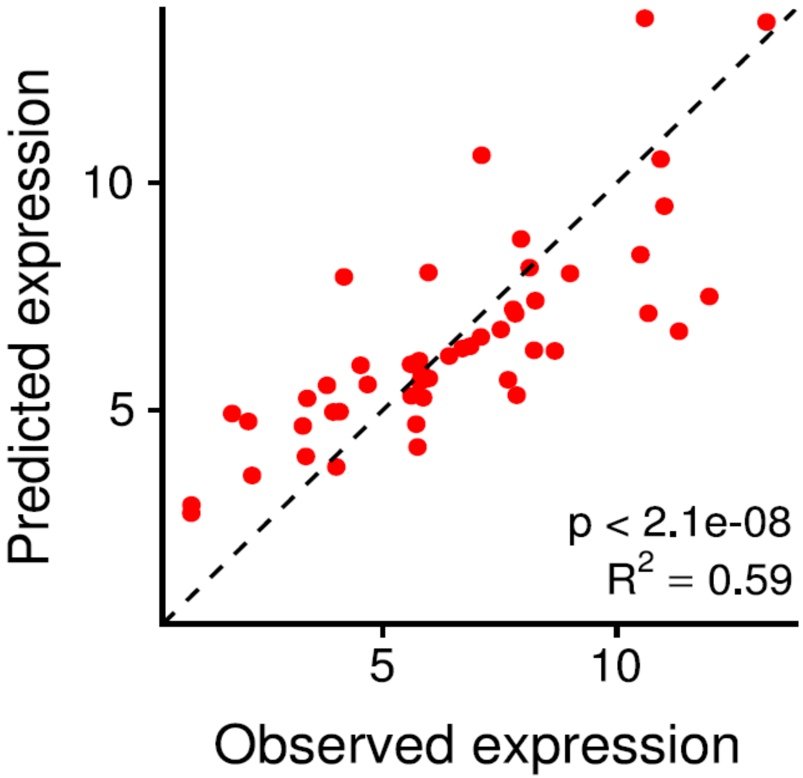
Cell cycle state

Dose of infection
(old viral RNA)

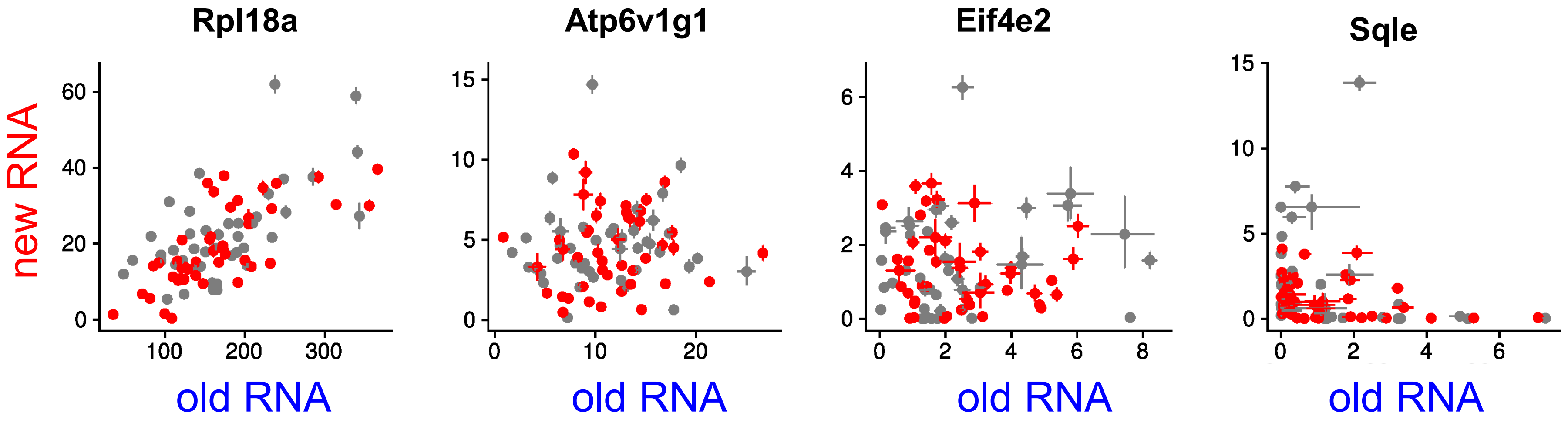


observed new viral gene expression

New viral gene expression

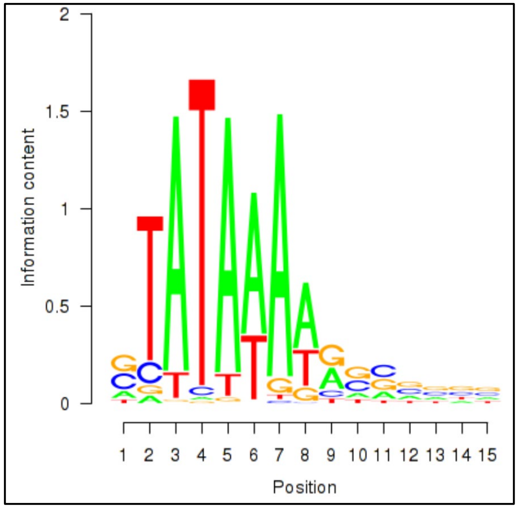


scSLAM-seq visualizes heterogeneity in transcriptional activity (bursts)

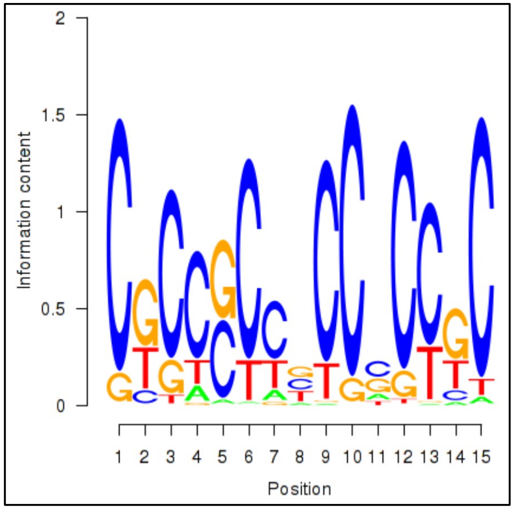


Hypothesis 1:	Heterogeneity reflects cell cycle, oscillation (e.g. NFkB)…
Hypothesis 2:	Transcription occurs in bursts with some promoters subsequently being temporally „non-permissive“ for hours

Promoter analyses reveals two major motifs



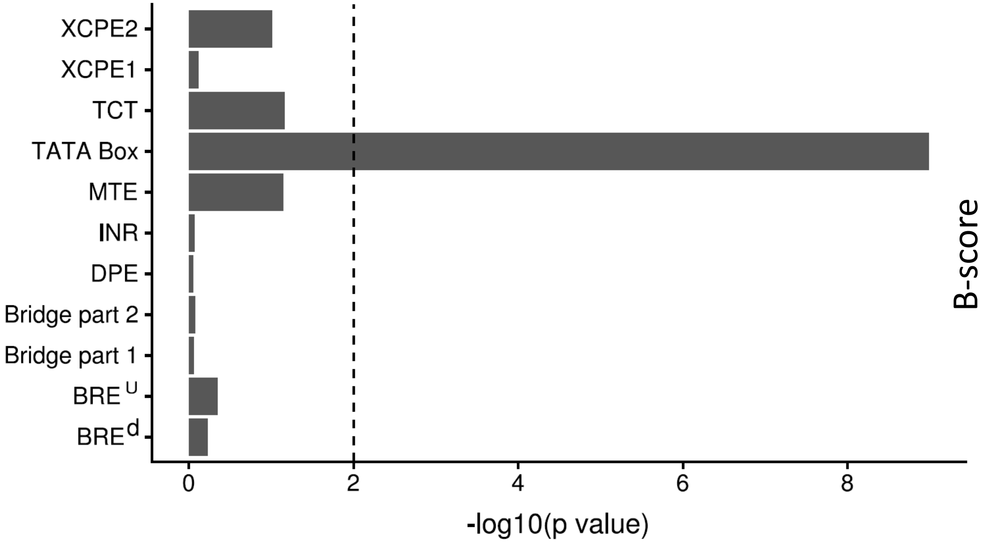
TATA-box



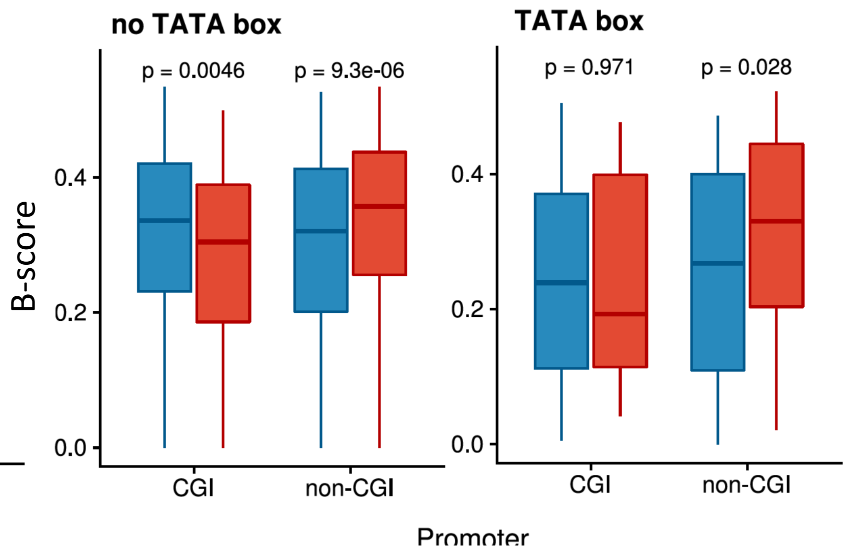
CpG Islands
(DNA methylation)

TATA-boxes (-) and CpG (+) methylation define heterogenious transcription

Strong TATA-box
⇒ contiuously active promoter



non-methylated
methylated

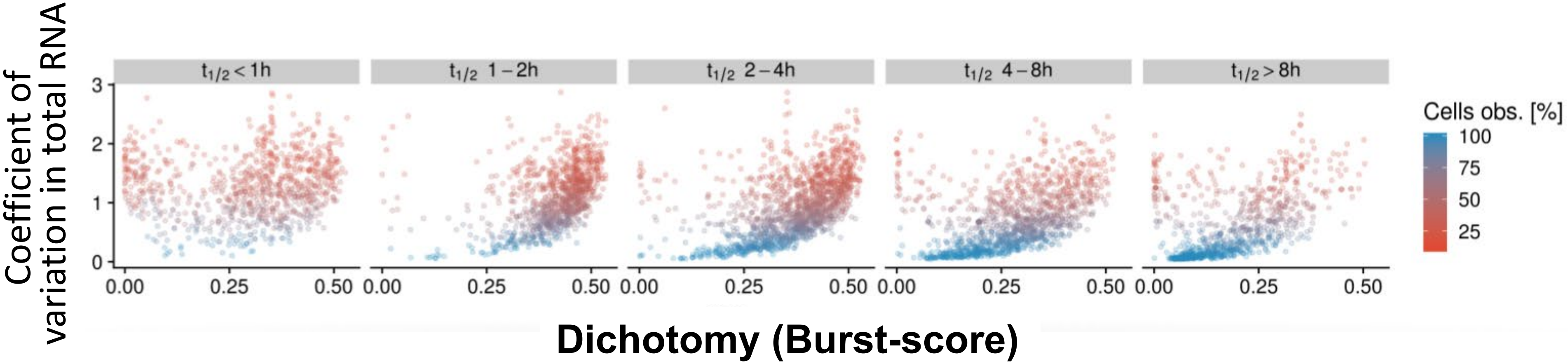


Temporal methylation of
non-CpG island promoters
⇒ Promoter temporally non-permissive

Dichotomous transcription is a gene (promoter)-intrinsic effect!

Dichotomous transcription explains intercellular heterogeneity

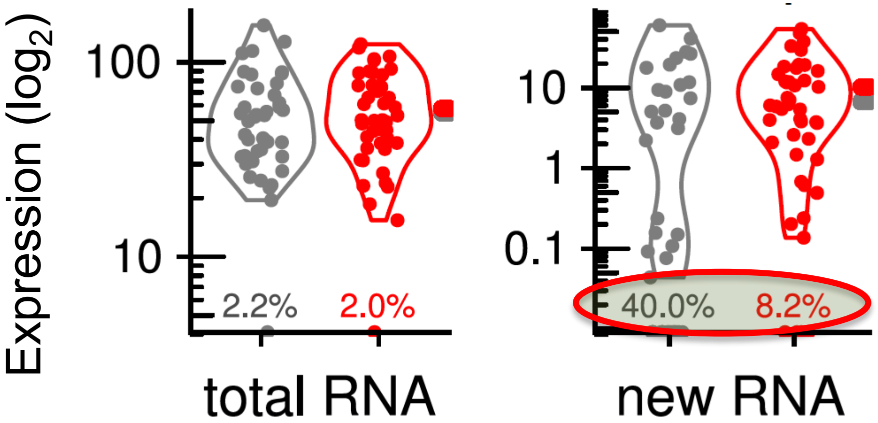
Cellular genes grouped according to mRNA half-life ($t_{1/2}$)

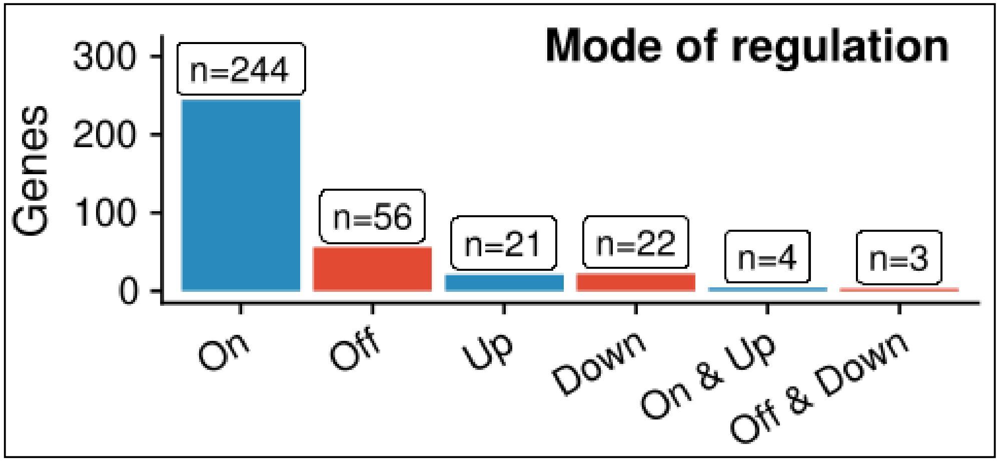


The more „On“-“Off“ is visible in „new“ RNA,
the larger the differences in total RNA levels between cells!

scSLAM-seq depicts „Off-On“ switches in the CMV-induced IFN response

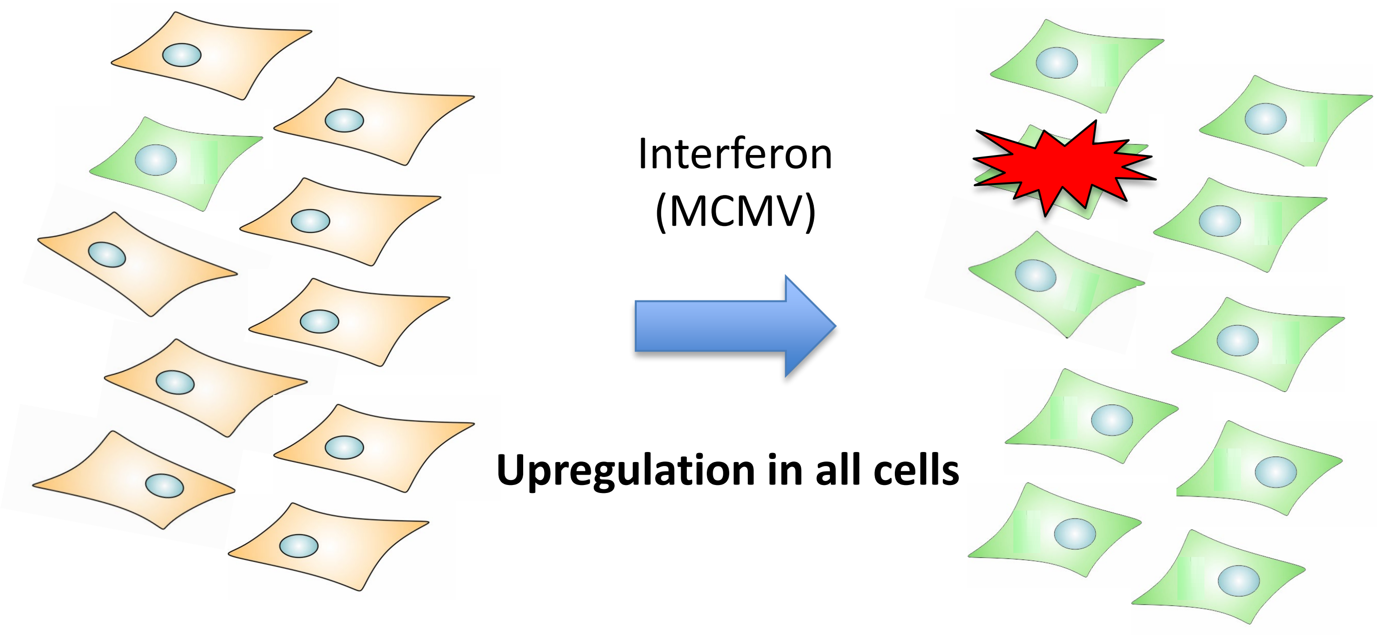
Example of virus-induced „On“ switch (Npc2)





„On-Off“ regulation may enable antiviral protection of all cells, while avoiding hyper-responsiveness

Proposed model

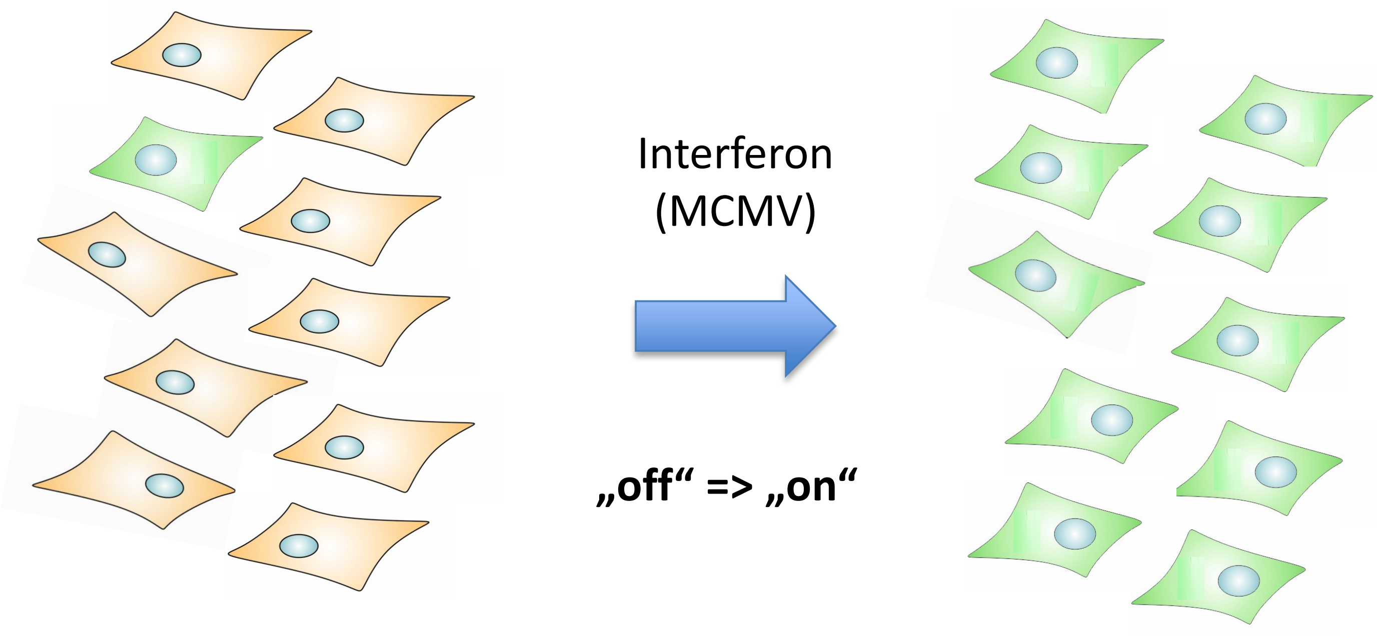


Interferon
(MCMV)

Upregulation in all cells

„On-Off“ regulation may enable antiviral protection of all cells, while avoiding hyper-responsiveness

Proposed model



Interferon
(MCMV)

„off“ => „on“

Systems biology is not about generating large amounts of data but about sharpening the questions!

The spirit of the woods
Sandro Del Prete, 1981

Cover of „Cosmic Encounter“
by Adolfo Sagastme