Skoltech

Skolkovo Institute of Science and Technology

Life Sciences program

Regulation of genome-wide transcription by essential factors that control promoter-proximal RNA polymerase II pausing in human cells.

Student: *Mariia Vlasenok*

Research Advisors: *Konstantin Severinov*

Sergei Nechaev

Yen Lee Loh

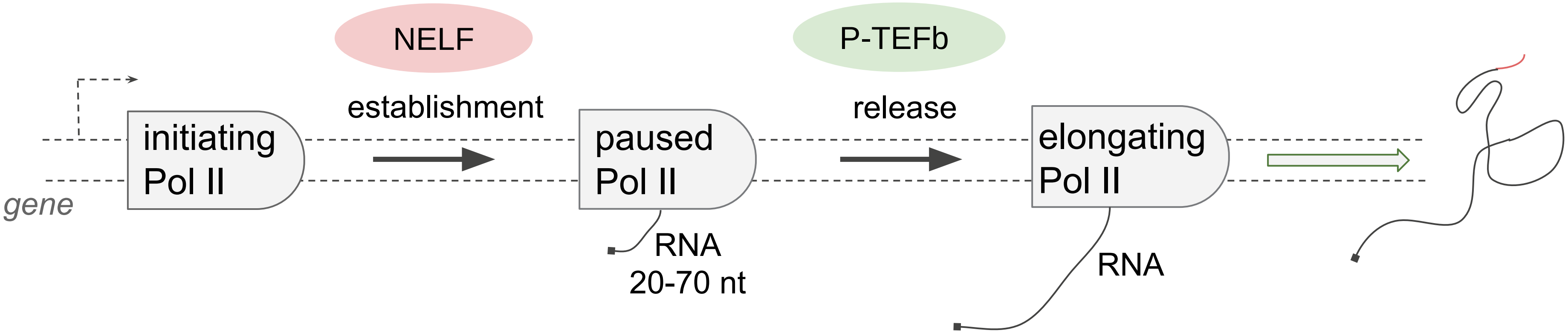
June, 2019

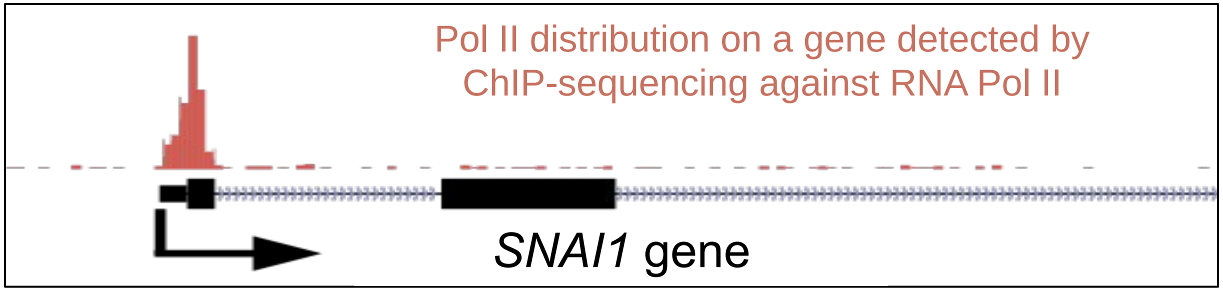
Background

- In eukaryotes, messenger RNA genes are transcribed by the RNA Polymerase II (Pol II).
- The main stages of Pol II transcription are initiation, elongation, and termination.
- Pol II transcription elongation is regulated as tightly as initiation.

Background

- Promoter-proximal Pol II pausing - halt of Pol II less than 100 nucleotides downstream of the start site.
- NELF and P-TEFb control, respectively, pausing establishment and release on every gene.





Pol II distribution on a gene detected by ChIP-sequencing against RNA Pol II

Samarakkody et al, NAR 2015

Background

- Pol II pausing is associated with active, not repressed genes.
- Pausing (and NELF + P-TEFb) is essential in higher eukaryotes.
- Pausing is implicated in organism development, cell differentiation and responses to stimuli, but the mechanisms are unknown.

How can essential factors regulate transcription?

Background

How can essential factors regulate transcription?

Overall idea:

Pausing regulates the distribution of Pol II on gene promoters across the genome.

Working hypothesis:

Limiting the levels of NELF and P-TEFb favors Pol II transcription at stronger promoters at the expense of less active ones.

Significance:

Limiting the levels of essential factors may be a common mechanism organizing transcription into stable genome-wide patterns.

Aims of the project

Aim 1

to analyze transcriptional responses of human cells to heat shock using global run-on sequencing data

bioinformatics

Aim 2

to model transcriptional consequences of NELF depletion in a cell as a closed system

mathematical modeling

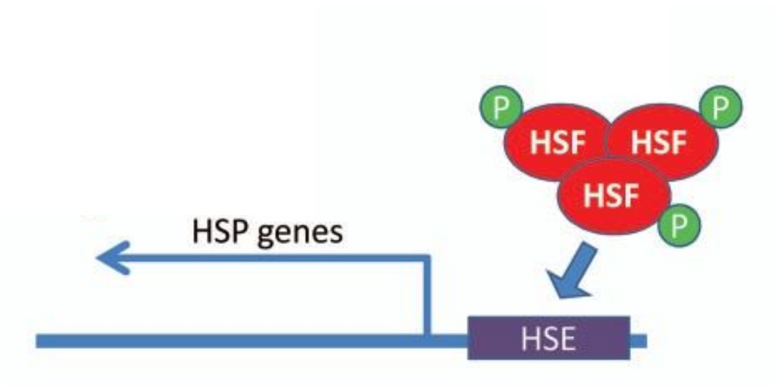
Aim 1

to analyze transcriptional responses of human cells to heat shock using global run-on sequencing data

to compare Heat Shock (HS) response in completely different human cells

- K562 - leukemia cell line
- MCF-7 - breast cancer cell line

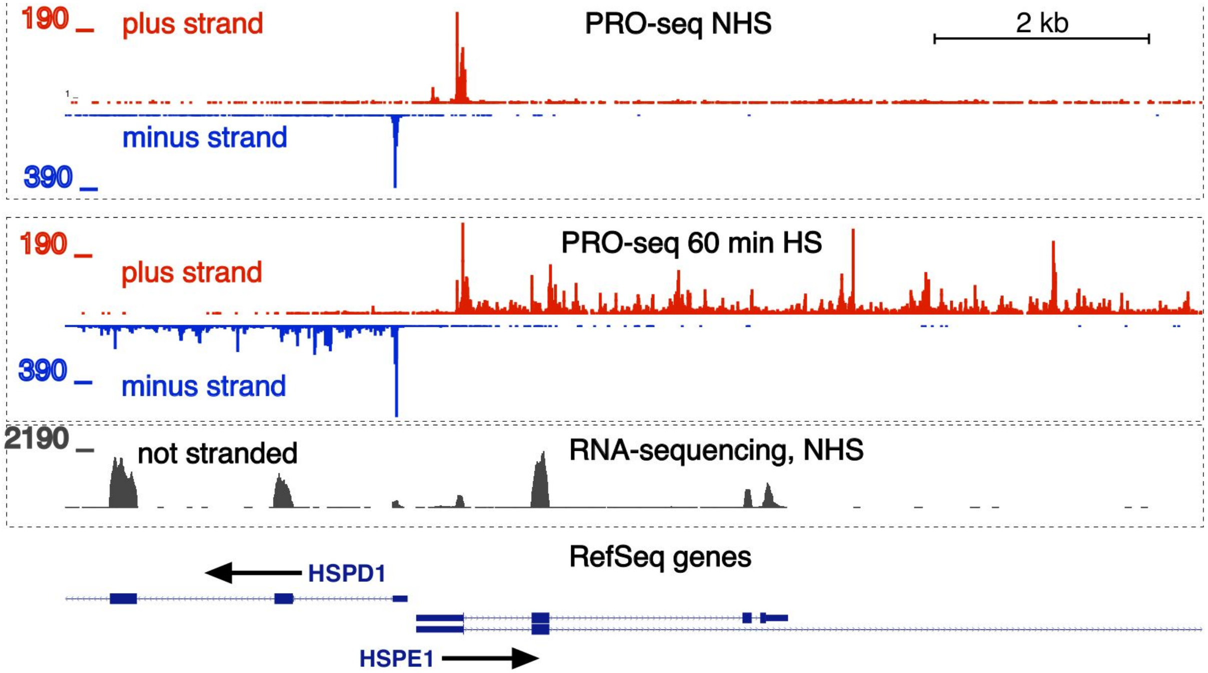
Heat Shock response includes conserved activation by Heat Shock Response factor (HSF)



Aim 1 methods

PRO-seq - Precision nuclear Run-On sequencing

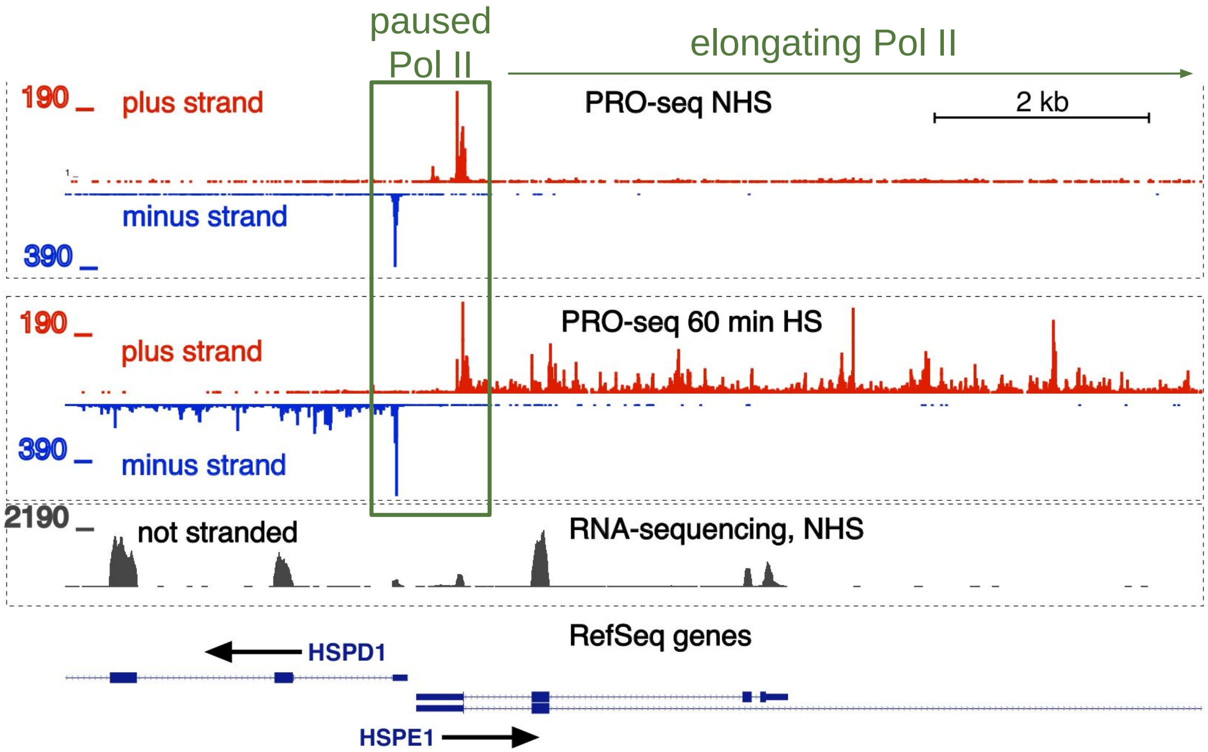
- uses biotin labeled nucleotides to detect nascent transcription
- enables genome-wide mapping of transcriptionally engaged RNA Polymerase with single nucleotide-resolution



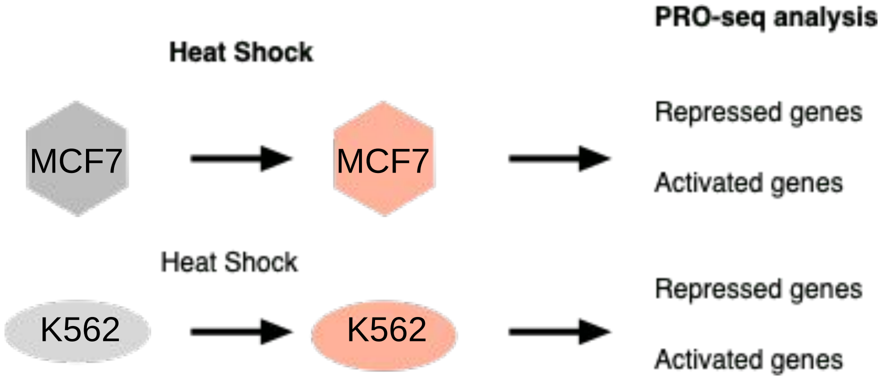
Aim 1 methods

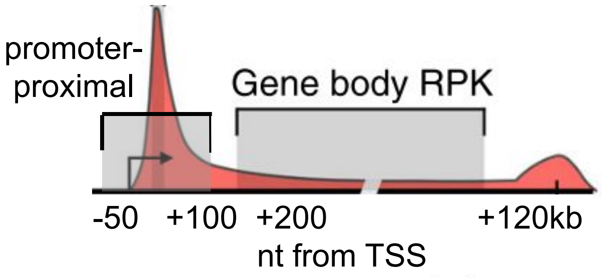
PRO-seq - Precision nuclear Run-On sequencing

- uses biotin labeled nucleotides to detect nascent transcription
- enables genome-wide mapping of transcriptionally engaged RNA Polymerase with single nucleotide-resolution



Aim 1 methods





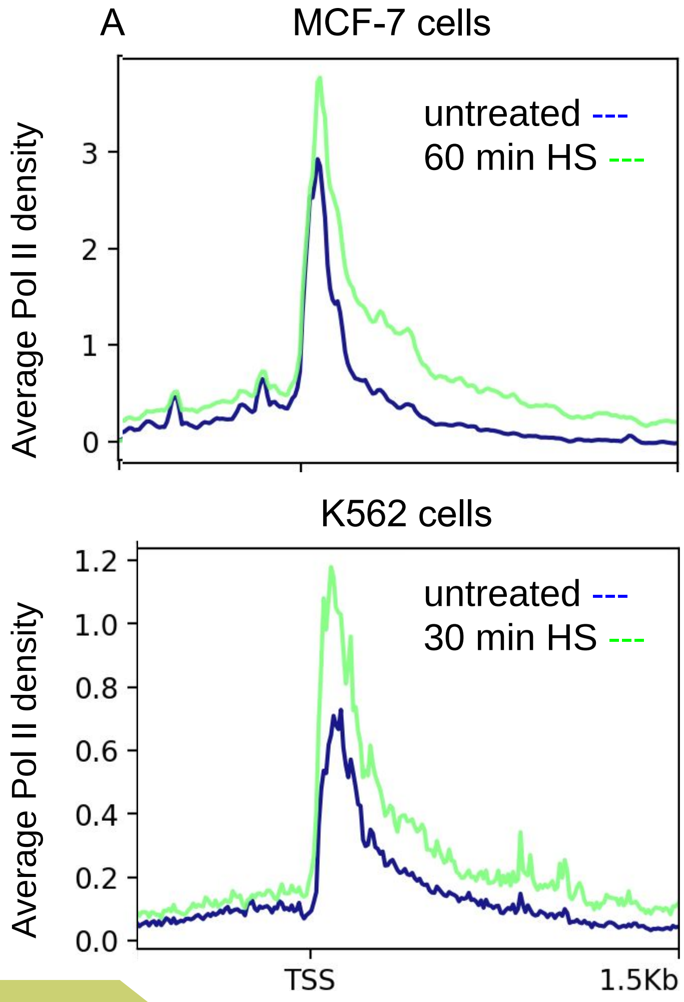
PRO-seq data analysis:

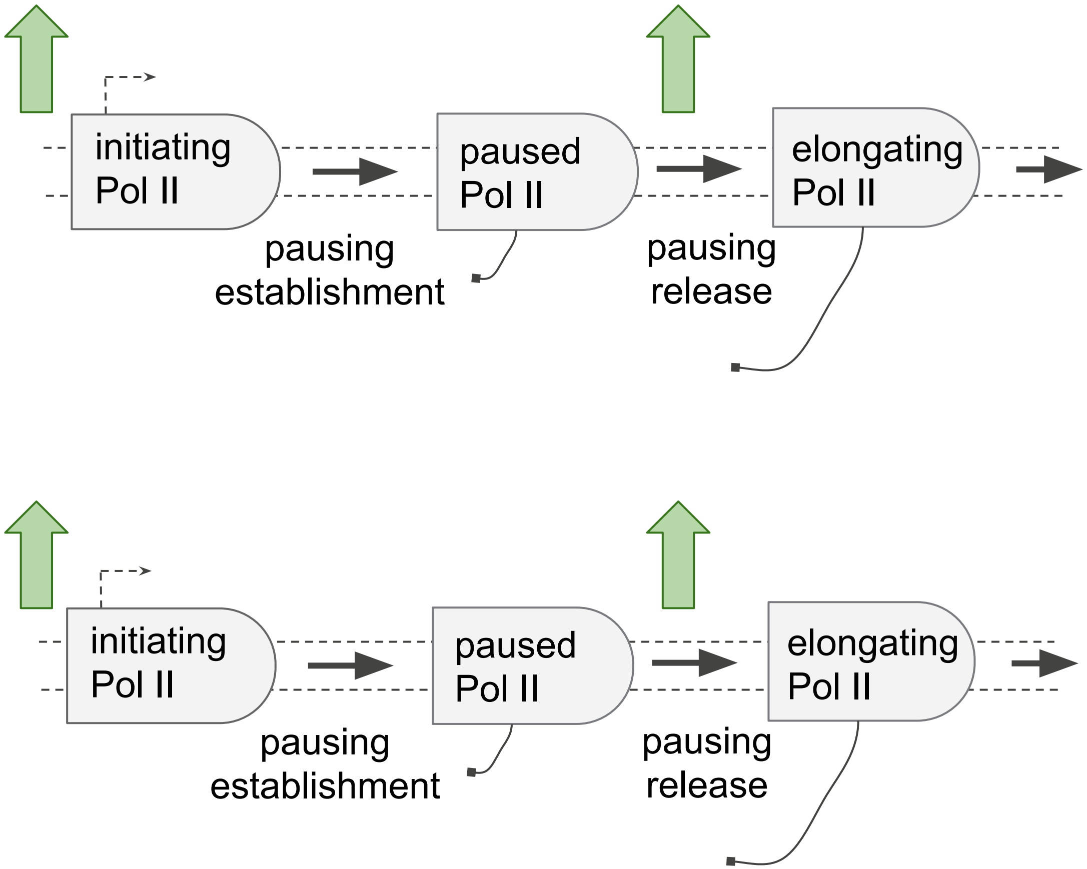
- ***Raw data:*** *FastQC, FASTX-toolkit, hisat2, samtools*
- ***genome arithmetic*** *- bedtools, deeptools*
- ***Normalization***
- ***Differential expression:*** *DESeq2*

Heat Shock response: similar mechanisms of activation

in two cell lines

Differentially expressed genes (p adj < 0.05) activated during HS.





initiating Pol II

paused Pol II

elongating Pol II

pausing establishment

pausing release

initiating Pol II

paused Pol II

elongating Pol II

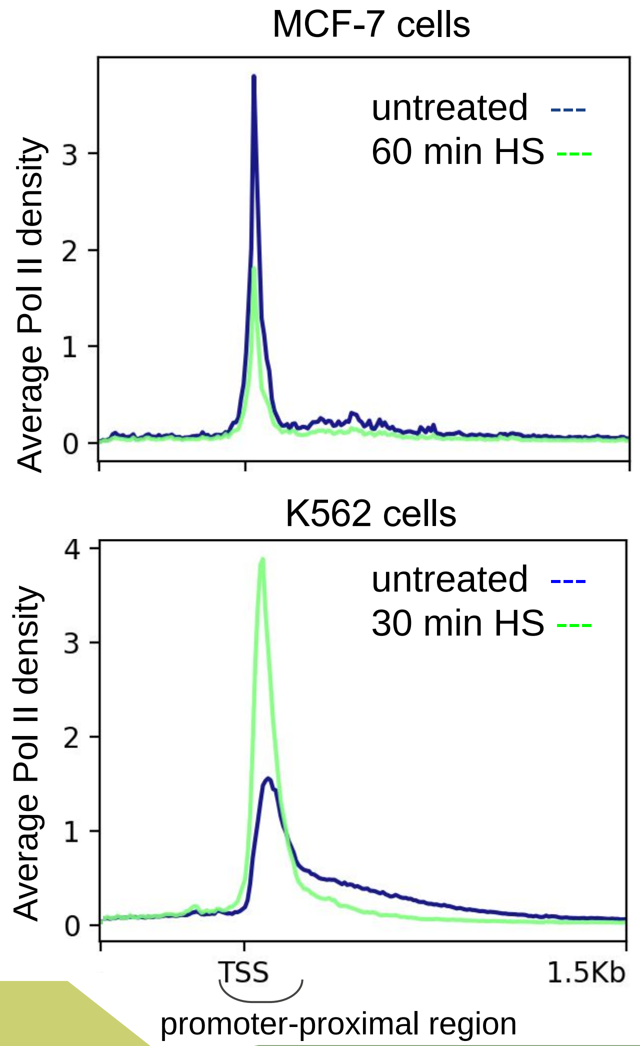
pausing establishment

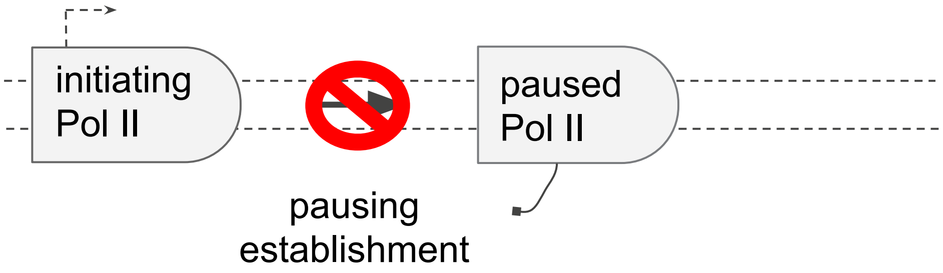
pausing release

HS response: different mechanisms of repression

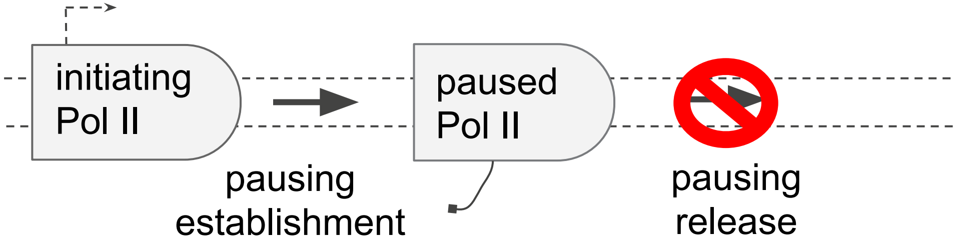
in two cell lines

Genes transcriptionally (p adj < 0.05) repressed during HS.





This mechanism of regulation without initiation repression was observed in both previously studied with pro-seq cell types (MEFs K562)

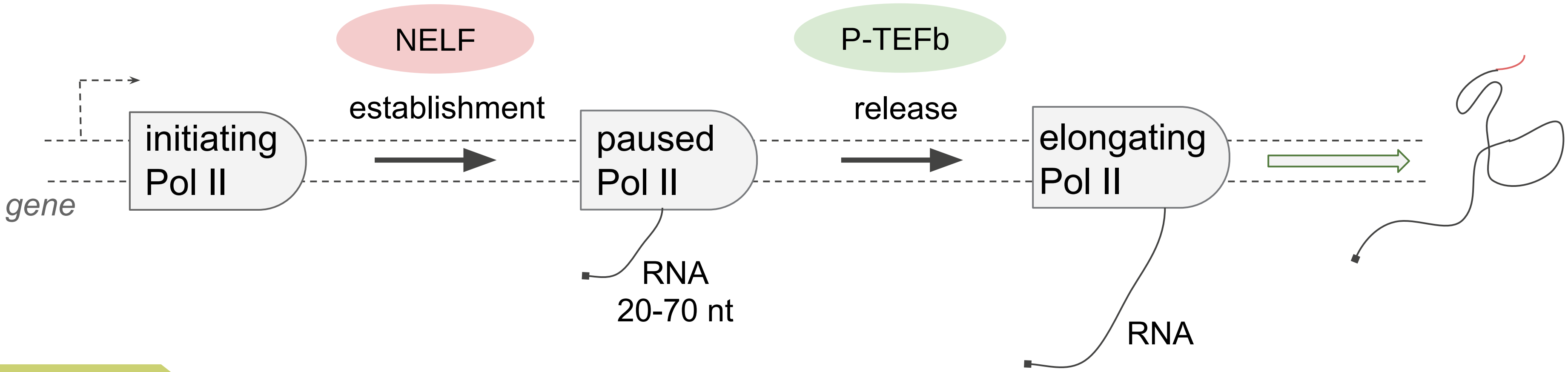


Aim 1 Conclusions

- We compared transcriptional response of distinct two human cell lines to the same stimulus.

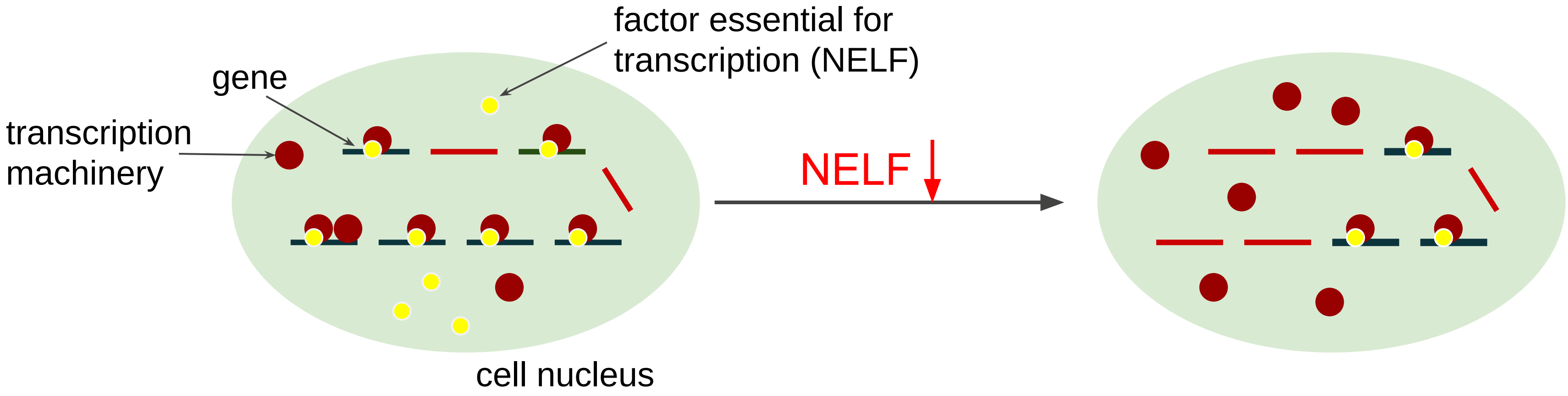
- We find major differences in mechanisms of gene repression. We show that repression can take place either at the level of Pol II recruitment to promoters or Pol II release from pausing.

- **SIGNIFICANCE.** The work provides evidence for transcription regulation by distribution of essential factors across active promoters.



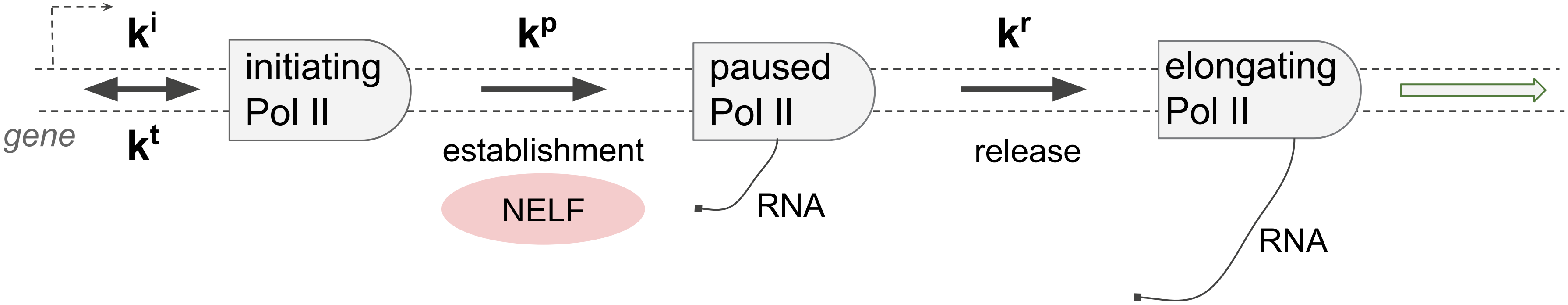
Aim 2

to model transcriptional consequences of NELF depletion in a cell as a closed system



We suggest that the effects of NELF depletion in the cell may be explained by competition of genes for the available essential factor. The number of genes that can be active is limited by the amount of the essential factor.

The chain of reactions used for simulation



for each gene g:

$$\begin{cases} P_g + e \leftrightarrow PIC_g \\ PIC_g + n \rightarrow PEC_g \\ PEC_g \rightarrow P_g + e + n \end{cases}$$

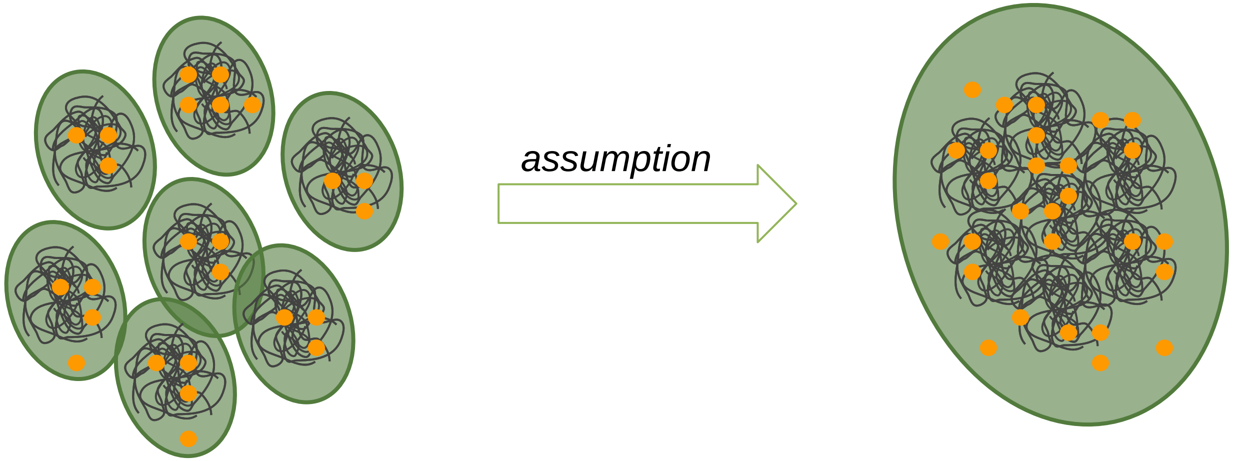
P - gene promoter
e - Pol II
n - NELF
PIC - pre-initiation complex
PEC - paused elongation complex
g - gene (out of 1000)

k^i - initiation frequency i.e.
~ promoter strength

k^p - pausing frequency i.e.
~ NELF preferences

Reaction Rate Equations Approach

RRE works for systems with large numbers of molecules of each species, which is clearly not true for promoters of each gene (p_g). So, we model many cells (nuclei) simultaneously as if they shared all genes and factors.



Reaction Rate Equations approach

steady state simulation

RRE allows to calculate concentrations in equilibrium:

$$\frac{d}{dt}[p_g] = \frac{d}{dt}[PIC_g] = \frac{d}{dt}[PEC_g] = \frac{d}{dt}[e] = \frac{d}{dt}[n] = 0$$

Results in:

We also can use the initial conditions:

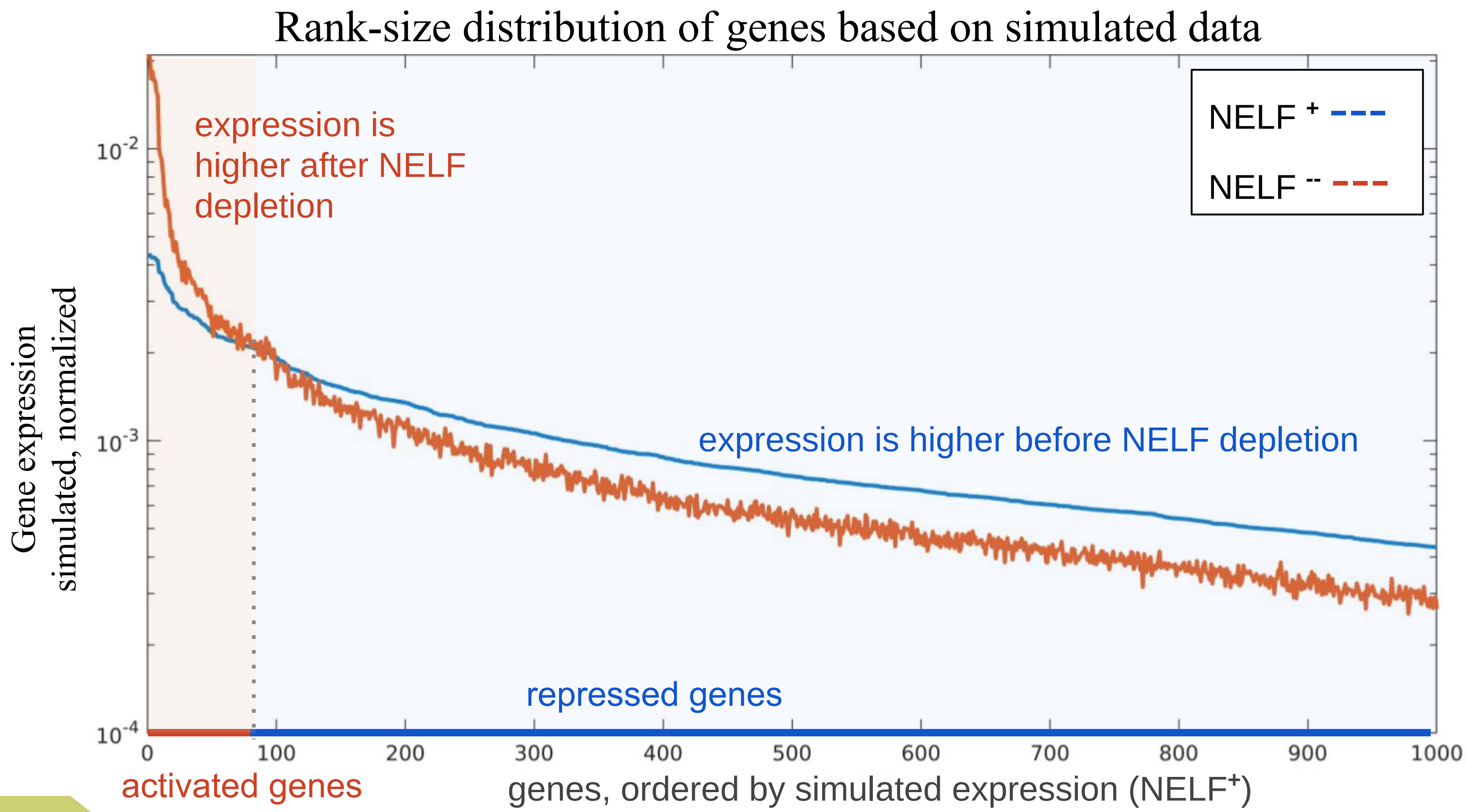
$$\begin{cases} for\ \forall\ g\ < G \\ [e]\frac{k_g^i}{k^t + [n]k_g^p} = \frac{[PIC_g]}{[P_g]} \\ [n]\frac{k_g^p}{k_g^r} = \frac{[PEC_g]}{[PIC_g]} \\ [PEC_g] + [PIC_g] + [P_g] = [P_g]_0 \\ \text{and for the whole system:} \\ [n] + \sum_g [PEC_g] = [n]_0 \\ [e] + \sum_g ([PEC_g] + [PIC_g]) = [e]_0 \end{cases}$$

for each gene g

Results of simulation:

nonlinear transcriptional response to depletion of NELF

The simulation prediction: most active genes are activated, but the rest - repressed.



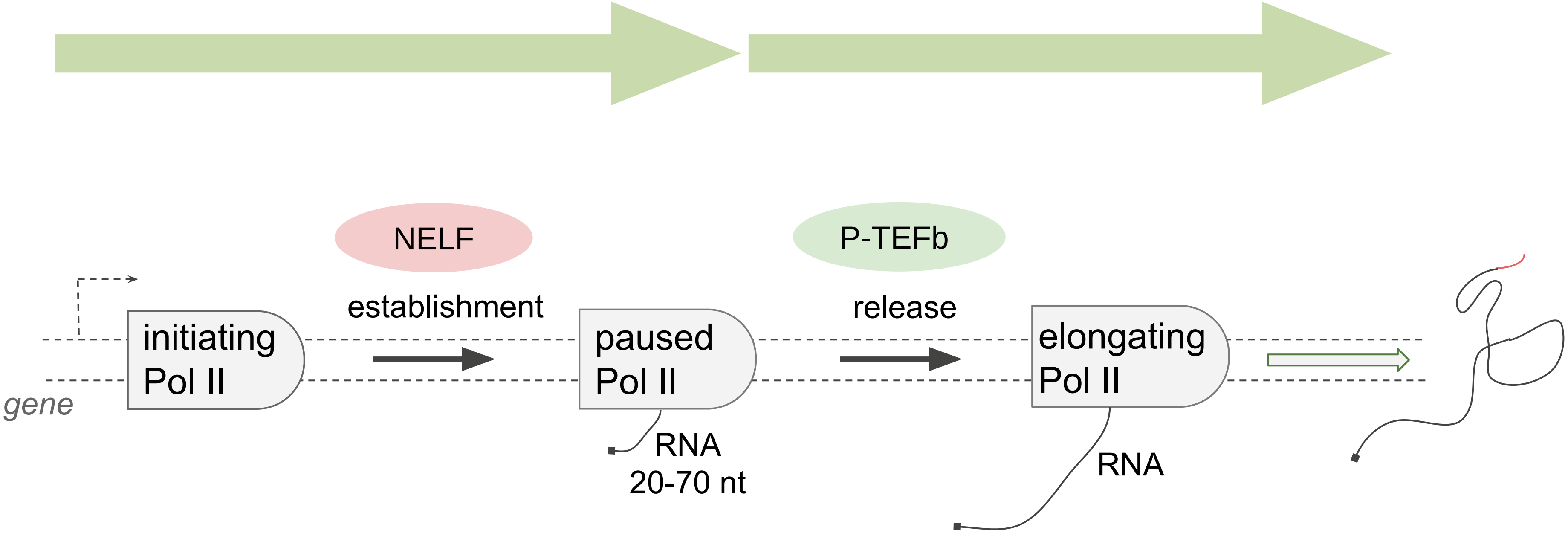
Picked rate constants for simulation based on previous studies

Aim 2 Conclusions

- Limiting the levels of a pausing factor NELF may be sufficient to enforce non-linearity in activity of promoters.
- **SIGNIFICANCE.** Limitation of factors at distinct steps of Pol II pausing may be a universal mechanism that stabilizes transcriptomes in different metazoan cell types.

My work thus far

Future directions



Acknowledgements

Skoltech
Skolkovo Institute of Science and Technology

Research advisors:

- *Konstantin Severinov, Skoltech*
- *Sergei Nechaev,*
  University of North Dakota School of Medicine
- *Yen Lee Loh, University of North Dakota*

Nechaev lab members:

- *Nii Koney-Kwaku Koney*
- *Sayantani Ghosh Dastidar*
- *Bo Lauckner*

Funding:
Skoltech Education to MV,
NSF CAREER 1750379 to SN

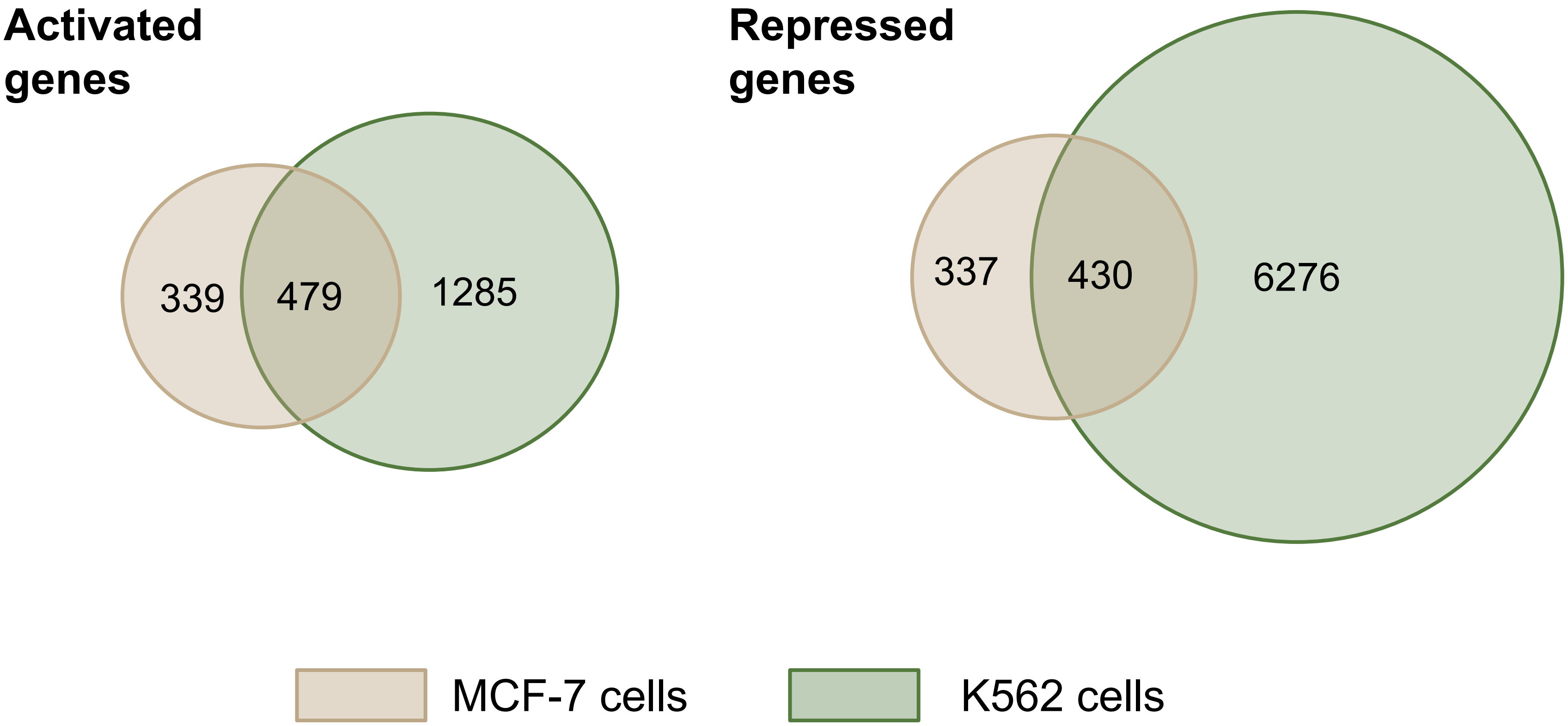
Thank you!

June, 2019

Student: *Mariia Vlasenok*

Heat-induced repression is much more extensive in K562 than in MCF-7 cells

Differentially expressed genes


Activated genes

Repressed genes

Reaction Rate Equations Approach

- **strengths**
  - “textbook” model for chemical kinetics
  - the least complex one that can account for thousands of genes
  - convenience when working with systems in equilibrium state
- **weakness**

  RRE works for systems with large numbers of molecules of each species, which is clearly not true for promoters of each gene (p_g).

Chemical Master Equations Approach

- **strengths**
  - cellular processes involve extremely small population sizes, where it is unrealistic to think in terms of concentration
  - vital when system exhibits bistability
- **weakness**
  - computational complexity

Imply constants k^p and k^i from experimental data

experimental data:

$$[PIC_g] + [PEC_g] = promoter\ \ occupancy$$

$$r_{release} = [PEC_g]k^r_g = [n]k^p_g[PIC]_g = expression$$

initial system:

$$(k^t + [n]k^p_g)[PIC_g] = e[P_g]k^i_g$$

$$[n]k^p_g[PIC]_g = [PEC_g]k^r_g$$

initial conditions:

$$[PEC_g] + [PIC_g] + [P_g] = [P_g]_0$$

$$[n] + \sum_g [PEC_g] = [n]_0$$

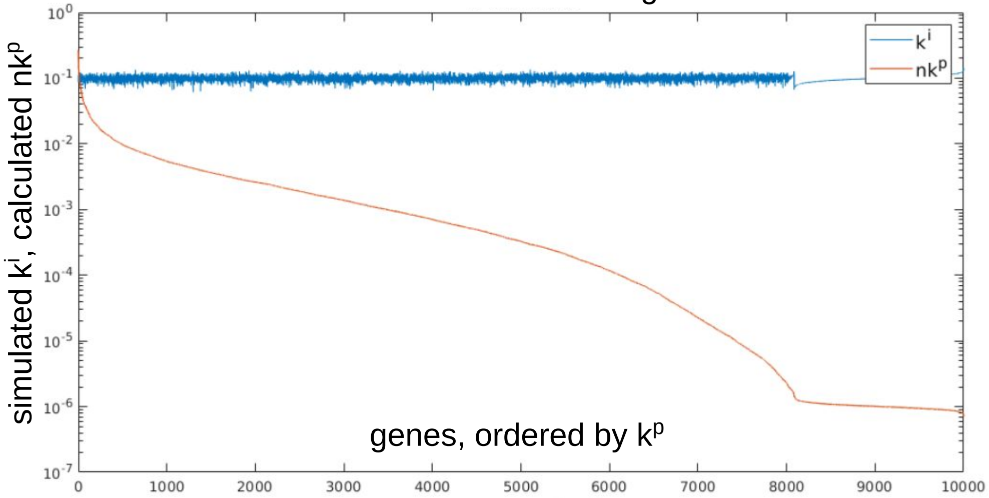
$$[e] + \sum_g ([PEC_g] + [PIC_g]) = [e]_0$$

System with *2g+2* variables

$$g+2 \left\{ \begin{array}{l} expr_g(nk^p_g + k^t) = nk^p_g(cells - prom_g)ek^i_g \\ e + \sum prom_g = e_0 \\ n + \sum prom_g - \sum \frac{expr_g}{nk^p_g} = n_0 \end{array} \right.$$

2g+4 variables and *2(g+2)* equations with NELF^{--} data.

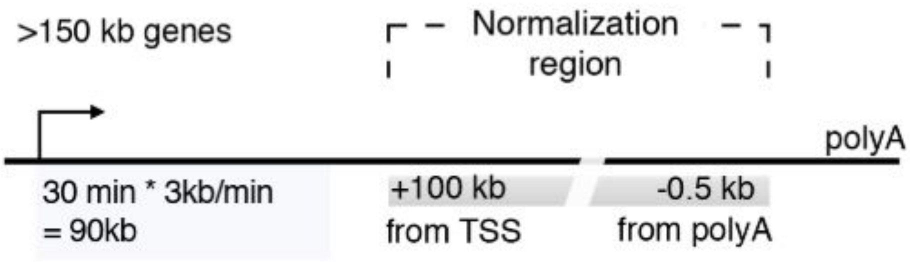
rank-size distribution of genes



k^p, k^i and k^t don't change after NELF depletion; *n*, *n_0* and *e* do.

Appendix *pipeline*

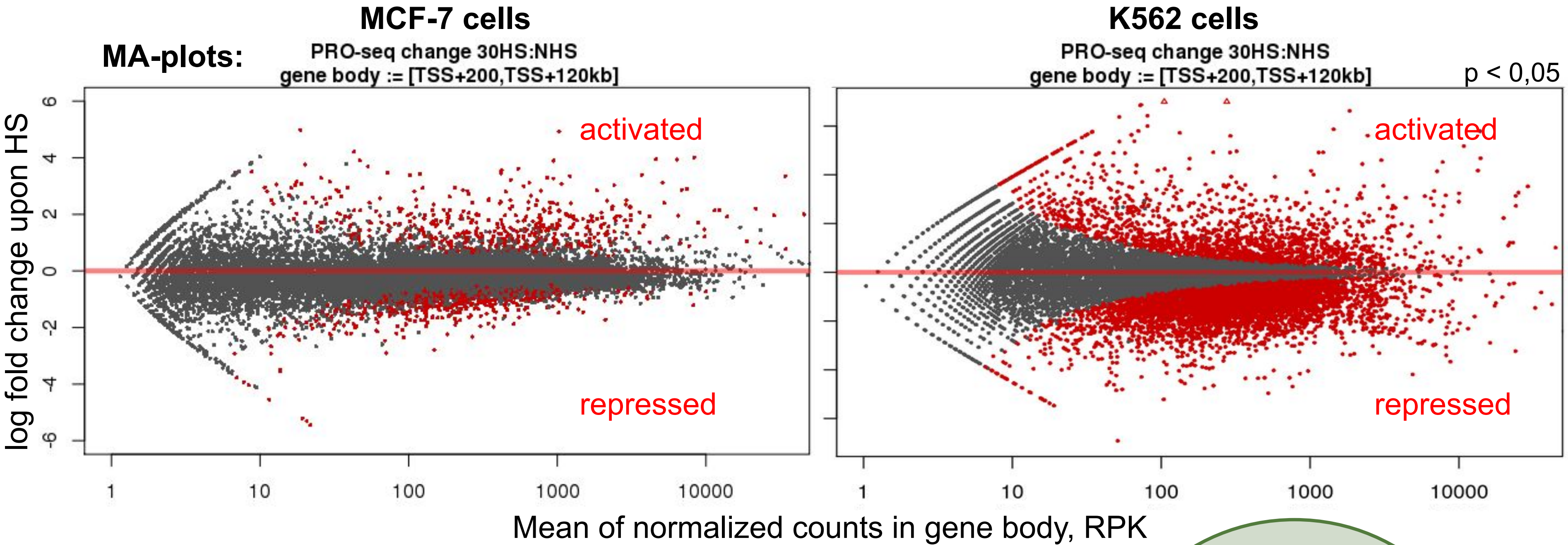
1. trim adaptor (3’ TGGAATTCTCGGGTGCCAAGG) and trim reads to min length of 15bp
2. exclude reads that map to ribosomal RNA (*hisat2)*
3. alignment to hg19 using *hisat2* (uniquely aligned reads with no more than 2 mismatches)
4. read counts for gene bodies
   a. 23k genelist
   b. gene body is considered to start at TSS+200bp and end at TSS+120kb or gene end, whichever is closer
   c. *bedtools intersect* were used to count overlaps
5. Normalization using 3’ ends of long genes
   a. used genes > 150kb
   b. normalization region: from TSS + 100kb to TTS - 0.5kb
6. DESeq2 on gene body read densities (reads per 10kb)
   a. p-value < 0.05
   b. no FC cutoff



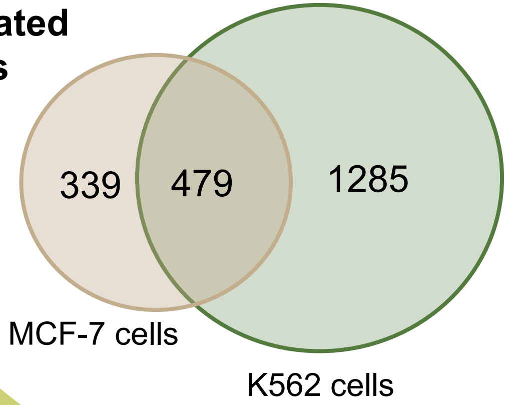
Vihervaara et al, 2017.

Heat Shock response includes conserved activation and variable repression

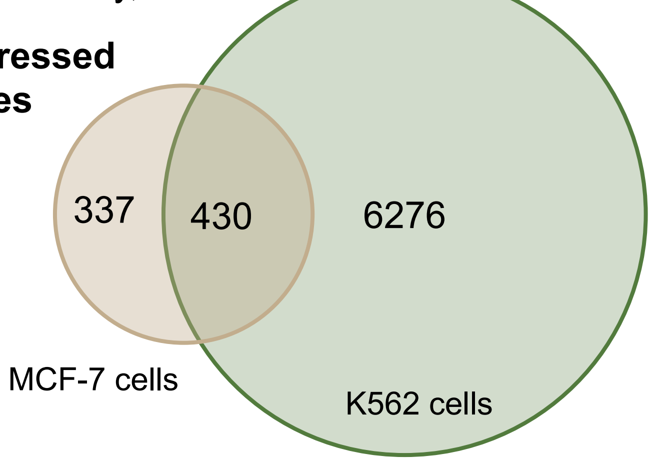
MA-plots:



Activated genes



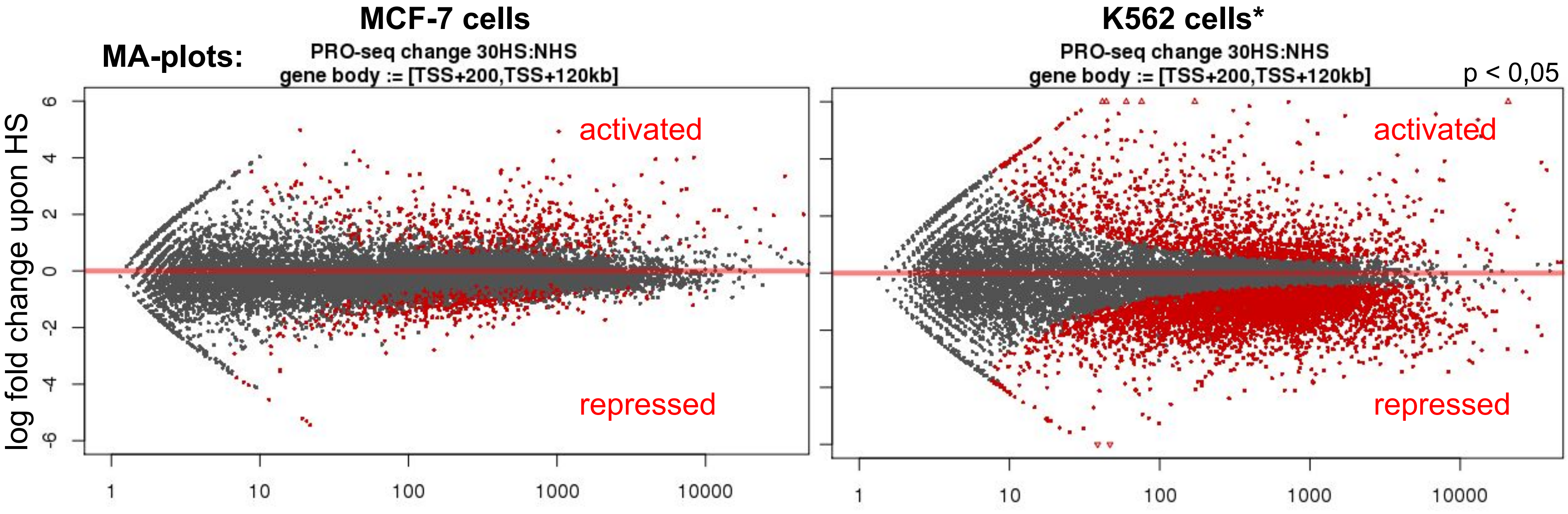
Repressed genes



Repressed genes differ between cell lines

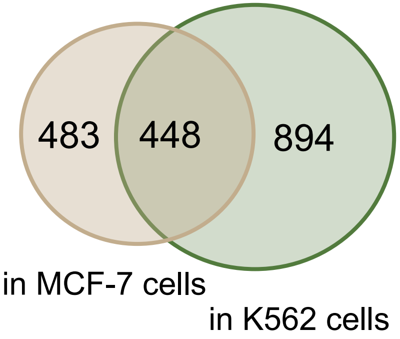
Skoltech

Skolkovo Institute of Science and Technology

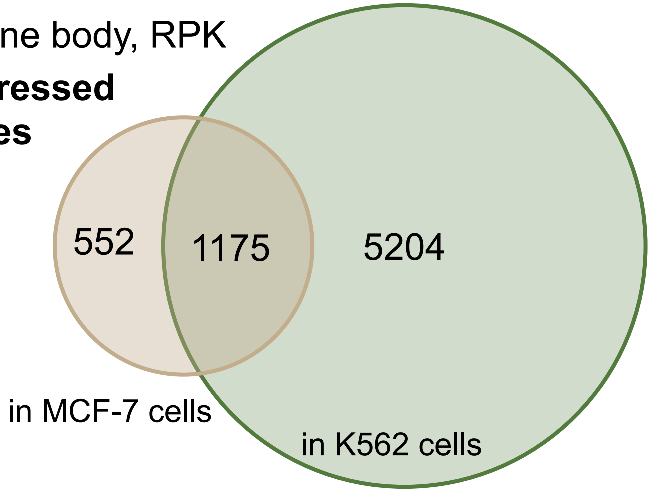


Mean of normalized counts in gene body, RPK

Activated genes



Repressed genes



** K562 data re-analyzed from Vihervaara et al, 2017. Analysis results are consistent.*

Reaction Rate Equations Approach

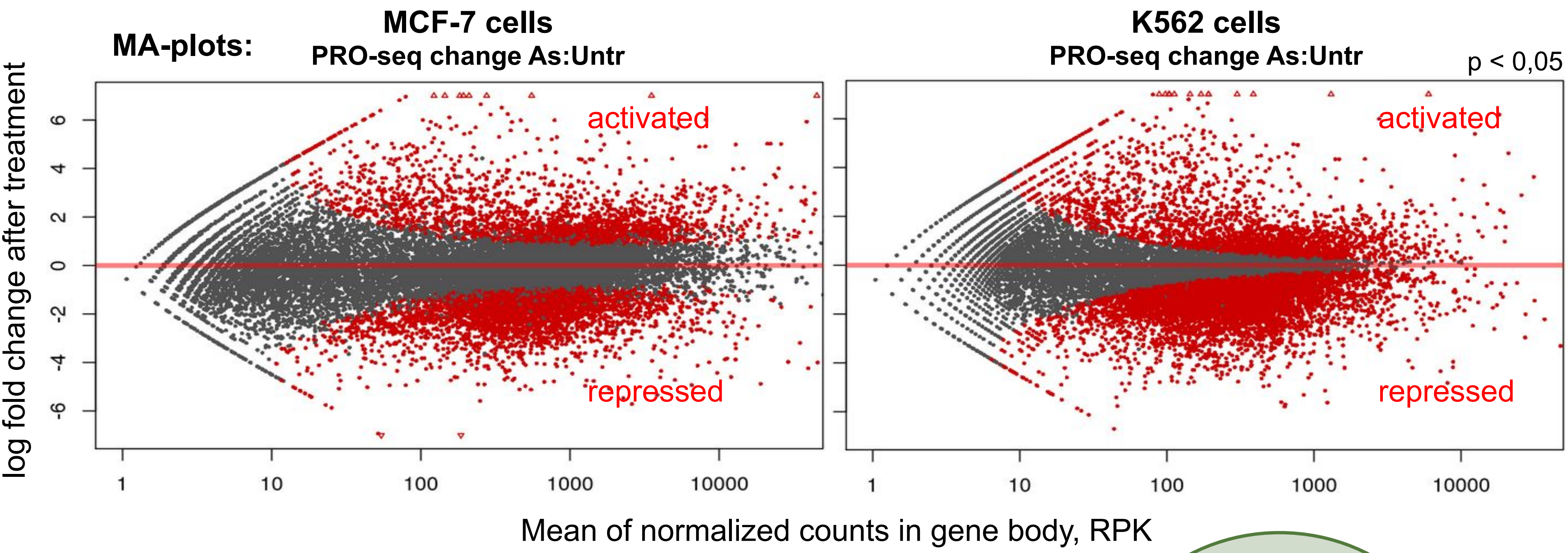
$$\begin{cases} P_g + e \leftrightarrow PIC_g \\ PIC_g + n \rightarrow PEC_g \\ PEC_g \rightarrow P_g + e + n \end{cases}$$

By the law of mass action

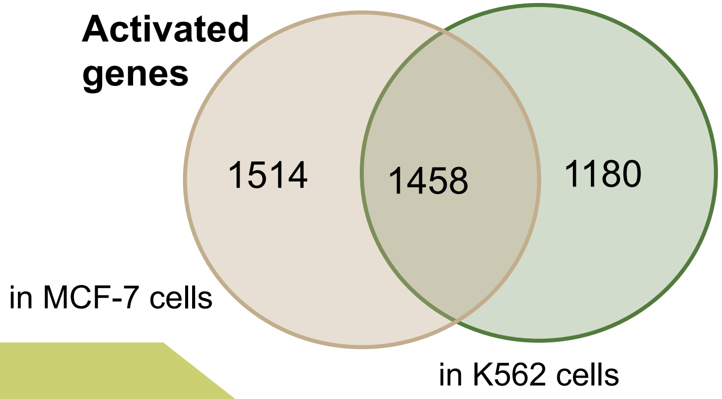
$$\begin{cases} for\ \forall\ g\ \ < G: \\ \frac{d}{dt}[p_g] = -k_g^i[p_g][e] + k^t[PIC_g] + k_g^r[PEC_g] \\ \frac{d}{dt}[PIC_g] = k_g^i[p_g][e] - k^t[PIC_g] - k_g^p[PIC_g][n] \\ \frac{d}{dt}[PEC_g] = k_g^p[PIC_g][n] - k_g^r[PEC_g] \\ \text{and for the whole system:} \\ \frac{d}{dt}[e] = \sum_g(-k_g^i[p_g][e] + k^t[PIC_g] + k_g^r[PEC_g]) \\ \frac{d}{dt}[n] = \sum_g(-k_g^p[n][PIC_g] + k_g^r[PEC_g]) \end{cases}$$

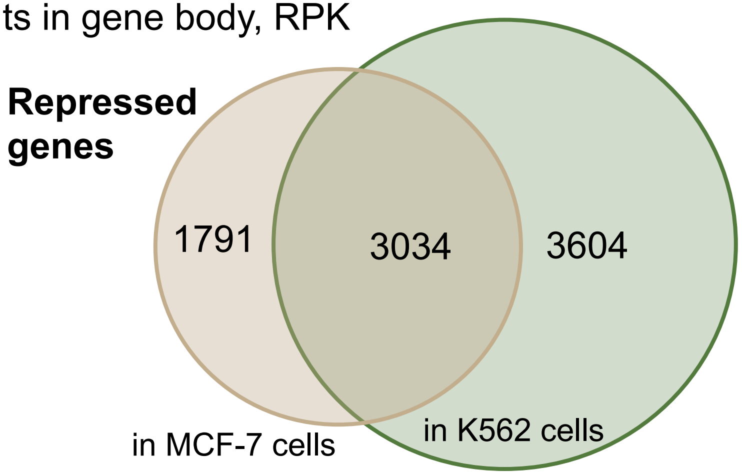
As treatment

MA-plots:


MCF-7 cells
PRO-seq change As:Untr

K562 cells
PRO-seq change As:Untr





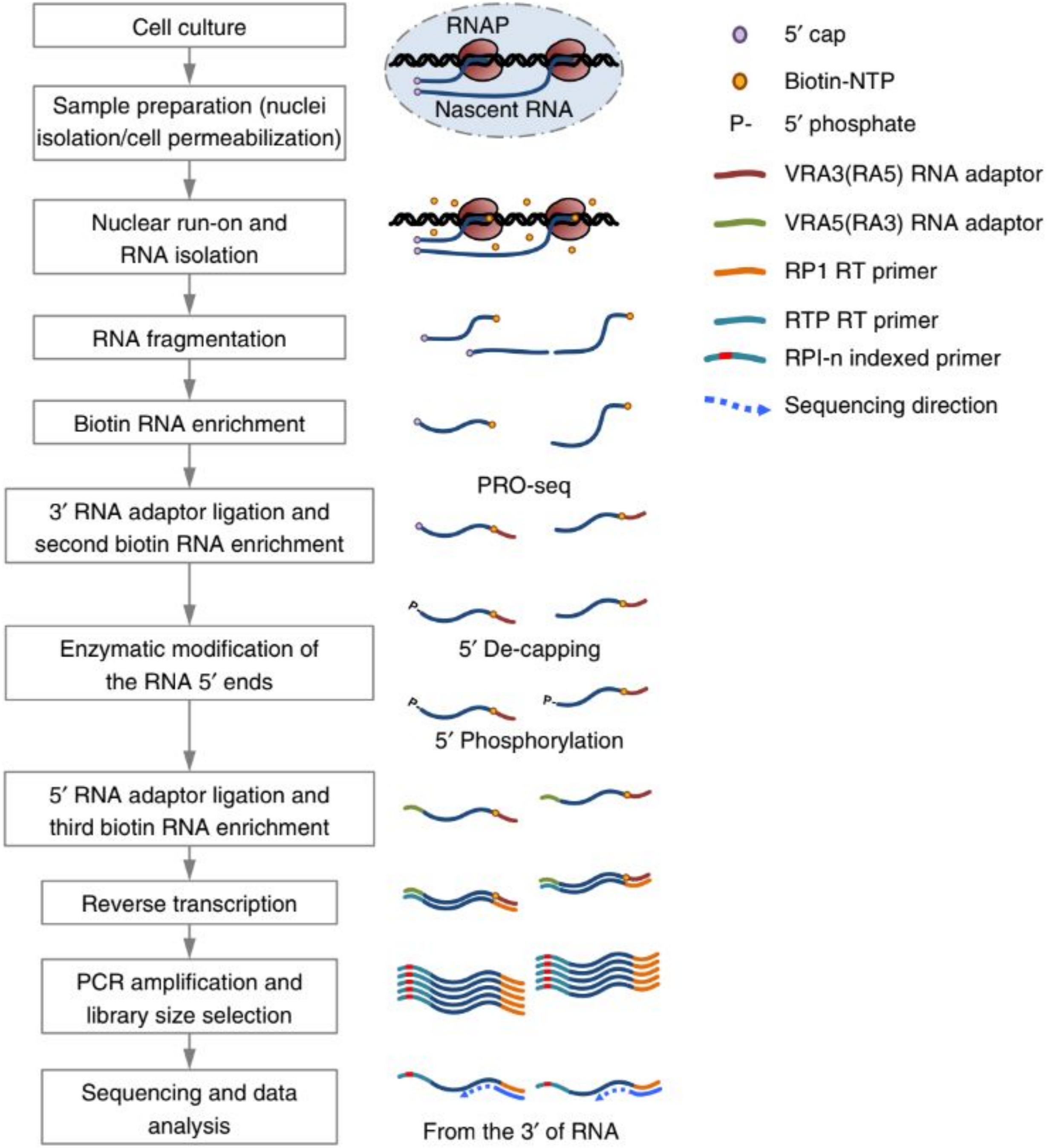
PRO-seq

map the location of active RNA polymerases (PRO-seq)

- *genome-wide*
- *strand-specific*
- *single nucleotide-resolution*

It is used for studying short-time transcriptional responses.

Nuclei are isolated from cells and in vitro transcriptionally engaged RNA polymerases incorporate one biotin-NTPs into the 3' end of nascent RNA. The biotin-labeled nascent RNA is used to prepare sequencing libraries, which are sequenced from the 3' end to provide high-resolution positional information for the RNA polymerases.



Cell culture

Sample preparation (nuclei isolation/cell permeabilization)

Nuclear run-on and RNA isolation

RNA fragmentation

Biotin RNA enrichment

3′ RNA adaptor ligation and second biotin RNA enrichment

Enzymatic modification of the RNA 5′ ends

5′ RNA adaptor ligation and third biotin RNA enrichment

Reverse transcription

PCR amplification and library size selection

Sequencing and data analysis