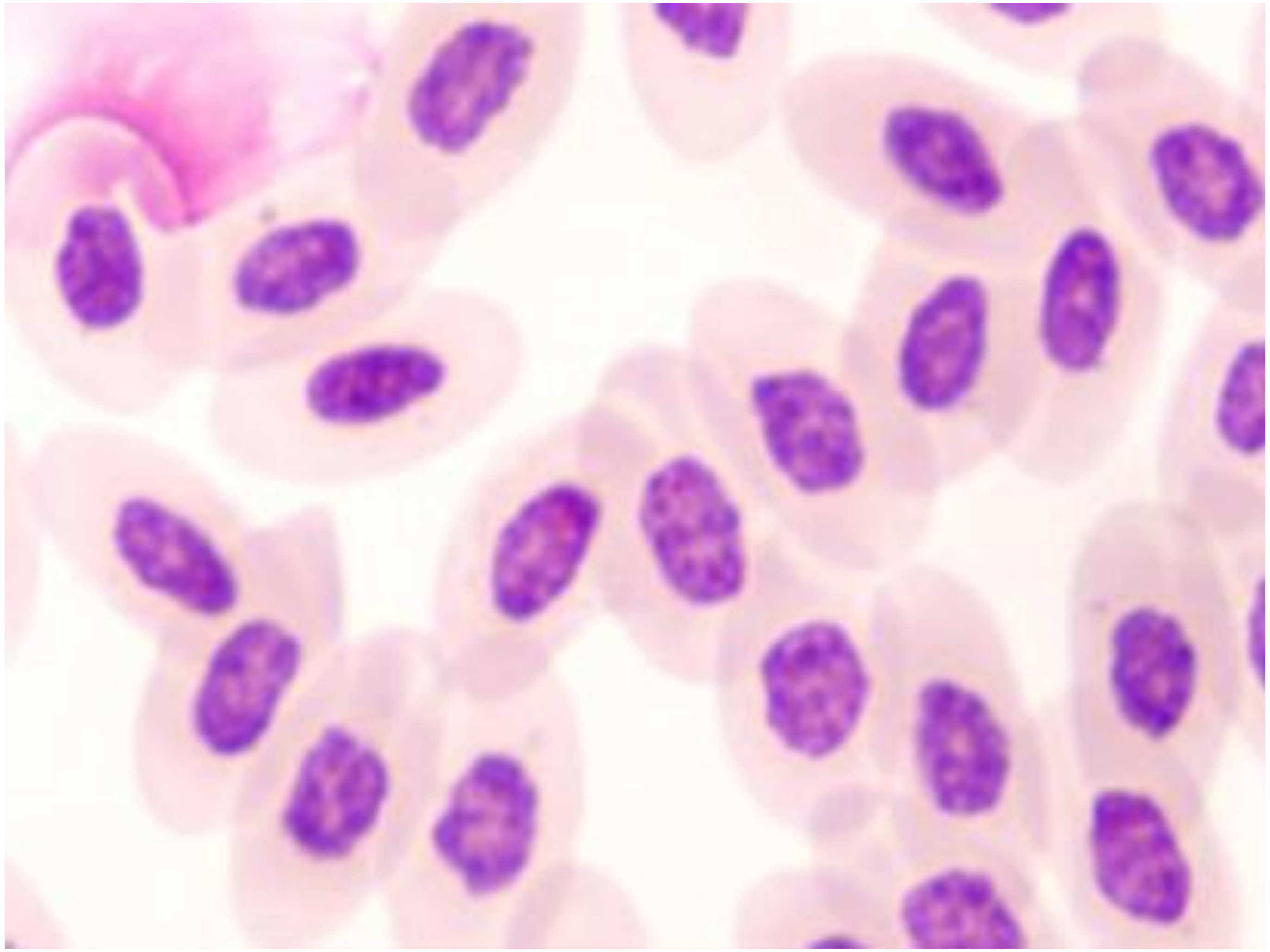
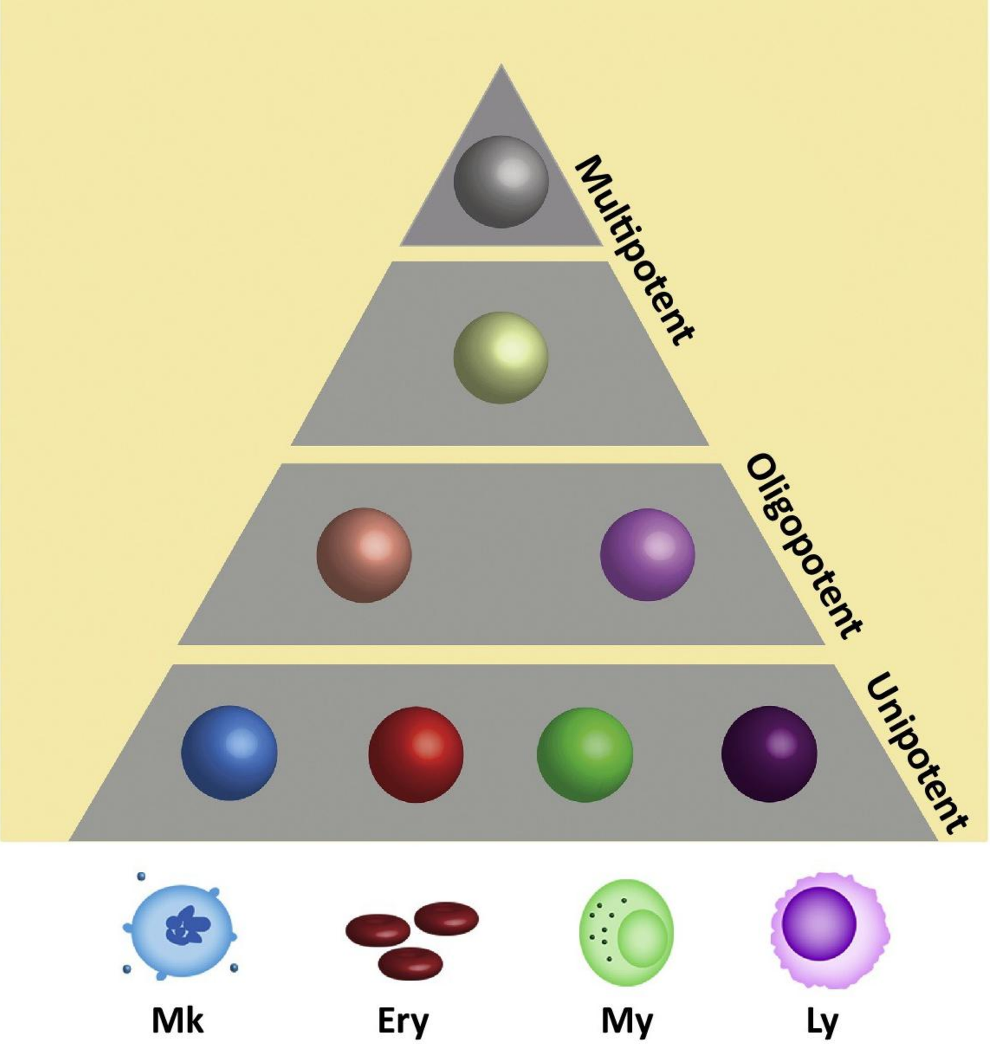
Inferring mechanistic gene regulatory networks from single cell data: a case study on erythropoiesis.

Arnaud Bonnaffoux (Vidium)
and
Olivier Gandrillon (LBMC)

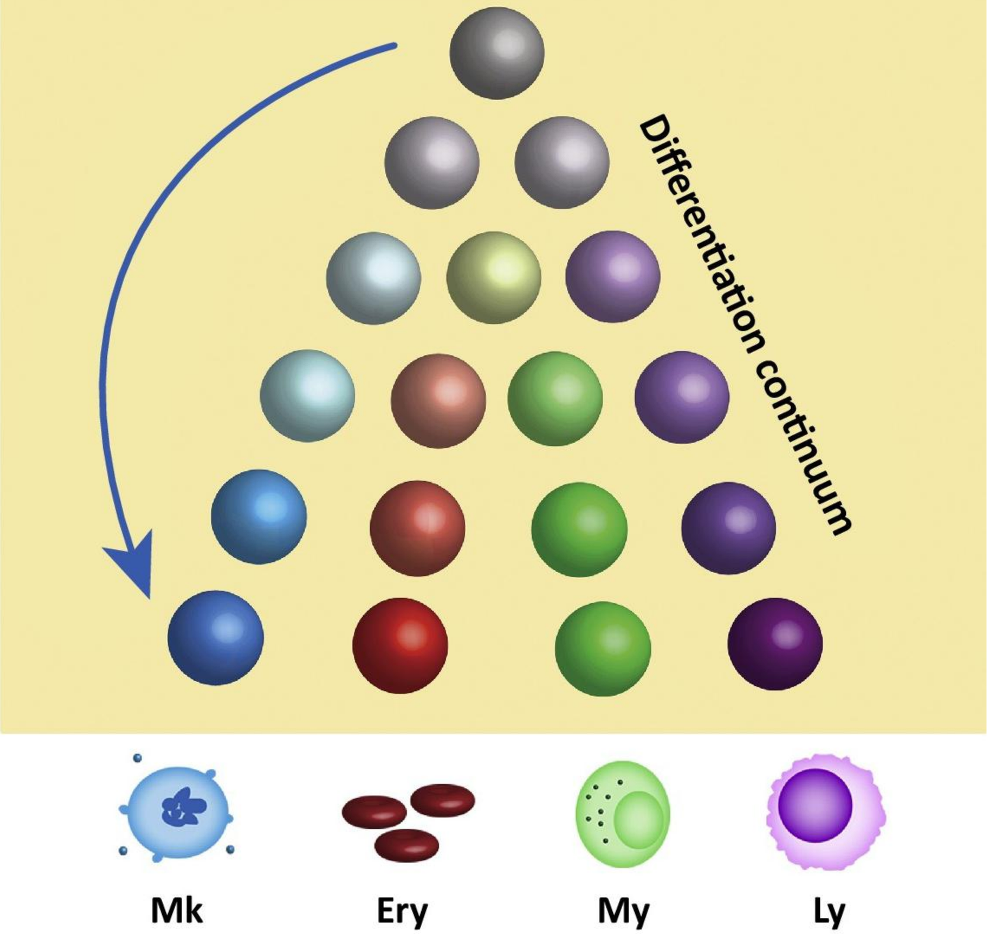
Erythropoiesis: the generation of erythrocytes (red blood cells) from bone marrow-harboured haematopoietic stem cells



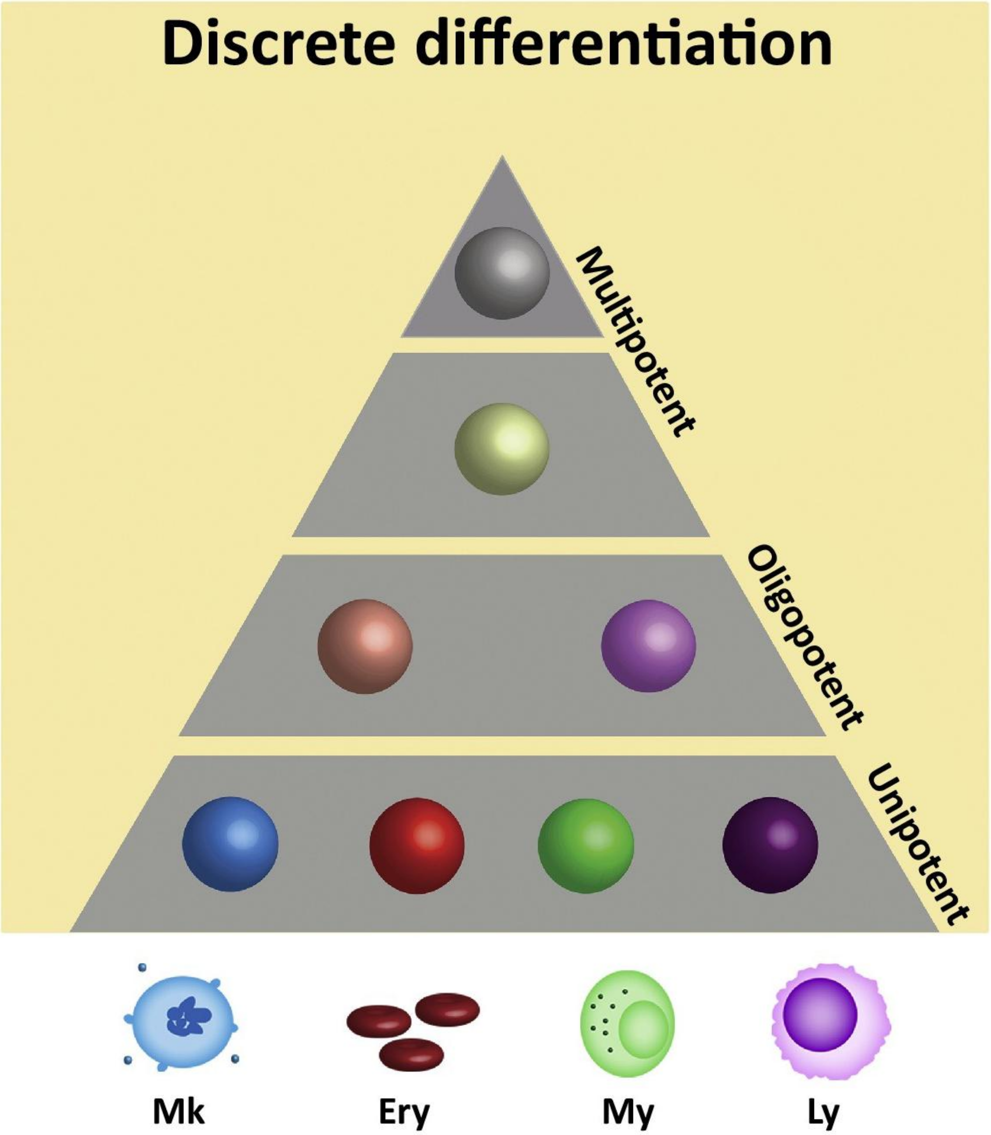
Discrete differentiation

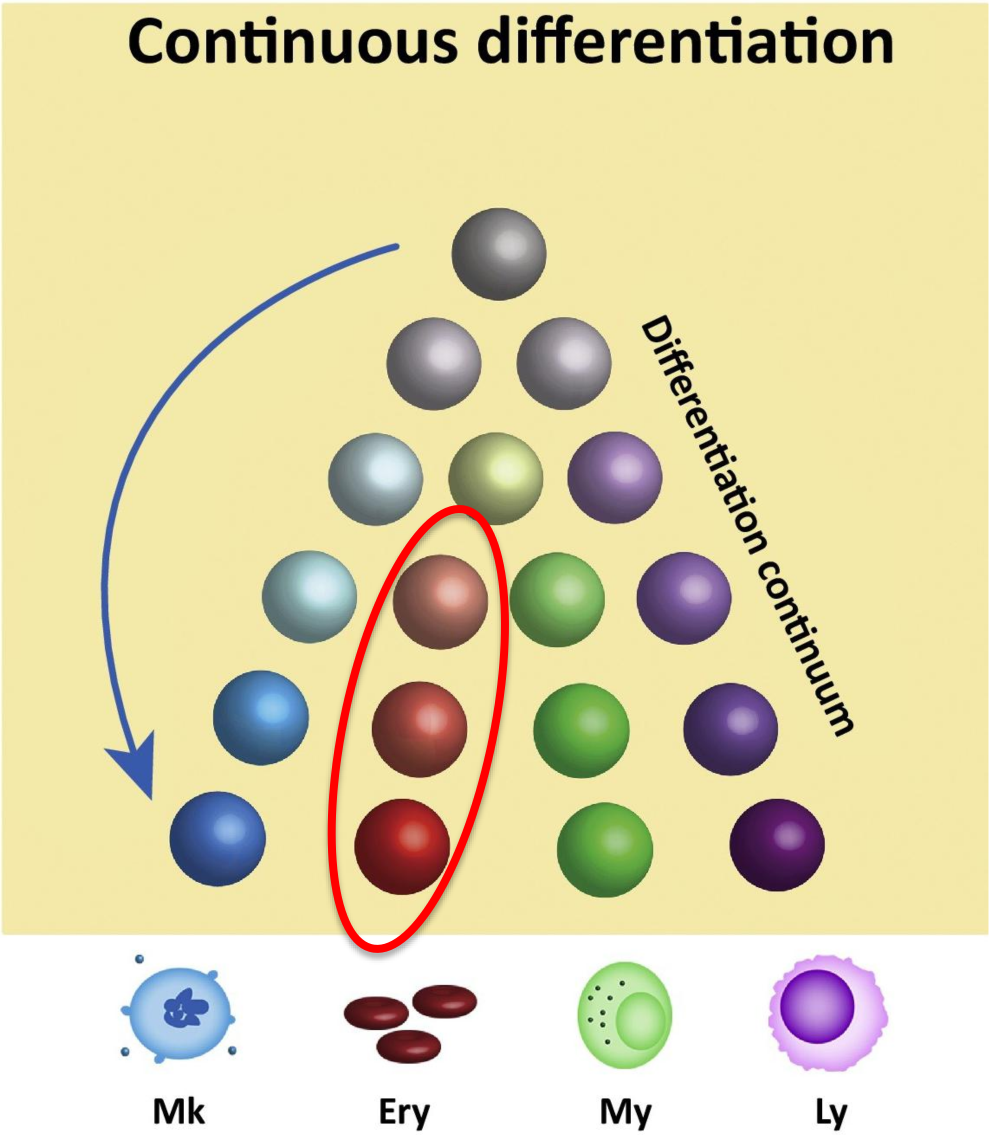


Continuous differentiation



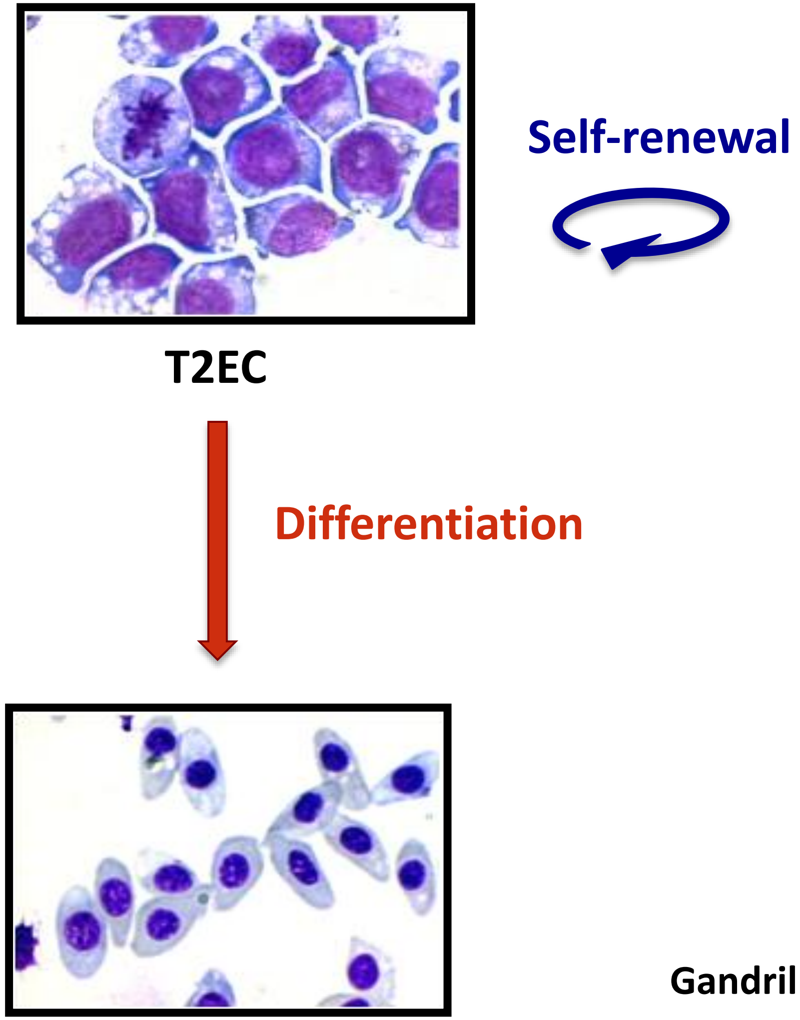
Zhang Y. et al. (2018). Trends Cell Biol 28:976-986.





Zhang Y. et al. (2018). Trends Cell Biol 28:976-986.

Our differentiation model: T2EC



T2EC

Self-renewal

Differentiation

Erythrocytes

Non genetically modified

Mono-ligneage-commited

Homogeneous cell population

Gandrillon *et al.* (1999). *EMBO Journal*

Our question: can we get the underlying molecular network controlling the erythroid differentiation sequence?

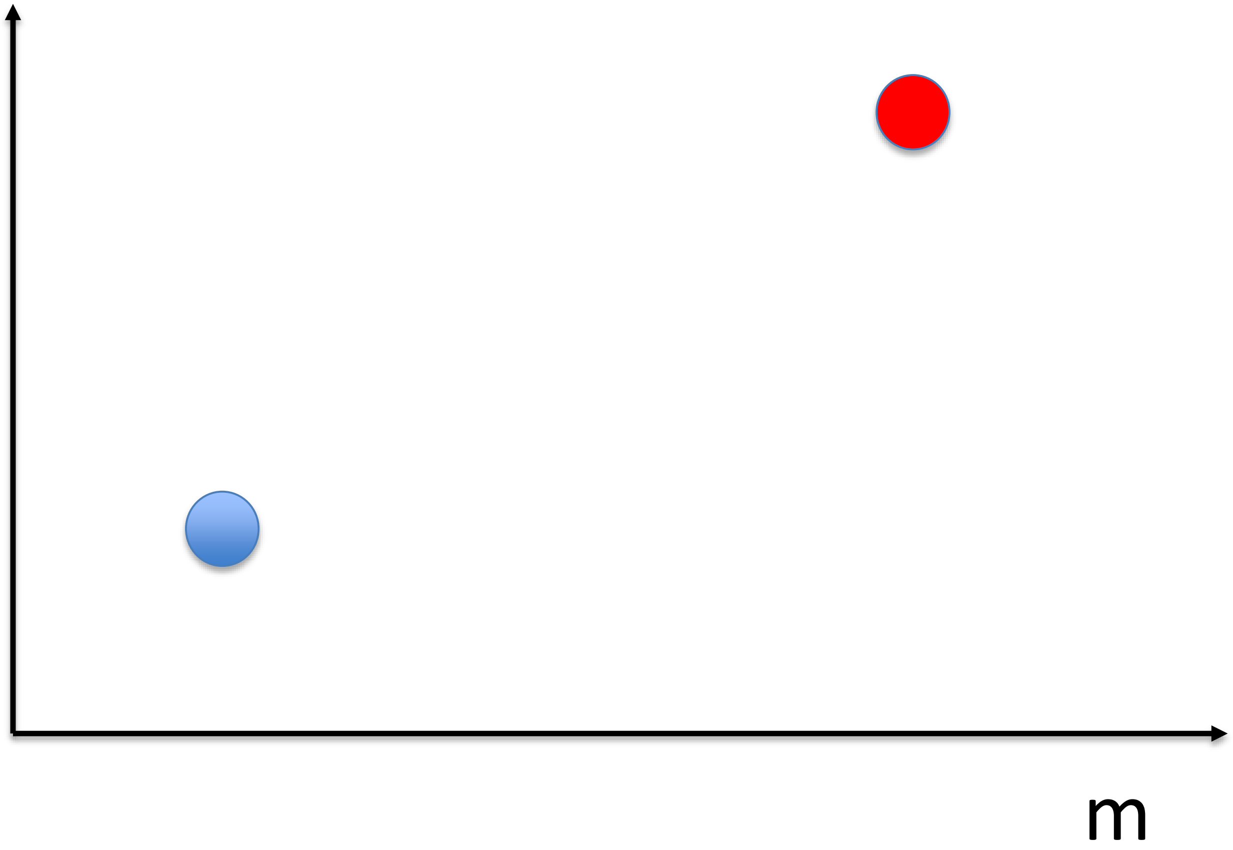
Clearly NOT a new question…

What is new (and we think is decisive) is that it can now be asked at the relevant level: the cell…

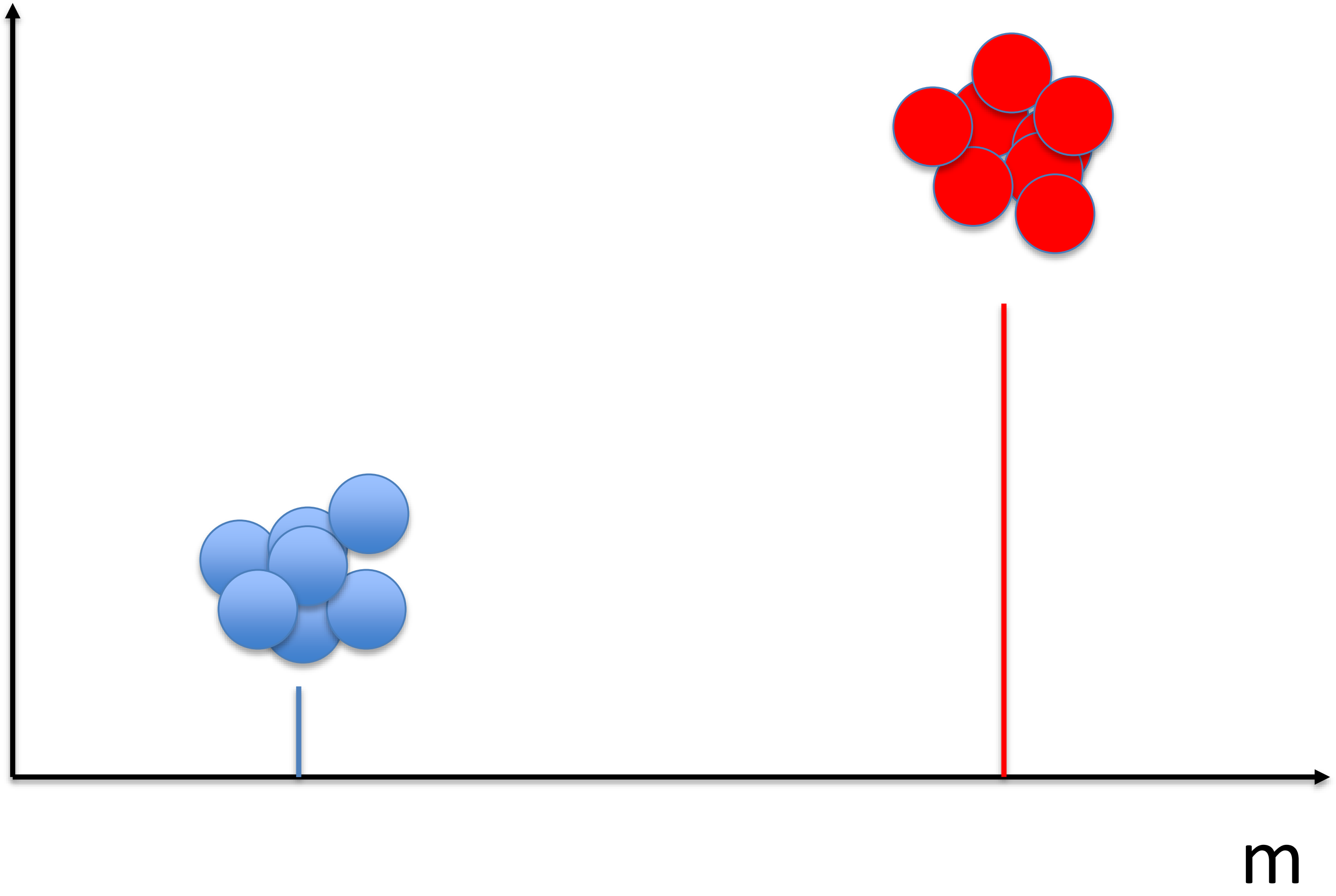
-> go for sc transcriptomics

Why go single cell?

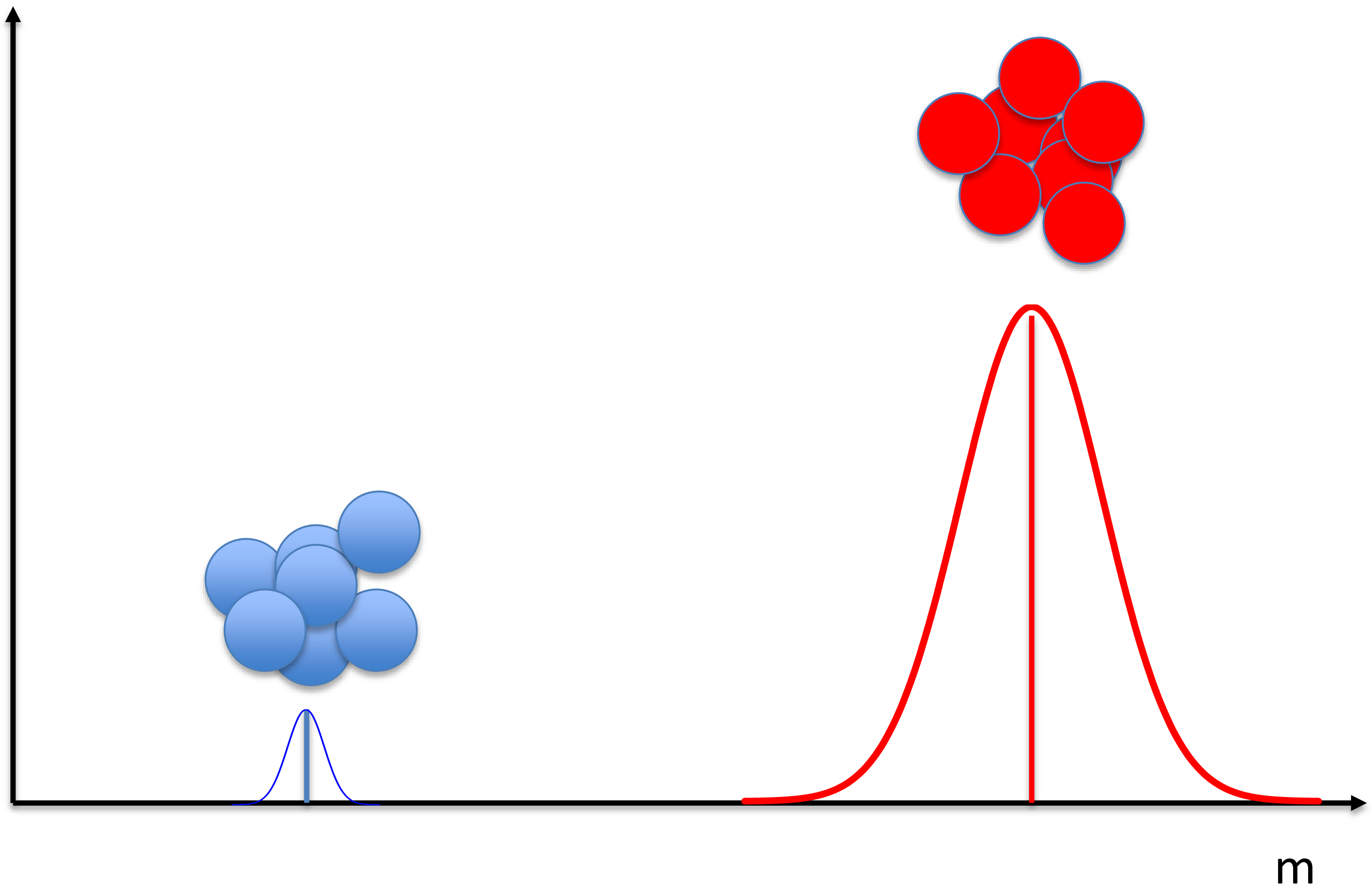
During erythropoiesis, (mean) beta-globin gene expression increases.



Averaged upon 10 million cells



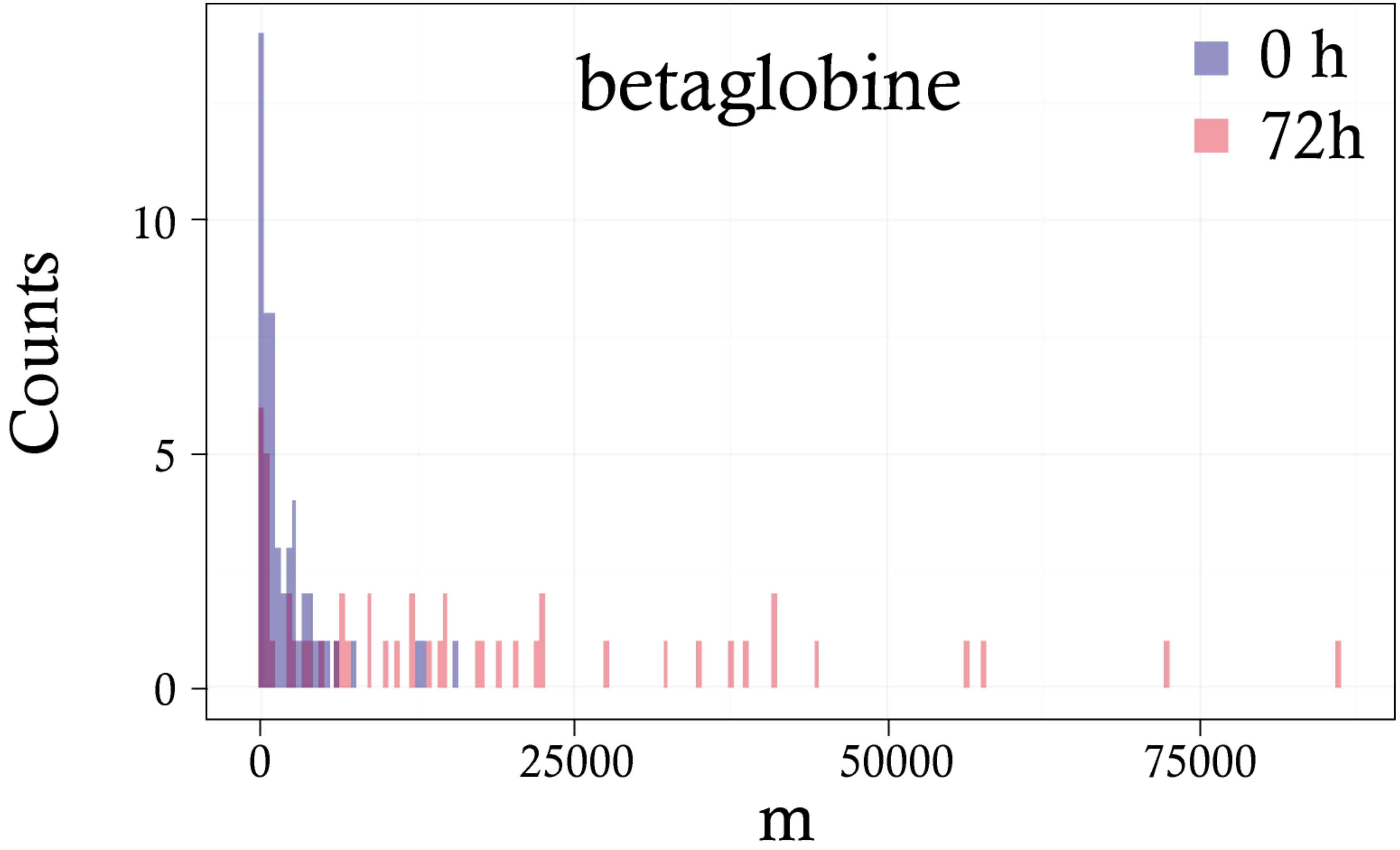
Let’s assume we are now looking at single cells, and assume some cell-to-cell variation



What is to be expected?

Is this true?

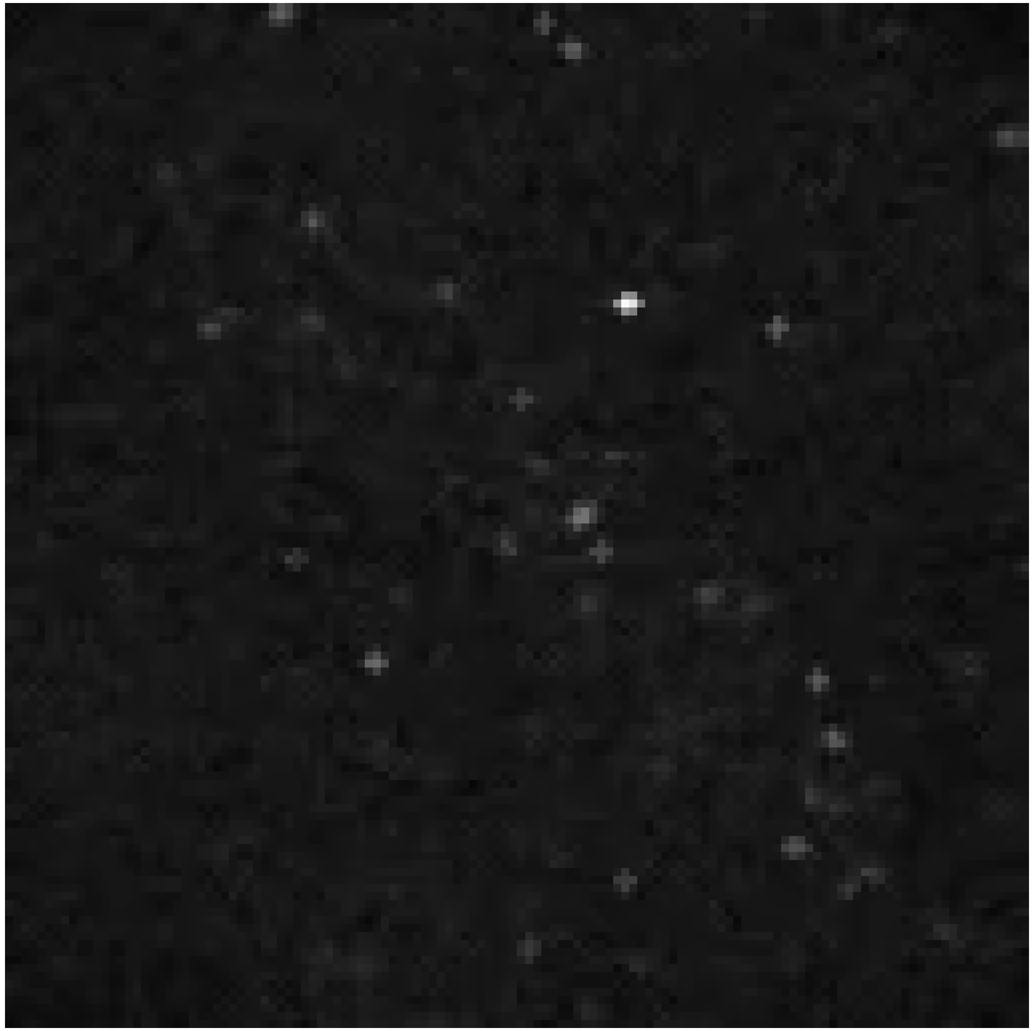
Not really…



Why isn’t it true?

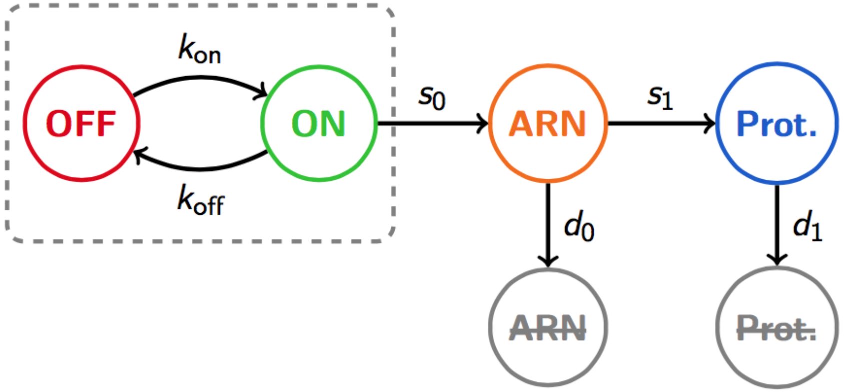
Gene expression is a bursty process

Luciferase



Suter et al.
(2011).
Science 332, pp.
472-474
(2 days movie)

Our proposal (1): describe genes as two-state models (PDMP)



$$G(t):\quad 0 \xrightarrow{k_{on}} 1 \;;\quad 0 \xrightarrow{k_{off}} 1$$

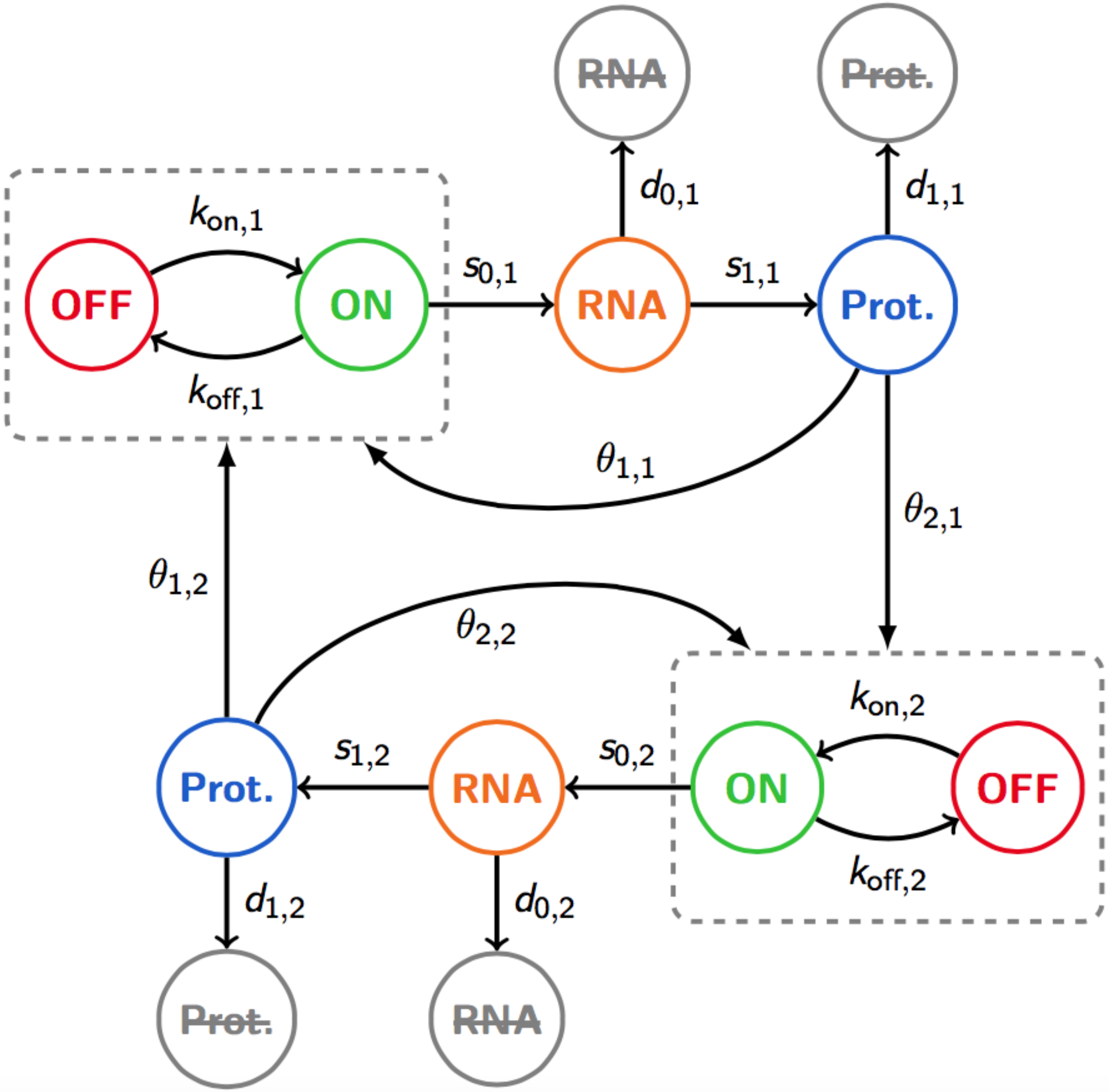
Probabilistic part

$$M'(t) = s_0G(t) - d_0M(t)$$

$$P'(t) = s_1M(t) - d_1P(t)$$

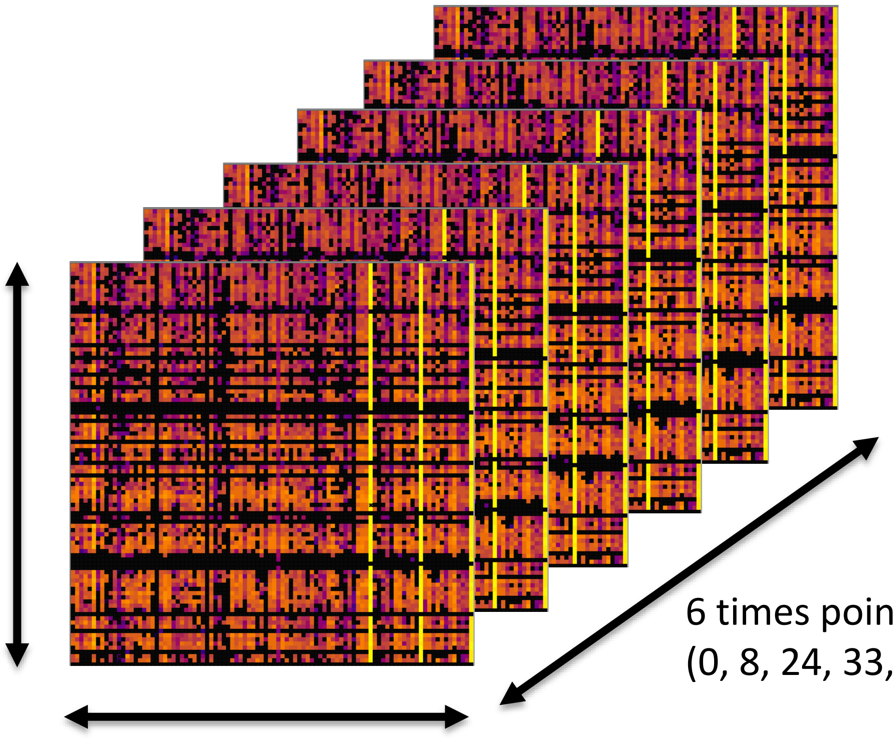
Deterministic part

Our proposal (2): couple the PDMPs to generate the network



→ Executable model

The data set

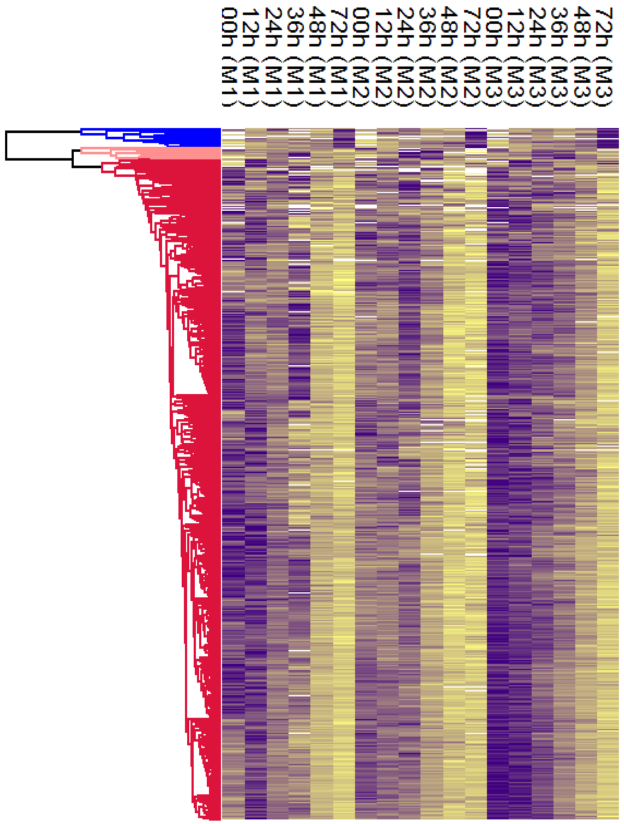


96 cells

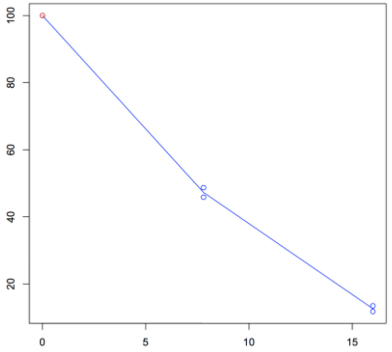
92 genes

6 times points
(0, 8, 24, 33, 48 and 72 hours)

+



+

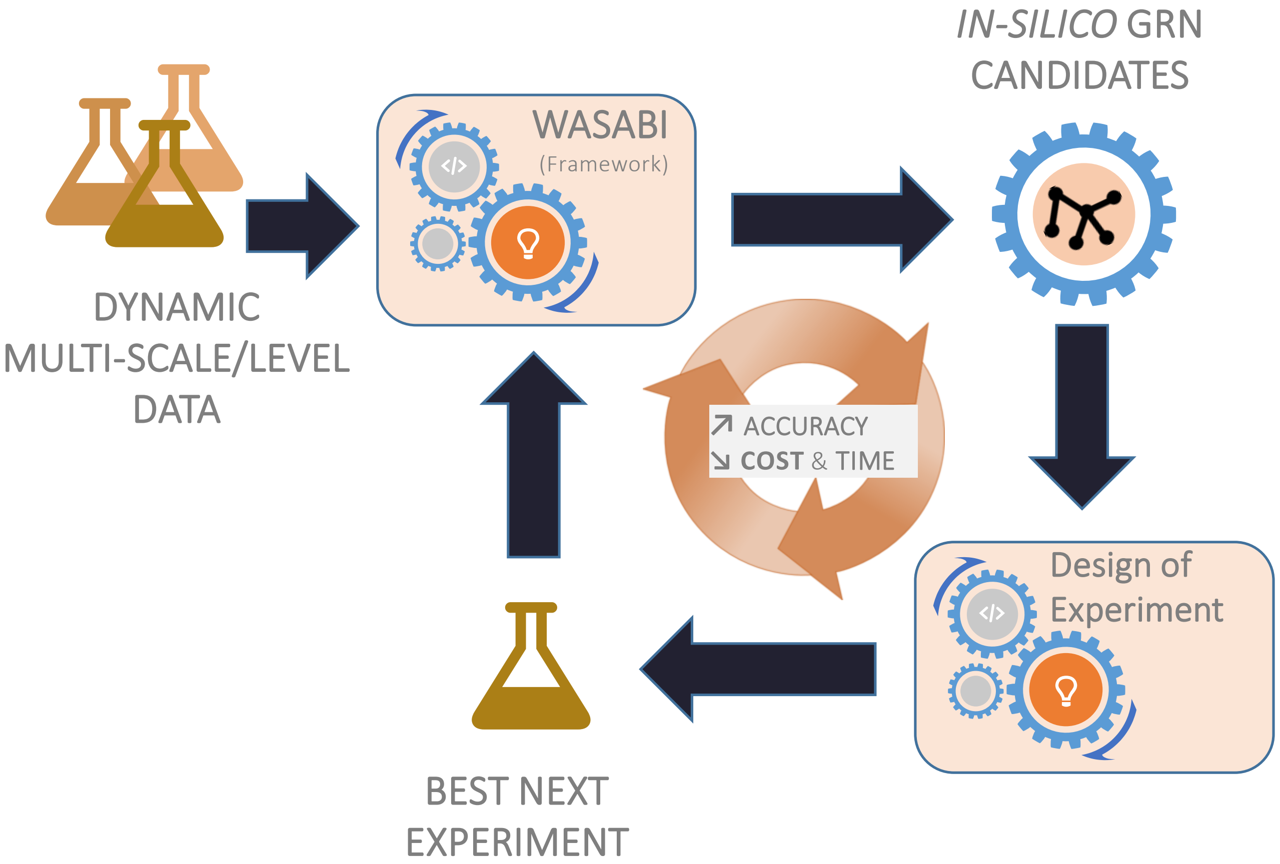


VIDIUM

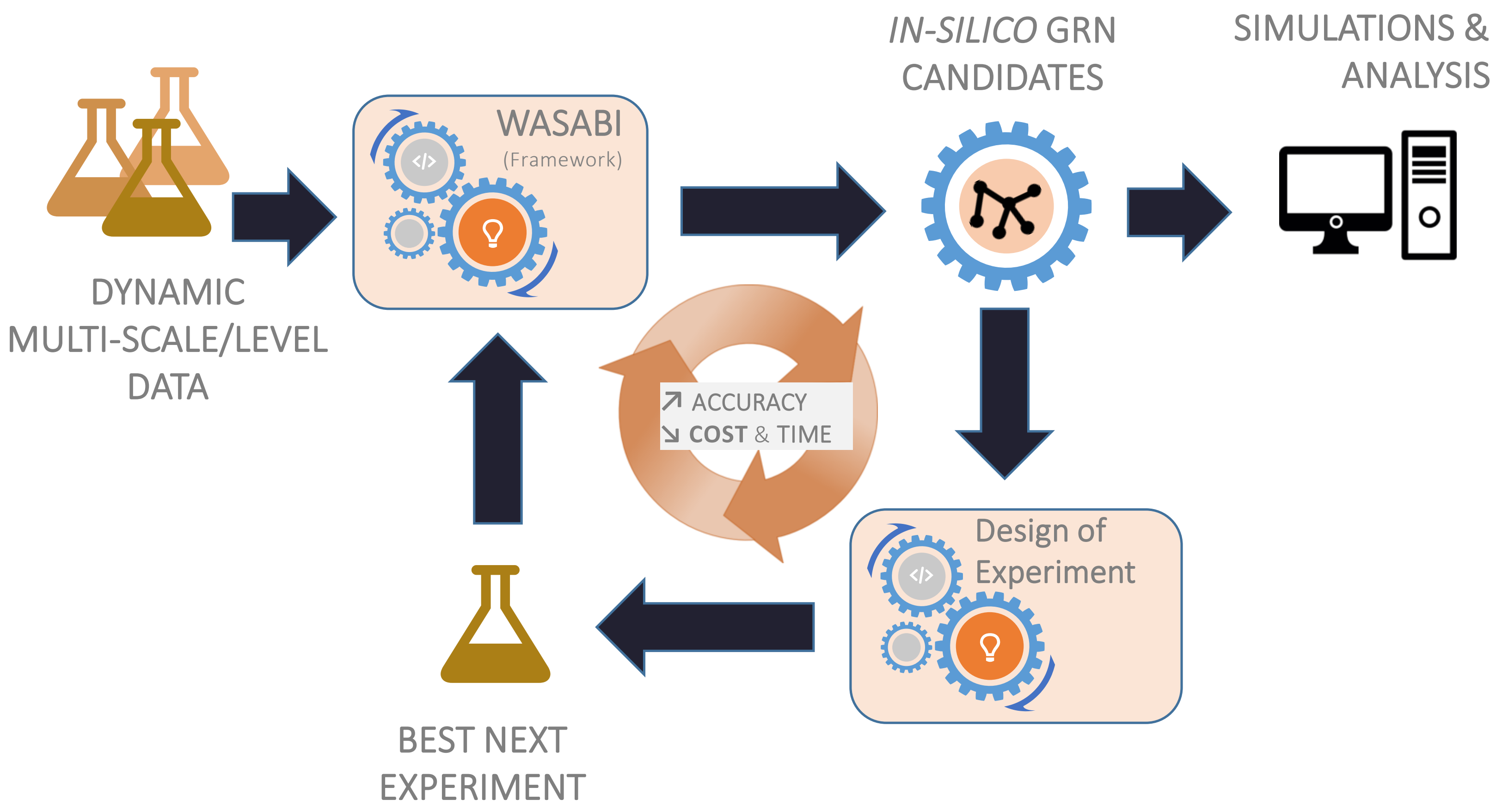
SYSTEMS BIOLOGY SOLUTIONS

Inferring Gene Regulatory Networks from dynamic multi-scale data

GRN INFERENCE FRAMEWORK



GRN INFERENCE FRAMEWORK



DYNAMIC MULTI-SCALE/LEVEL DATA

WASABI (Framework)

IN-SILICO GRN CANDIDATES

SIMULATIONS & ANALYSIS

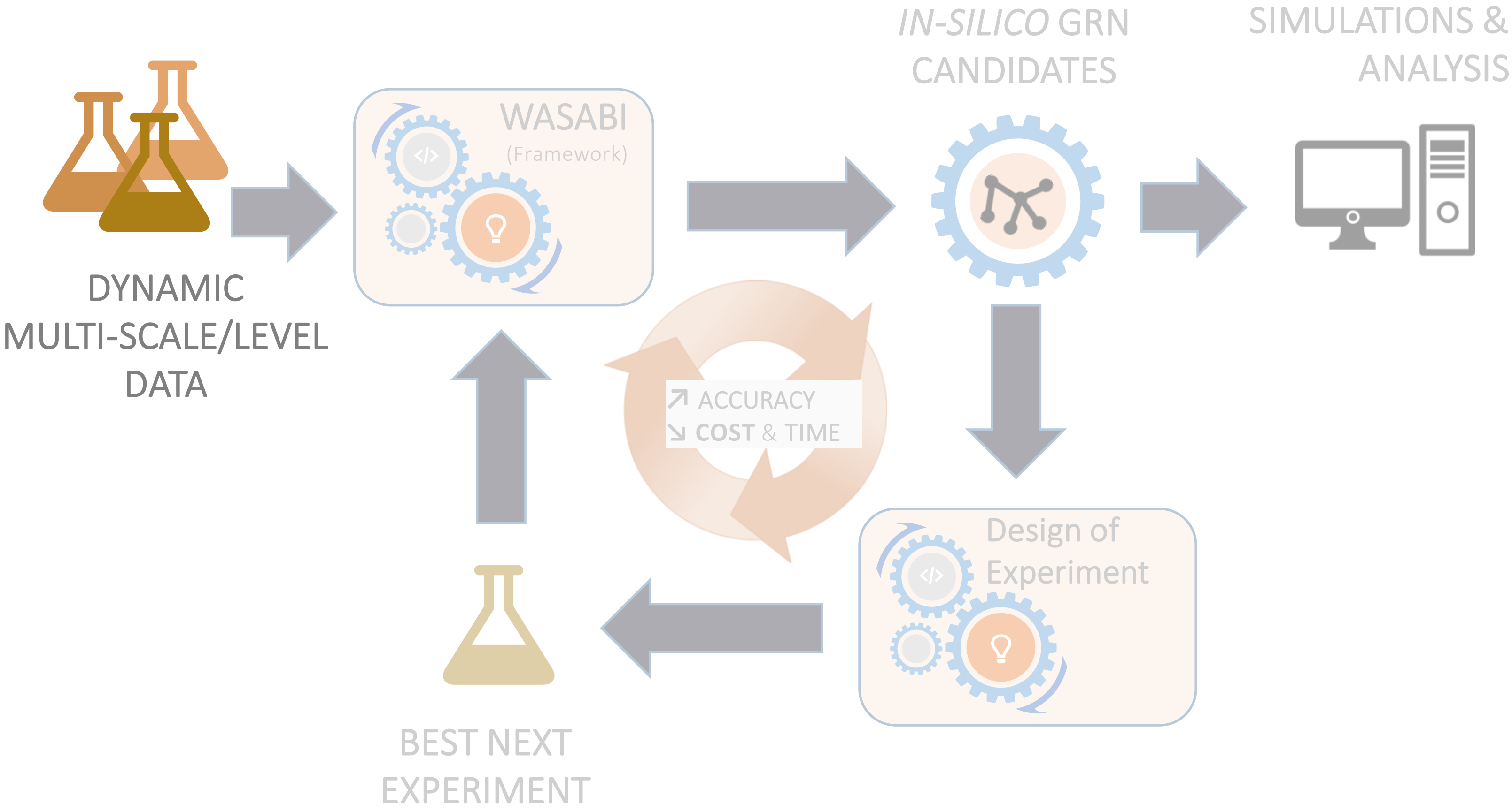
↗ ACCURACY
↘ **COST** & TIME

Design of Experiment

BEST NEXT EXPERIMENT

- **Identification** of new therapeutic targets
- **Prediction** of treatment efficiency
- **Limitation** of side effects
- Personalised Medecine
- Diagnostic
- …

WHY DYNAMIC MULTI-SCALE/LEVEL DATA ?



WHY DYNAMIC MULTI-SCALE/LEVEL DATA ?

DYNAMIC :

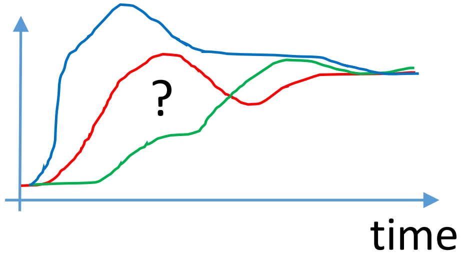
MULTI-LEVEL :

MULTI-SCALE :

WHY DYNAMIC MULTI-SCALE/LEVEL DATA ?

DYNAMIC :

- **Causality** hides in **transient**



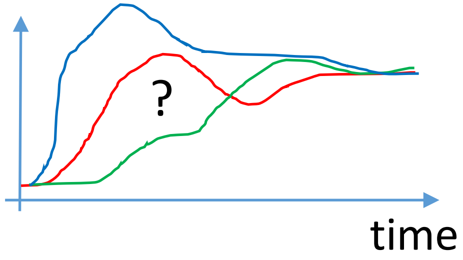
MULTI-SCALE :

MULTI-LEVEL :

WHY DYNAMIC MULTI-SCALE/LEVEL DATA ?

DYNAMIC :

- **Causality** hides in **transient**



MULTI-SCALE :

MULTI-LEVEL :

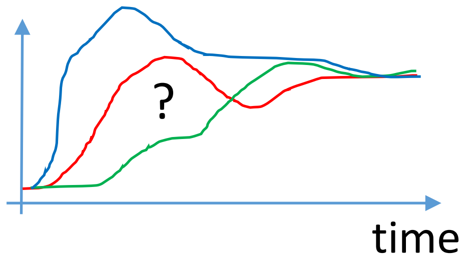
Promoter/RNA/Protein/Cell

- Multi-level **Regulation**
- Multi-omic data **integration**

WHY DYNAMIC MULTI-SCALE/LEVEL DATA ?

DYNAMIC :

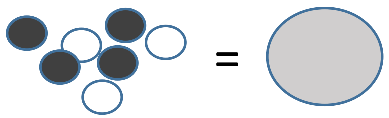
- **Causality** hides in **transient**

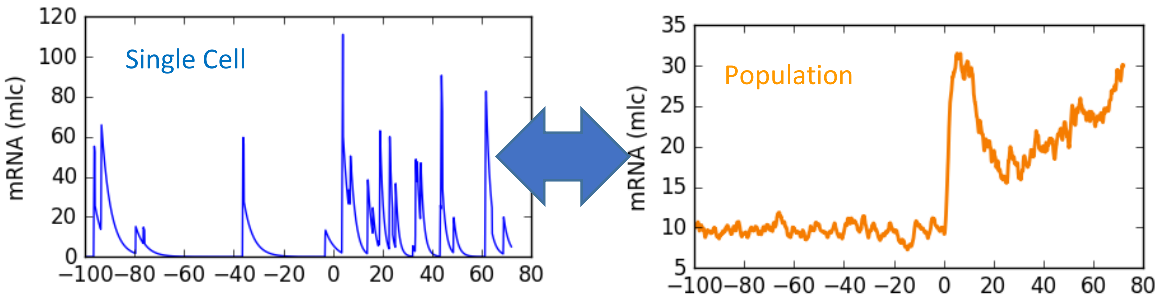


MULTI-SCALE :

Single-cell VS population

- Cellular **heterogeneity**
- Inner cell **stochasticity** > RNA burst



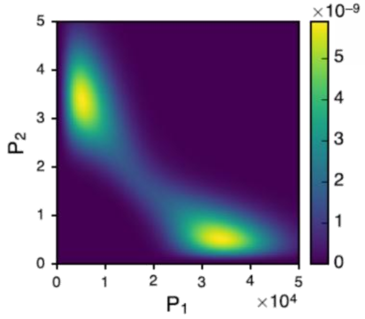


MULTI-LEVEL :

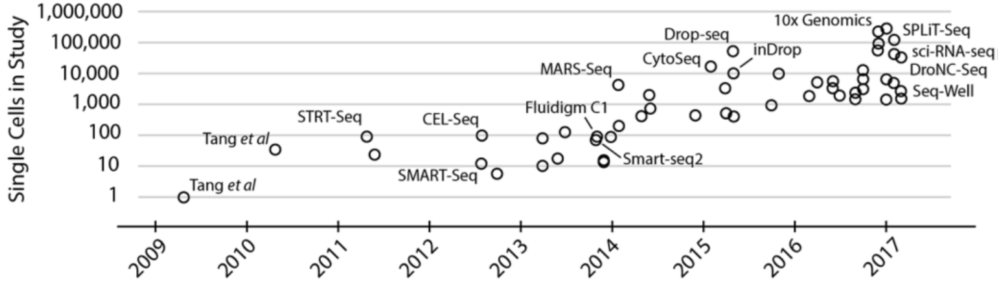
Promoter/RNA/Protein/Cell

- Multi-level **Regulation**
- Multi-omic data **integration**

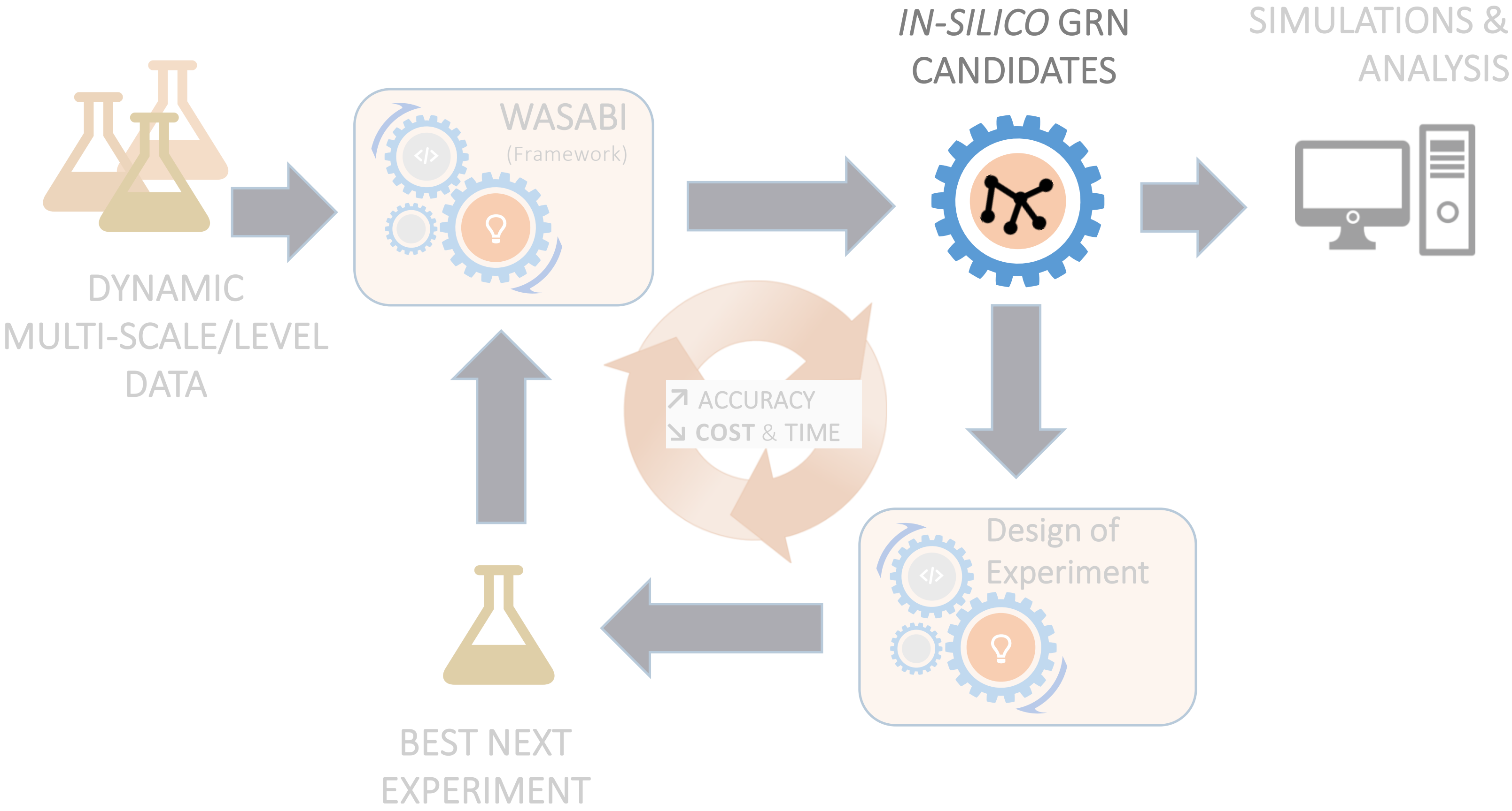
- **Joint** distribution



- Statistical power > single-Cell **Moore's law**



GENE REGULATORY NETWORK MODEL



IN-SILICO GRN CANDIDATES

SIMULATIONS & ANALYSIS

DYNAMIC MULTI-SCALE/LEVEL DATA

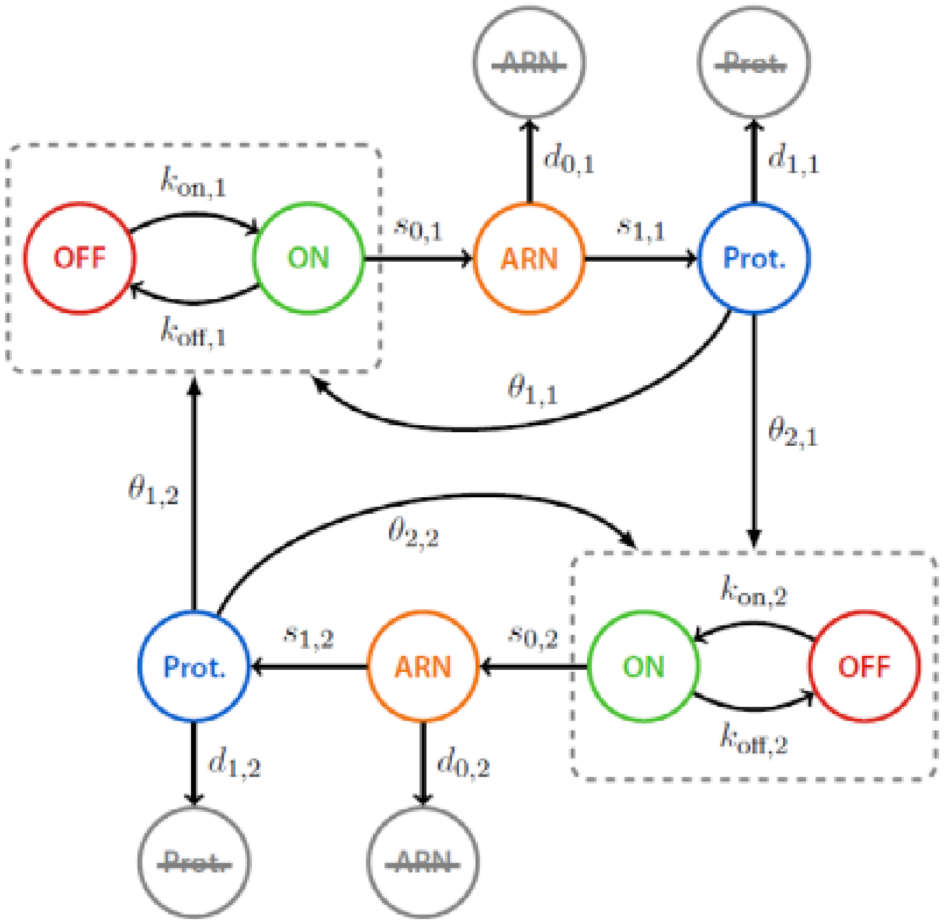
WASABI (Framework)

↗ ACCURACY
↘ **COST** & TIME

Design of Experiment

BEST NEXT EXPERIMENT

MODELLING AT SINGLE CELL SCALE

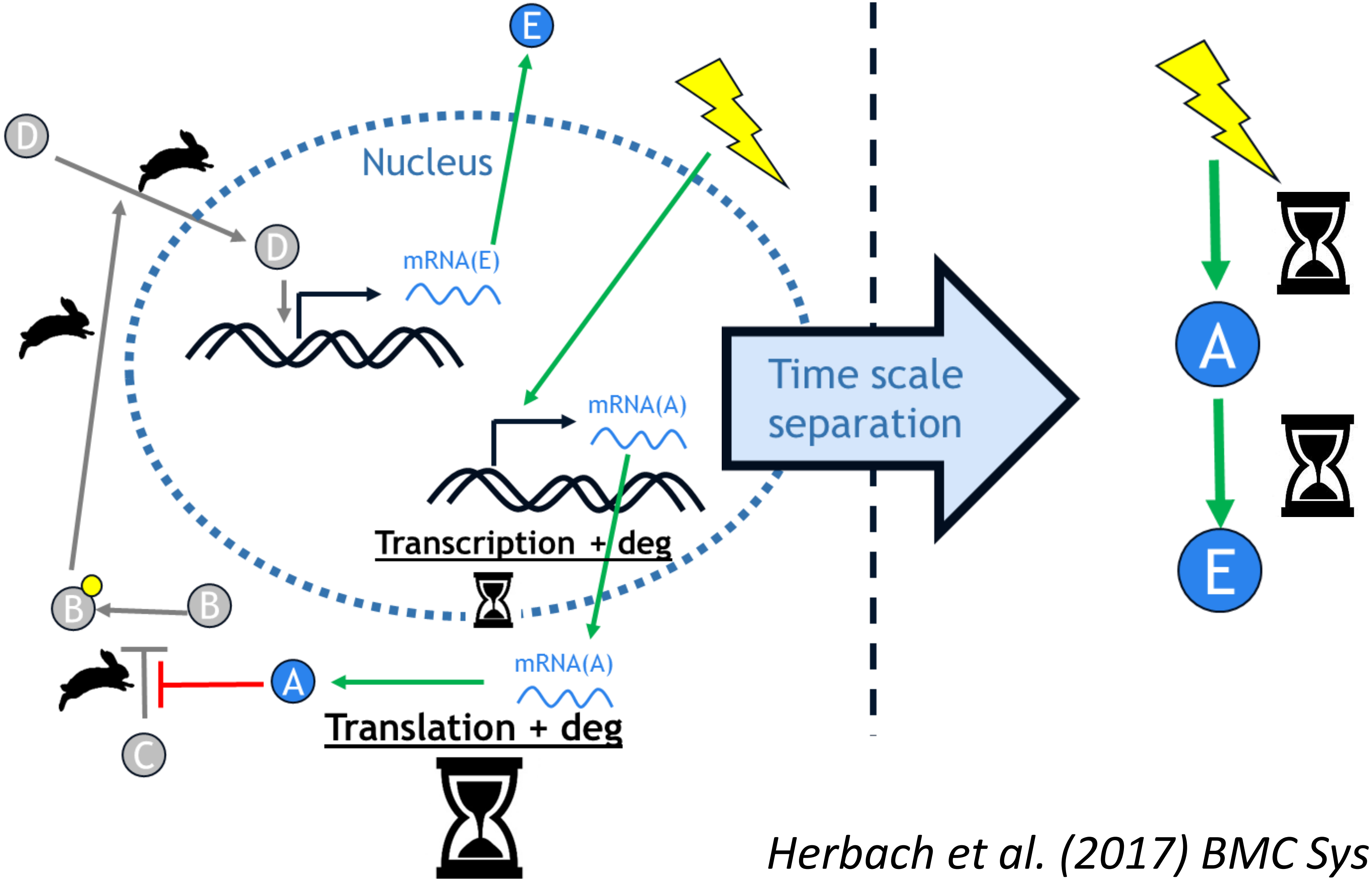


$$Kon(t) = Kon_{basal} + \overbrace{\sum F(\theta_i, P_i(t)) + \sum F(\theta_i, Stim(t))}^{\text{Network model}}$$

$$Koff(t) = Koff_{basal} + \sum F(\theta_i, P_i(t)) + \sum F(\theta_i, Stim(t))$$

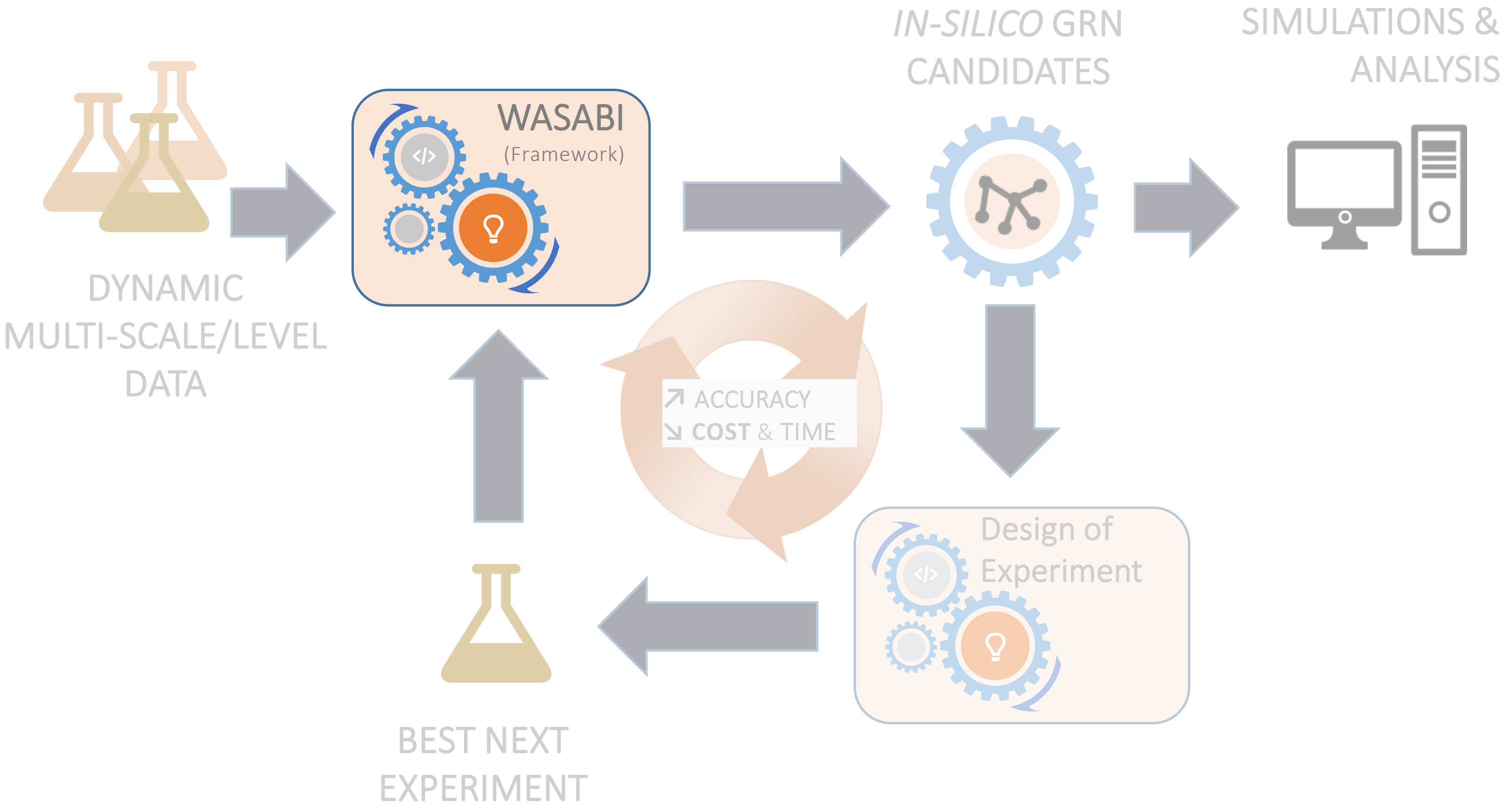
Herbach et al. (2017) BMC Systems Biology

GRN INTERACTION IS FUNCTIONAL



Herbach et al. (2017) BMC Systems Biology

WASABI



DYNAMIC MULTI-SCALE/LEVEL DATA

WASABI (Framework)

IN-SILICO GRN CANDIDATES

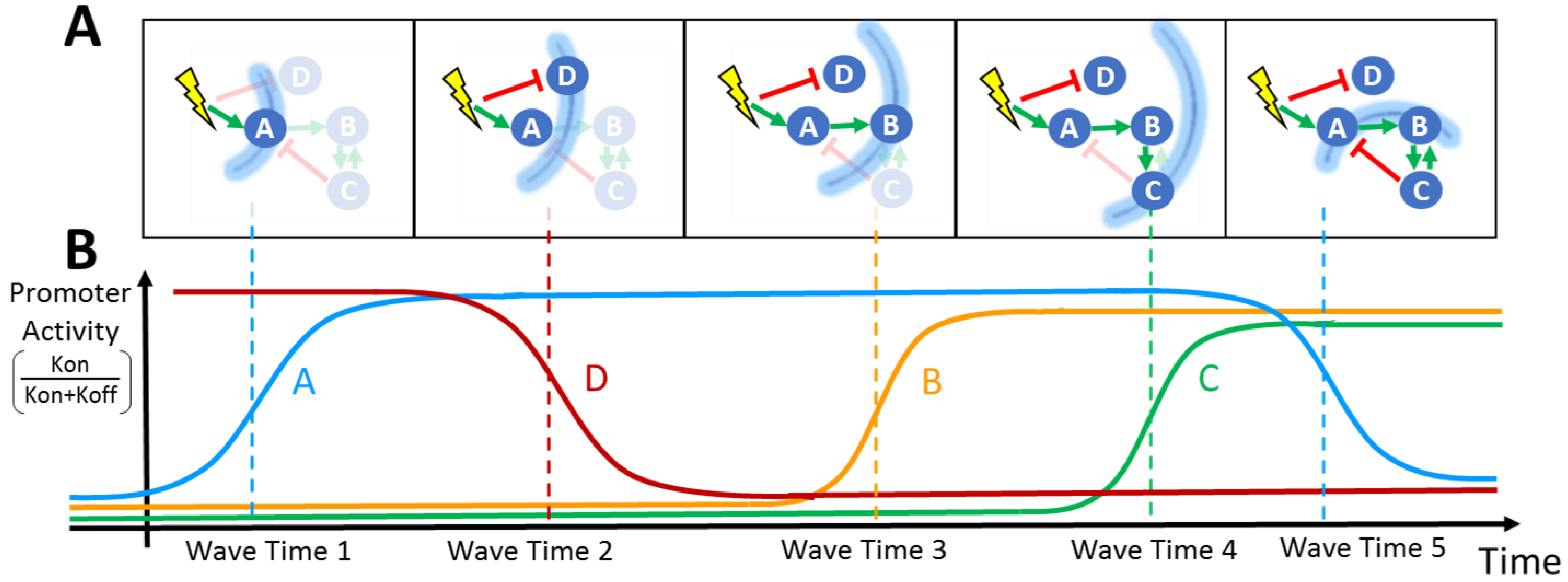
SIMULATIONS & ANALYSIS

↗ ACCURACY
↘ COST & TIME

Design of Experiment

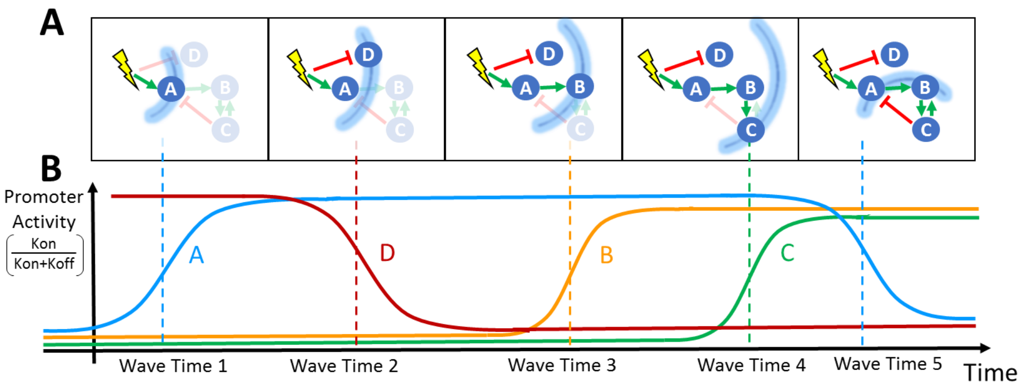
BEST NEXT EXPERIMENT

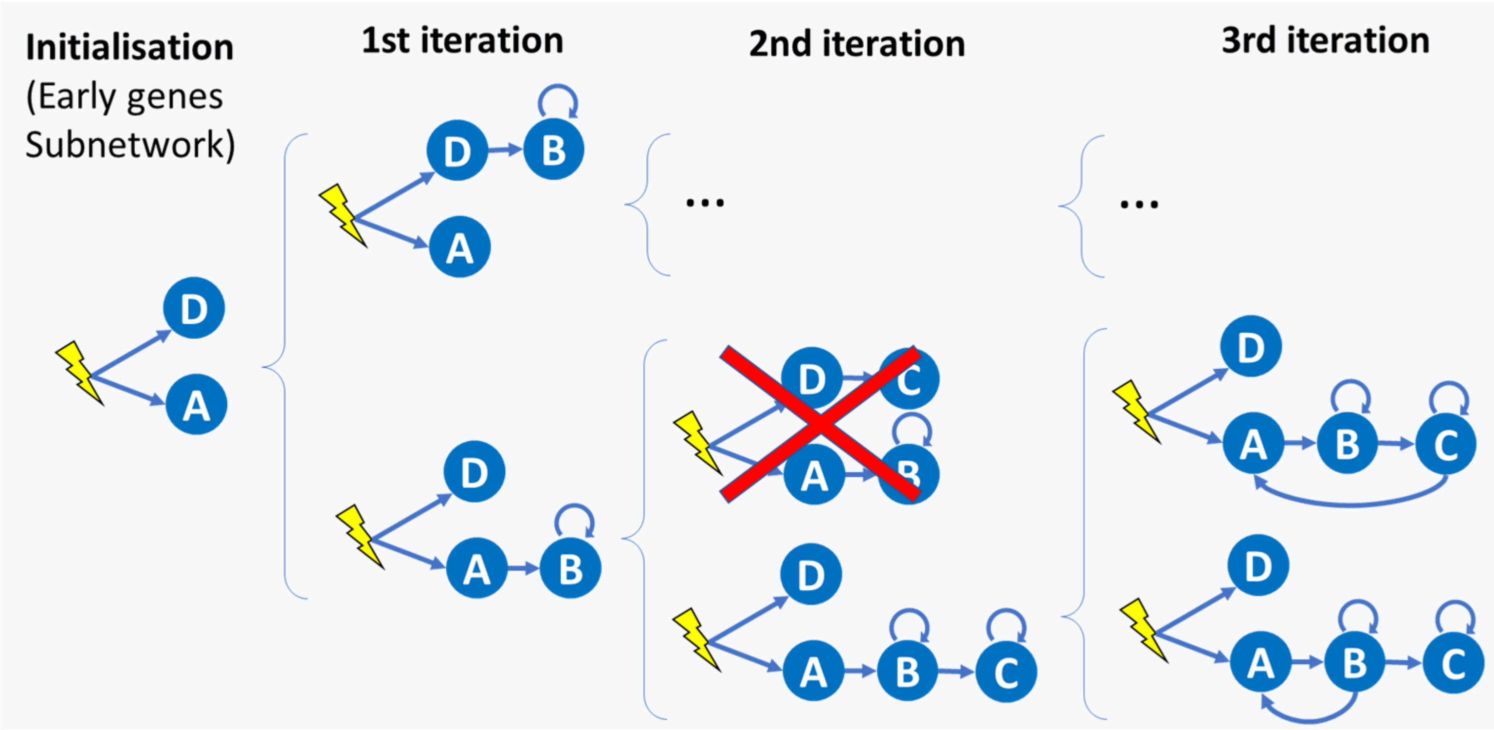
WASABI = WAVES ANALYSIS BASED INFERENCE



Bonnaffoux et al. (2019), BMC Bioinformatics

WASABI SPLITS & PARALELLIZE GRN INFERENCE PROBLEM

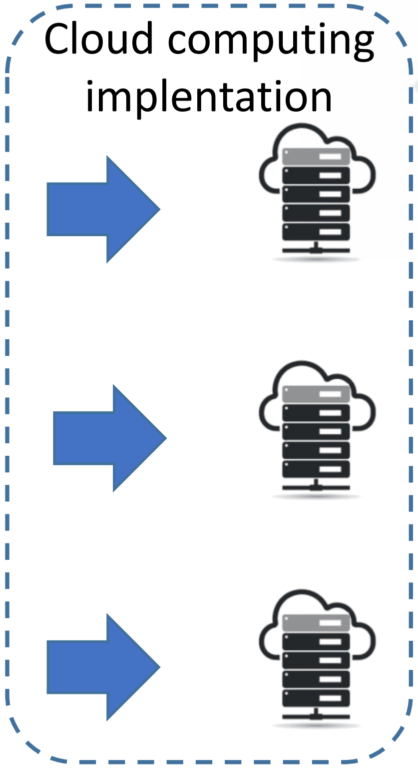



Initialisation
(Early genes Subnetwork)

1st iteration

2nd iteration

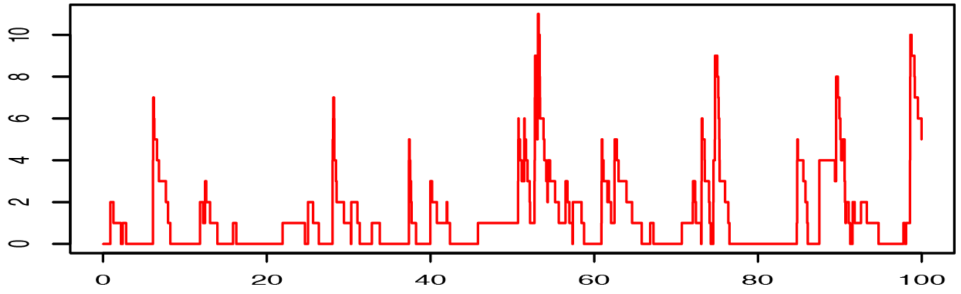
3rd iteration


Cloud computing implentation

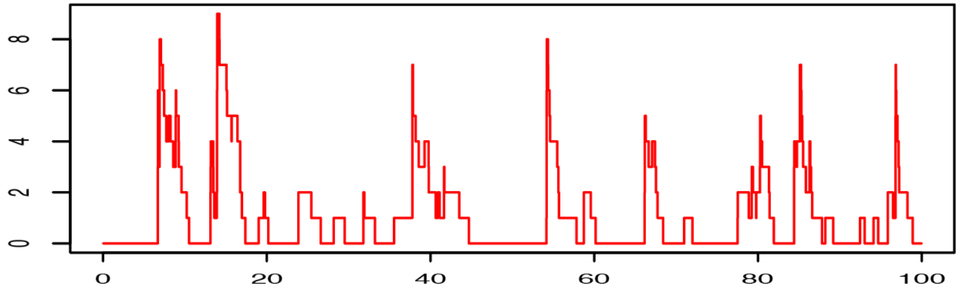
Bonnaffoux et al. (2019), BMC Bioinformatics

Inference fitting : Distribution distance

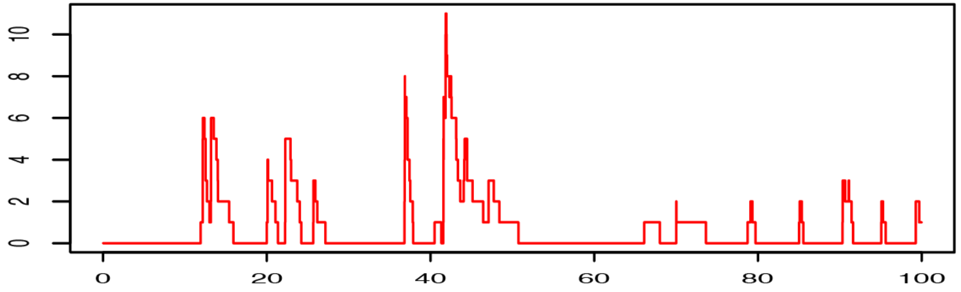
Cell n° 1



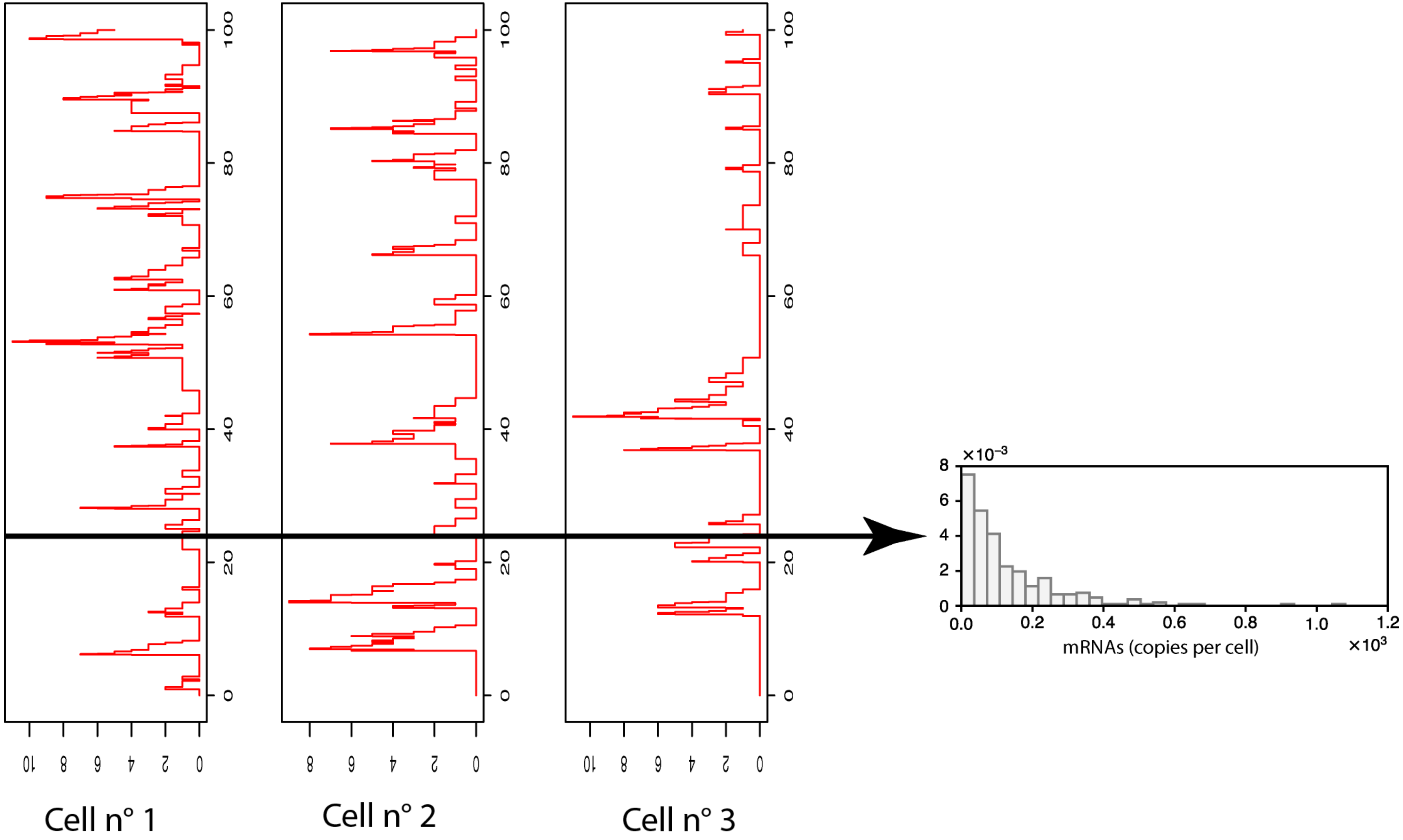
Cell n° 2



Cell n° 3



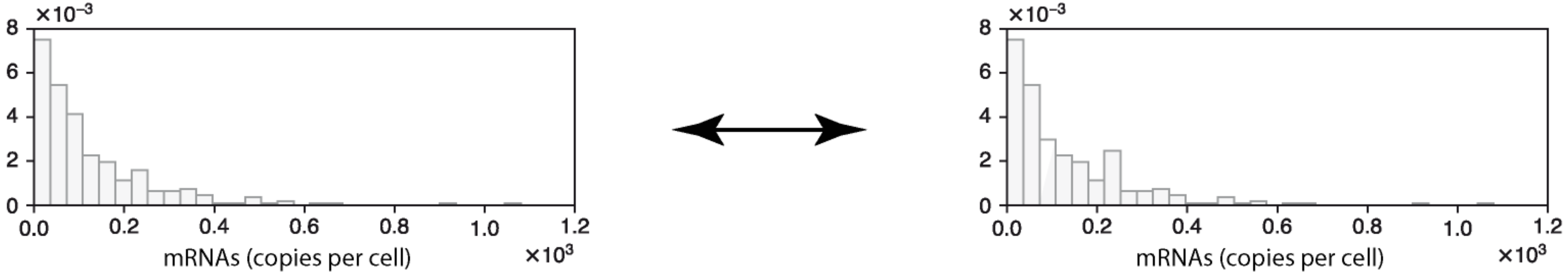
Inference fitting : Distribution distance

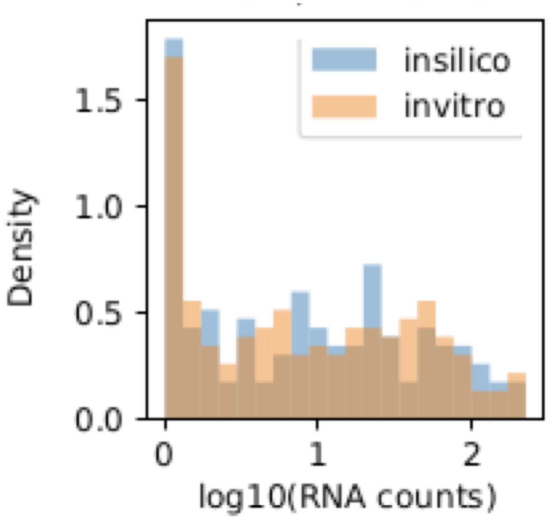


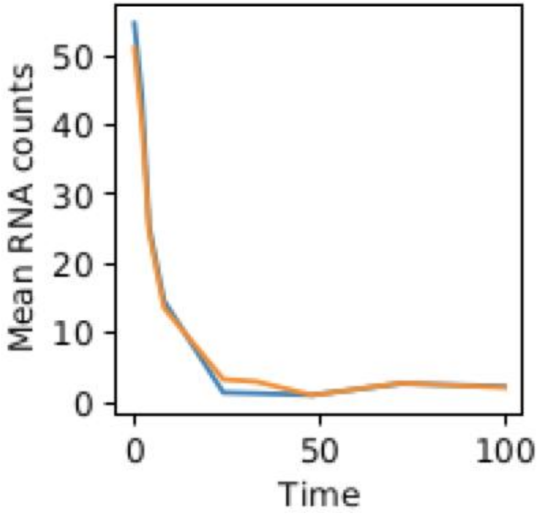
Inference fitting : Distribution distance

Model-generated distribution

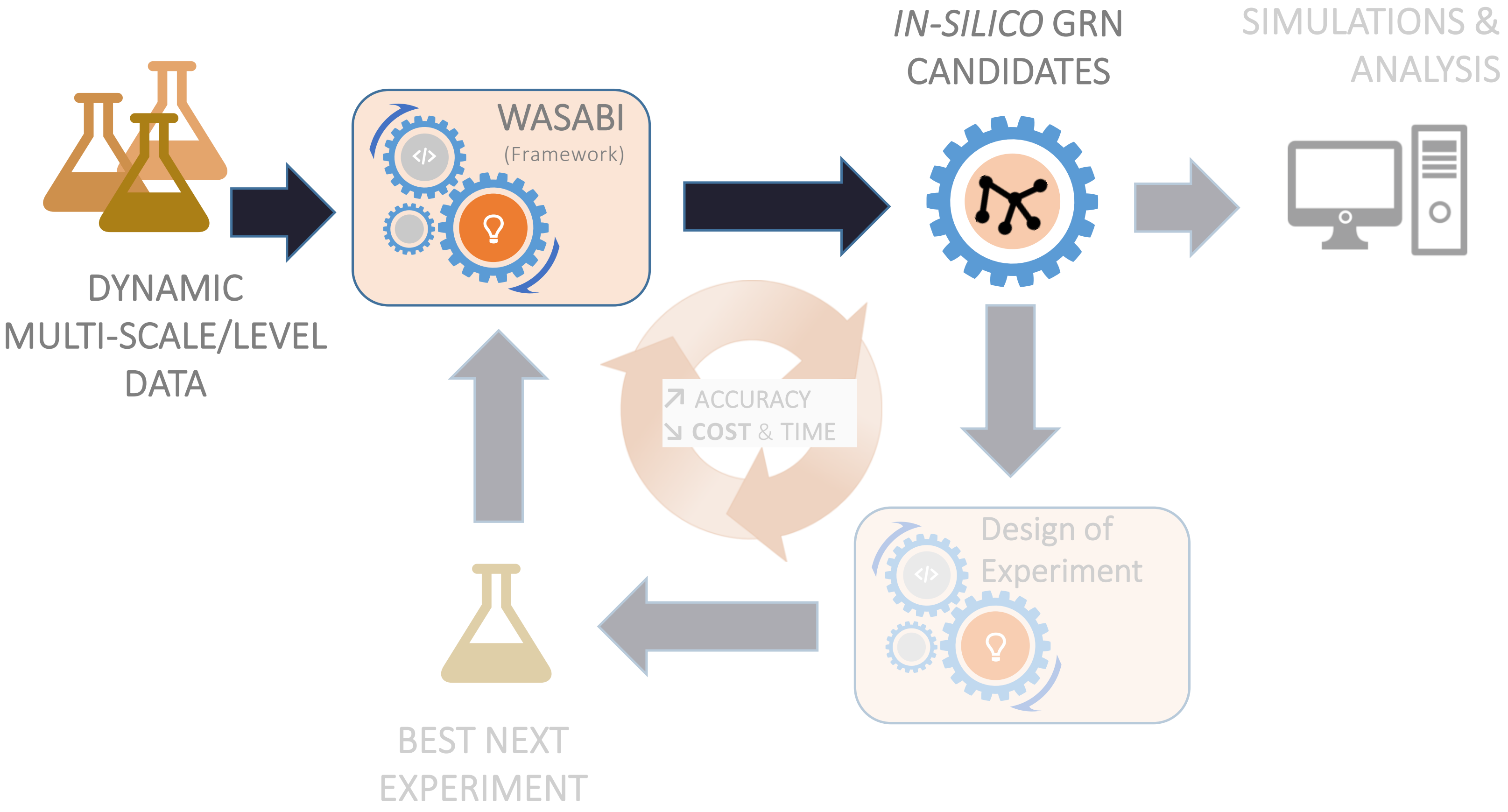
Experimentally-observed distribution



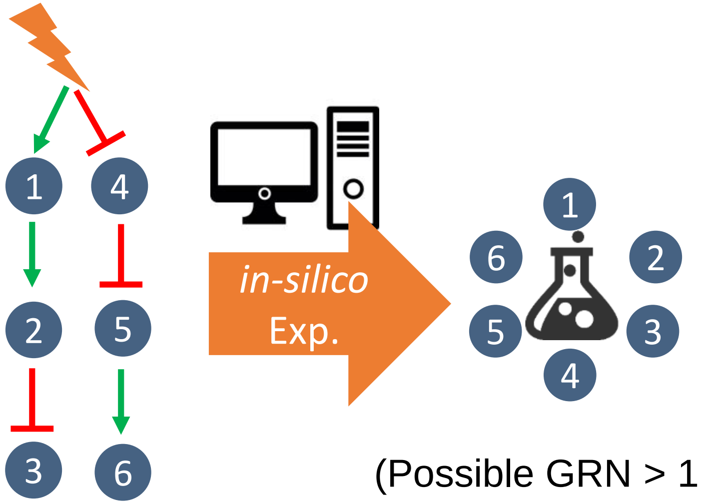




WASABI : IN-SILICO VALIDATION



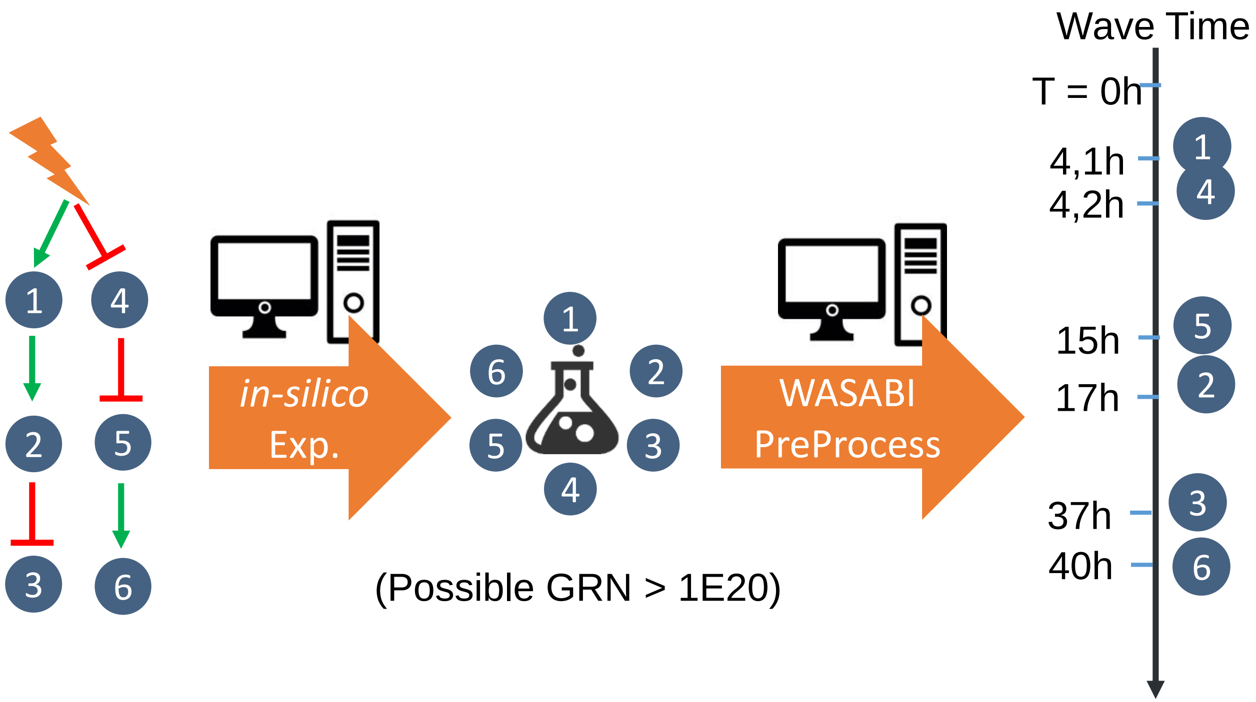
IN-SILICO VALIDATION



(Possible GRN > 1E20)

Bonnaffoux et al. (2019), BMC Bioinformatics

IN-SILICO VALIDATION



in-silico Exp.

WASABI PreProcess

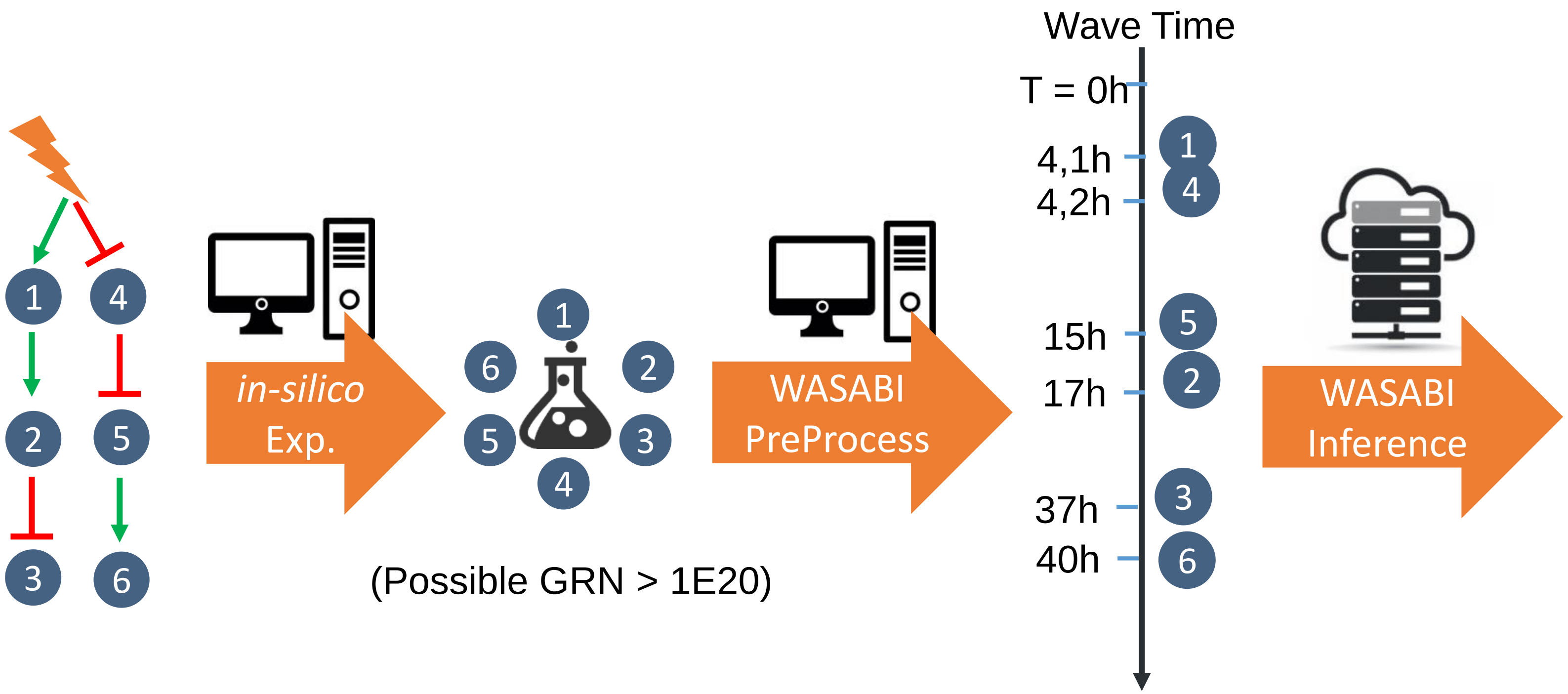
(Possible GRN > 1E20)

Wave Time

T = 0h
4,1h 1
4,2h 4
15h 5
17h 2
37h 3
40h 6

Bonnaffoux et al. (2019), BMC Bioinformatics

IN-SILICO VALIDATION



in-silico Exp.

(Possible GRN > 1E20)

WASABI PreProcess

Wave Time

T = 0h

4,1h

4,2h

15h

17h

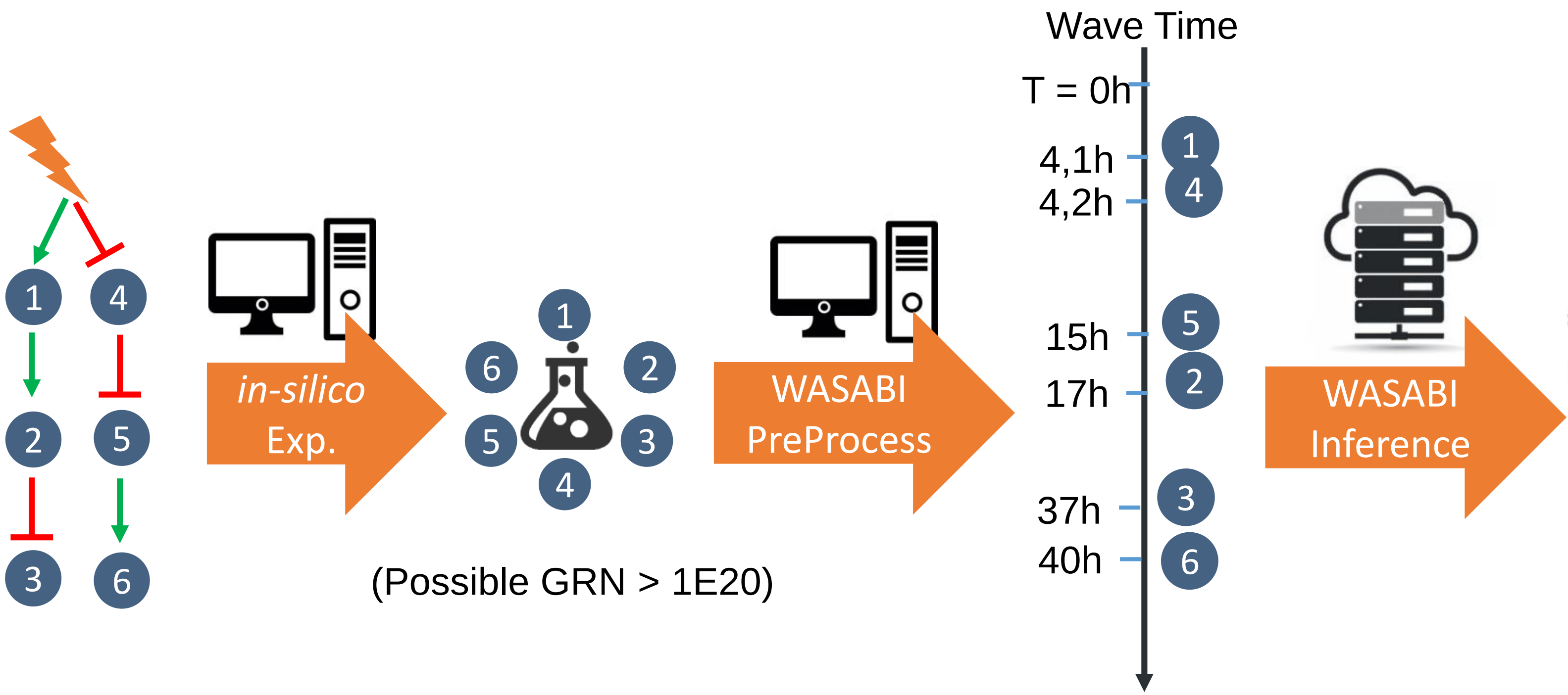
37h

40h

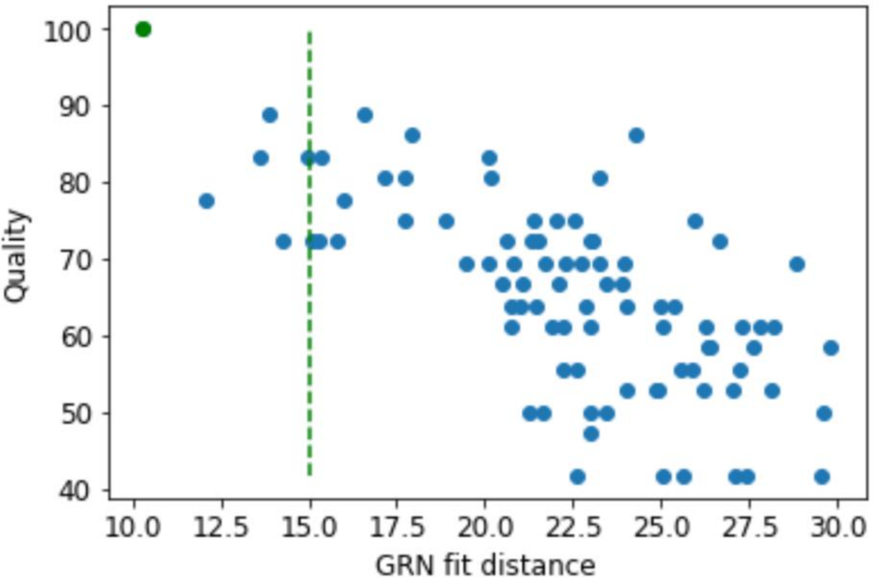
WASABI Inference

Bonnaffoux et al. (2019), BMC Bioinformatics

IN-SILICO VALIDATION

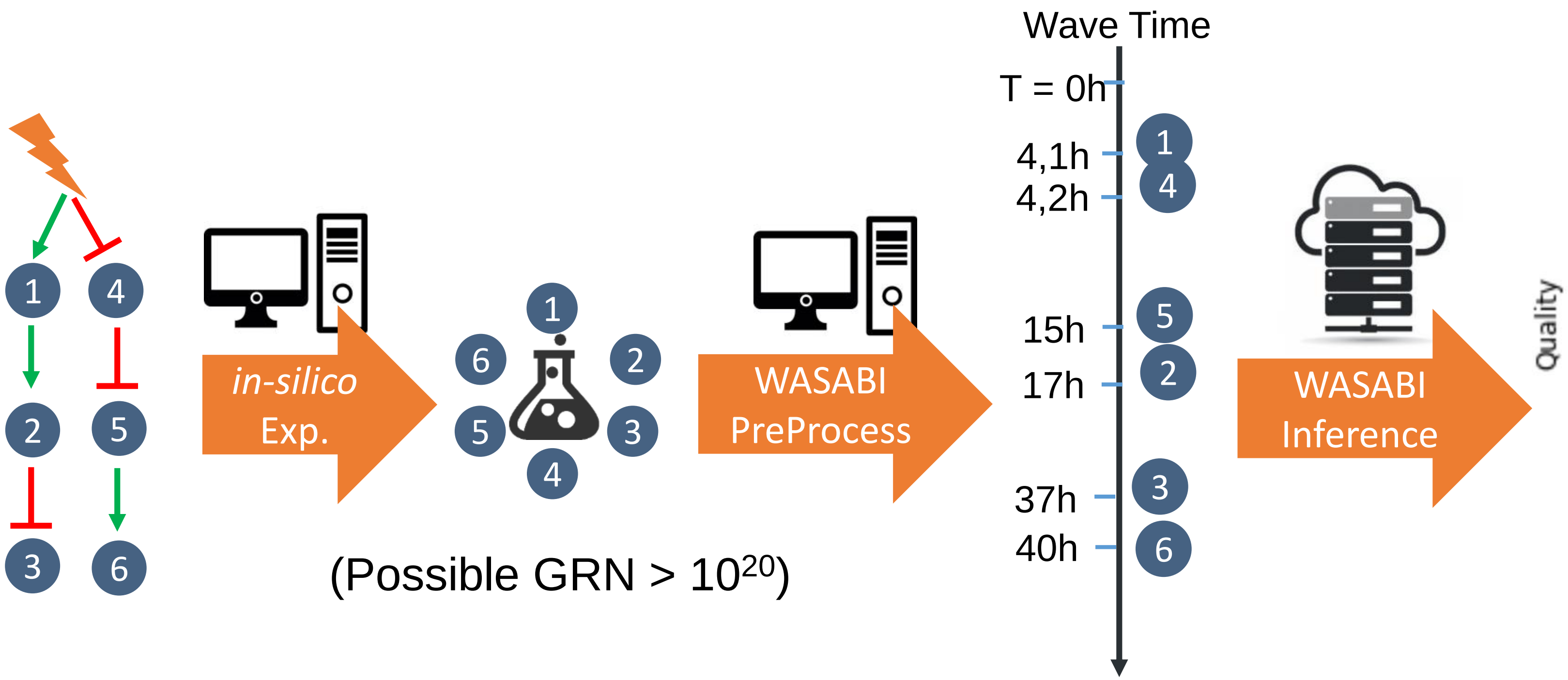


(Possible GRN > 1E20)

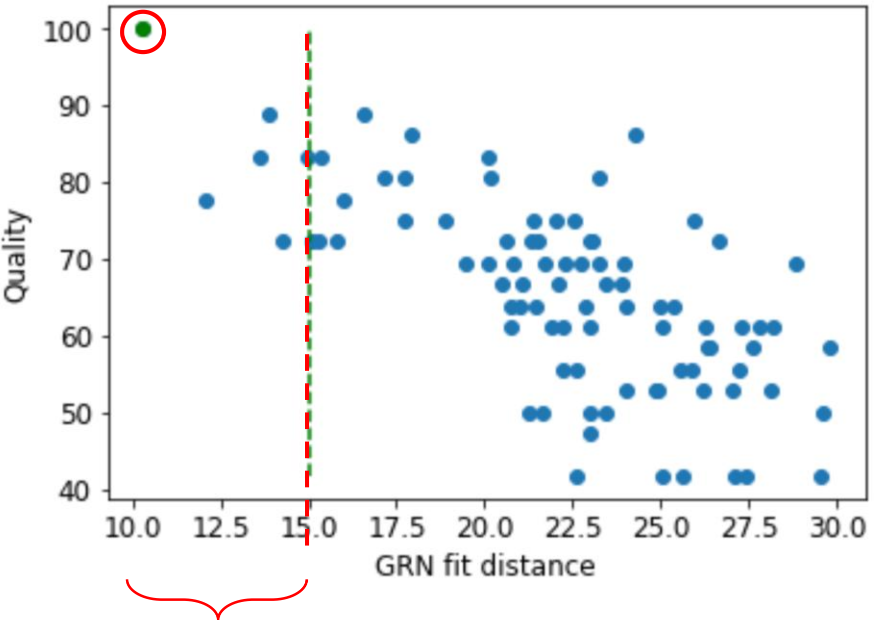


Bonnaffoux et al. (2019), BMC Bioinformatics

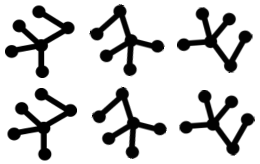
IN-SILICO VALIDATION



(Possible GRN > 10^{20})

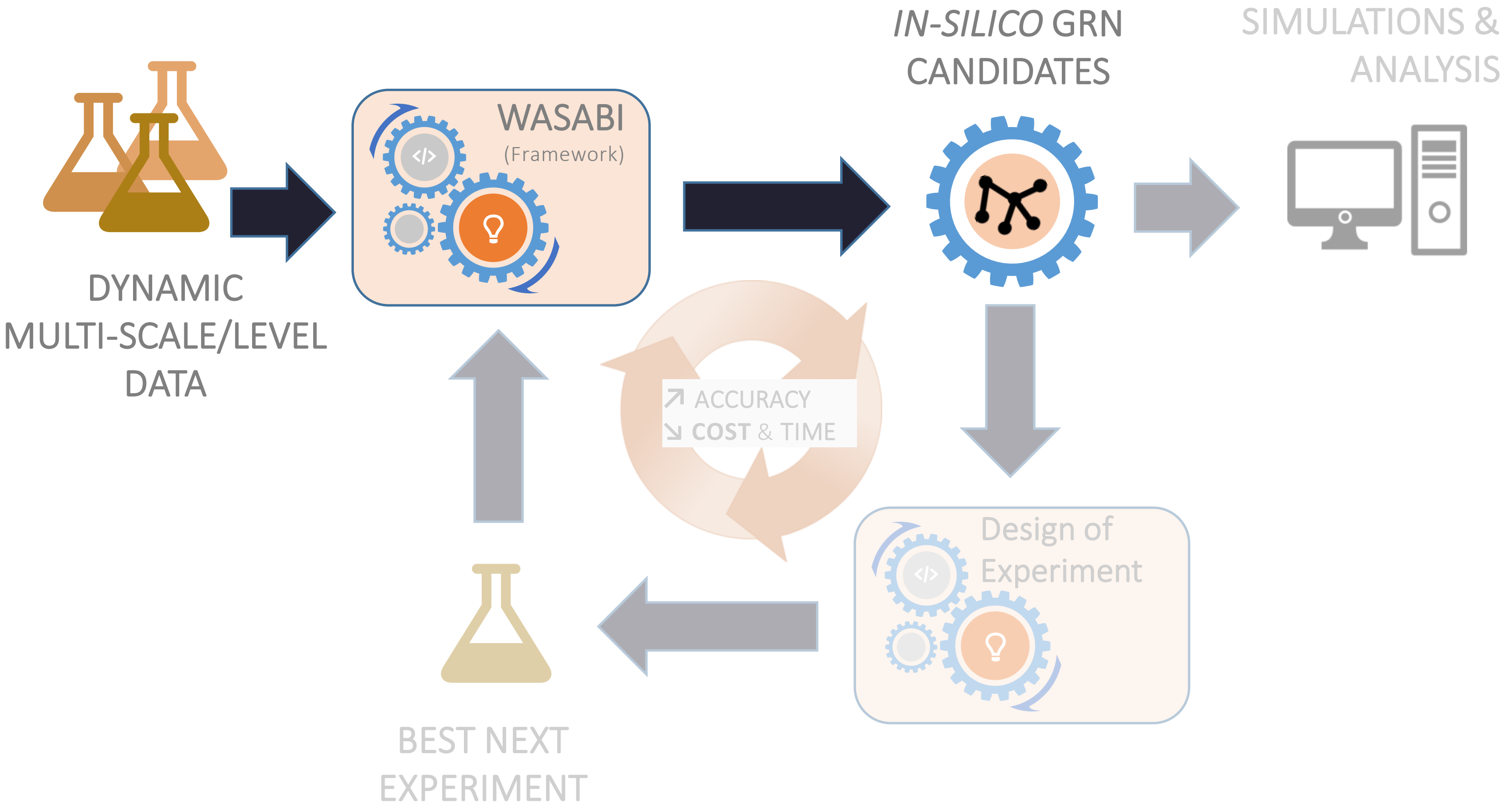


Set of 6 GRN candidates



Bonnaffoux et al. (2019), BMC Bioinformatics

WASABI : IN-VITRO VALIDATION



DYNAMIC MULTI-SCALE/LEVEL DATA

WASABI (Framework)

IN-SILICO GRN CANDIDATES

SIMULATIONS & ANALYSIS

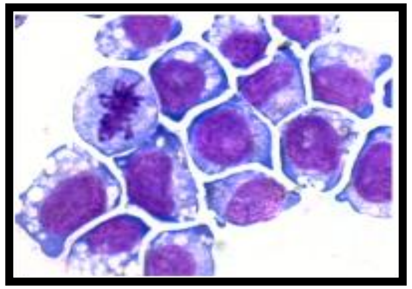
↗ ACCURACY
↘ **COST** & TIME

Design of Experiment

BEST NEXT EXPERIMENT

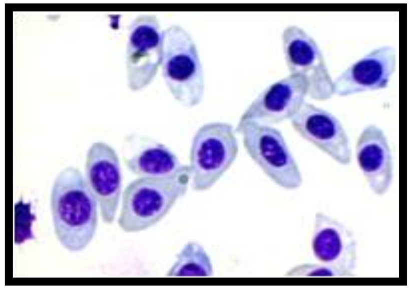
BIOLOGICAL *IN-VITRO* MODEL + DATA

self-renewal



T2EC

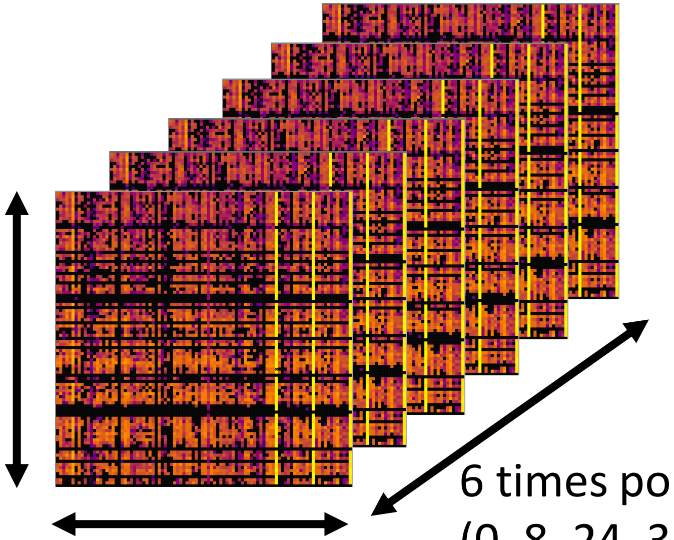
Differentiation



Erythrocytes

Single Cell RTqPCR data

(Richard et al. (2016) PLoS Biology)

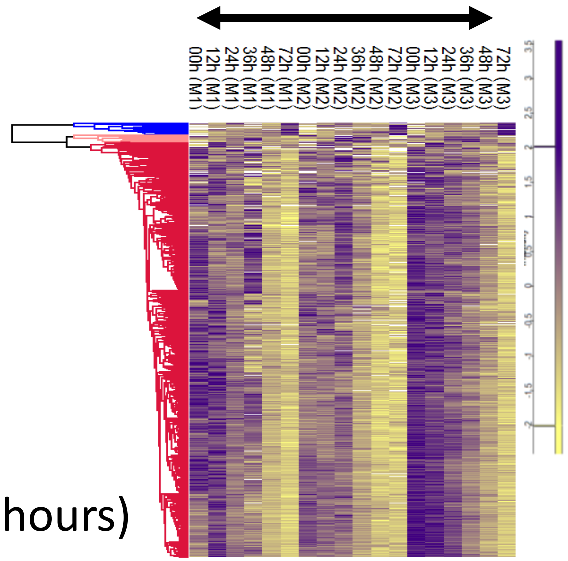


6 times points
(0, 8, 24, 33, 48 and 72 hours)

+

Pop. Proteomic data

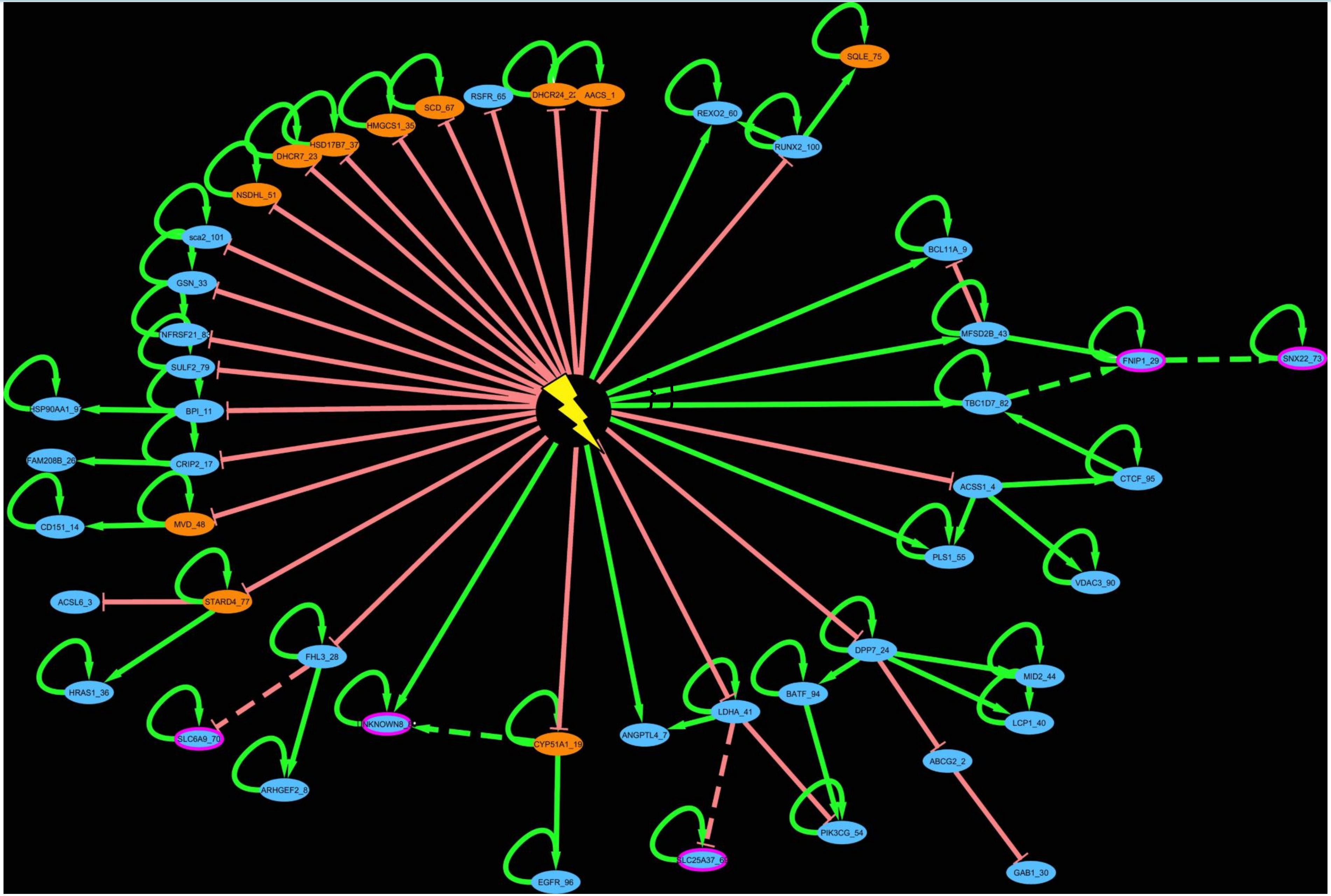
6 times points (0, 8, 24, 33, 48 and 72 hours)



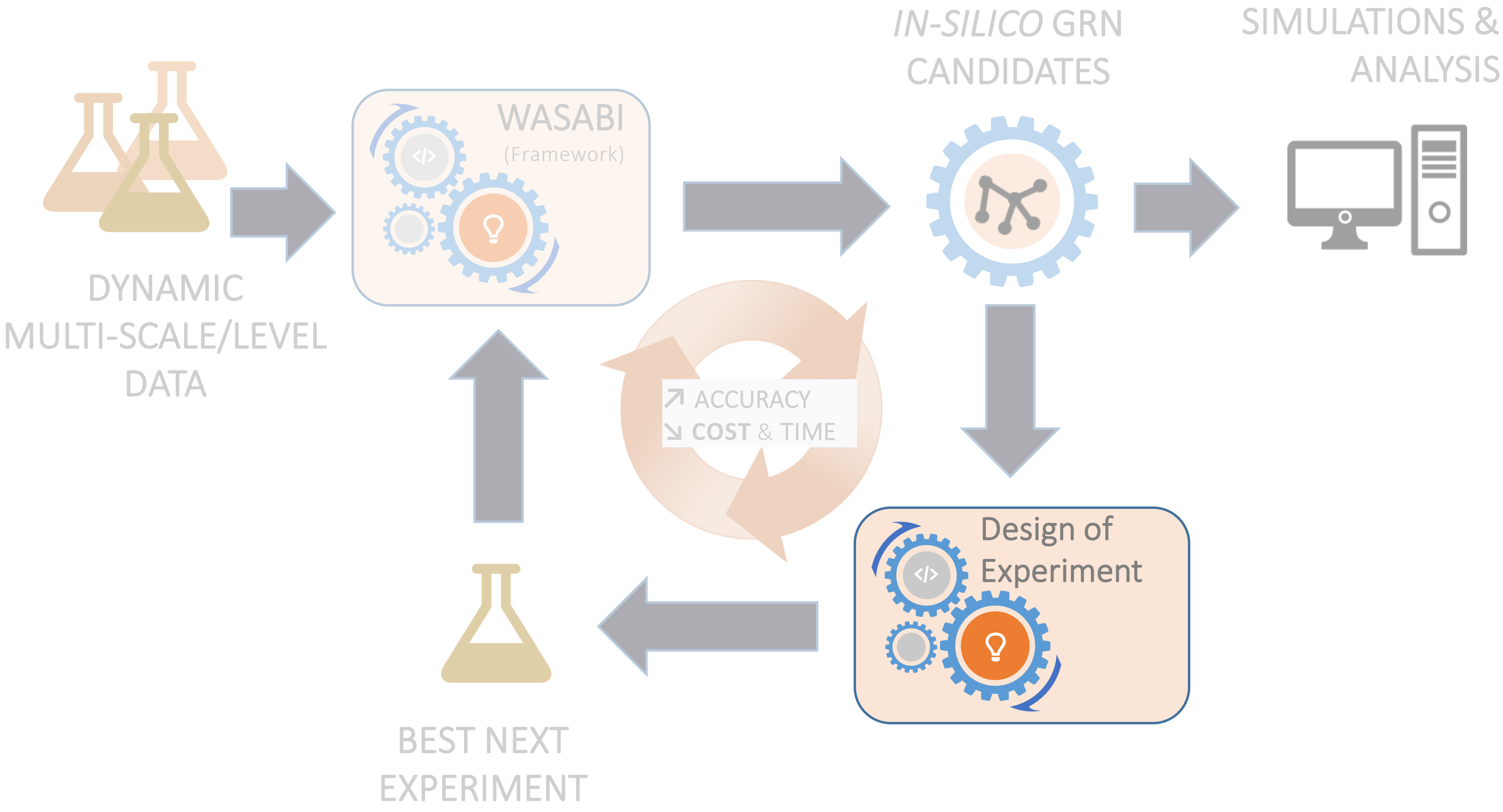
The output from WASABI (1)

364 candidates

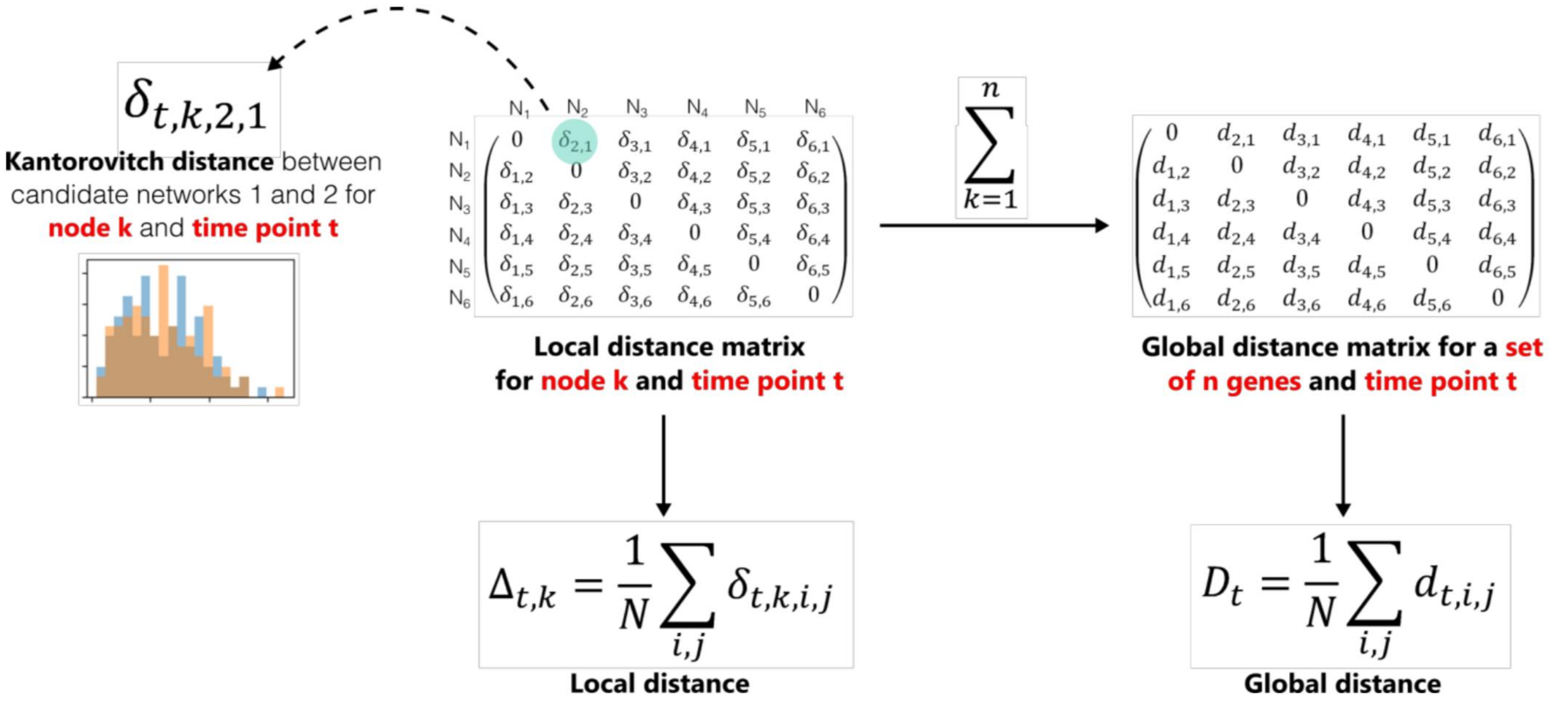
The output from WASABI (2)



DESIGN OF EXPERIMENT



DESIGN OF EXPERIMENT: define a distance between networks



$\delta_{t,k,2,1}$

Kantorovitch distance between candidate networks 1 and 2 for **node k** and **time point t**

$$\begin{array}{c|cccccc} & N_1 & N_2 & N_3 & N_4 & N_5 & N_6 \\ \hline N_1 & 0 & \delta_{2,1} & \delta_{3,1} & \delta_{4,1} & \delta_{5,1} & \delta_{6,1} \\ N_2 & \delta_{1,2} & 0 & \delta_{3,2} & \delta_{4,2} & \delta_{5,2} & \delta_{6,2} \\ N_3 & \delta_{1,3} & \delta_{2,3} & 0 & \delta_{4,3} & \delta_{5,3} & \delta_{6,3} \\ N_4 & \delta_{1,4} & \delta_{2,4} & \delta_{3,4} & 0 & \delta_{5,4} & \delta_{6,4} \\ N_5 & \delta_{1,5} & \delta_{2,5} & \delta_{3,5} & \delta_{4,5} & 0 & \delta_{6,5} \\ N_6 & \delta_{1,6} & \delta_{2,6} & \delta_{3,6} & \delta_{4,6} & \delta_{5,6} & 0 \end{array}$$

Local distance matrix for node k and time point t

$$\xrightarrow{\sum_{k=1}^{n}}$$

$$\begin{pmatrix} 0 & d_{2,1} & d_{3,1} & d_{4,1} & d_{5,1} & d_{6,1} \\ d_{1,2} & 0 & d_{3,2} & d_{4,2} & d_{5,2} & d_{6,2} \\ d_{1,3} & d_{2,3} & 0 & d_{4,3} & d_{5,3} & d_{6,3} \\ d_{1,4} & d_{2,4} & d_{3,4} & 0 & d_{5,4} & d_{6,4} \\ d_{1,5} & d_{2,5} & d_{3,5} & d_{4,5} & 0 & d_{6,5} \\ d_{1,6} & d_{2,6} & d_{3,6} & d_{4,6} & d_{5,6} & 0 \end{pmatrix}$$

Global distance matrix for a set of n genes and time point t

$$\Delta_{t,k} = \frac{1}{N}\sum_{i,j} \delta_{t,k,i,j}$$

Local distance

$$D_t = \frac{1}{N}\sum_{i,j} d_{t,i,j}$$

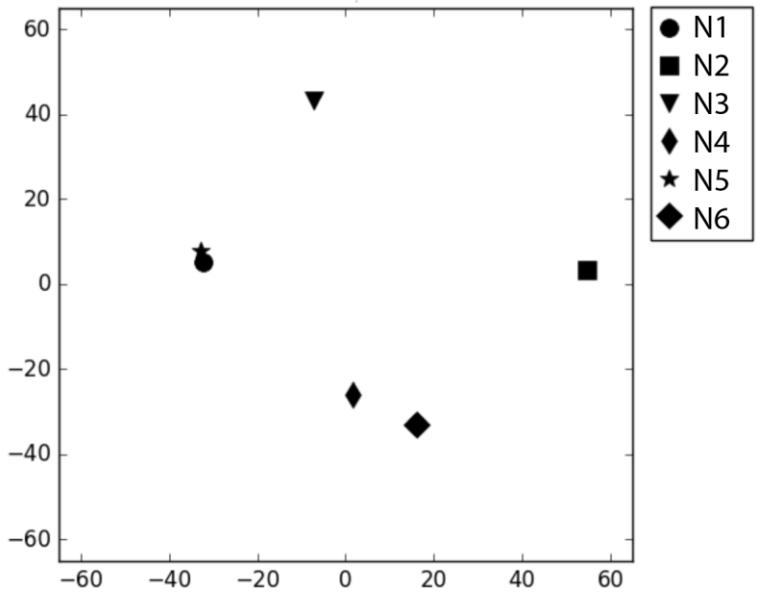
Global distance

DESIGN OF EXPERIMENT: project in a 2D space

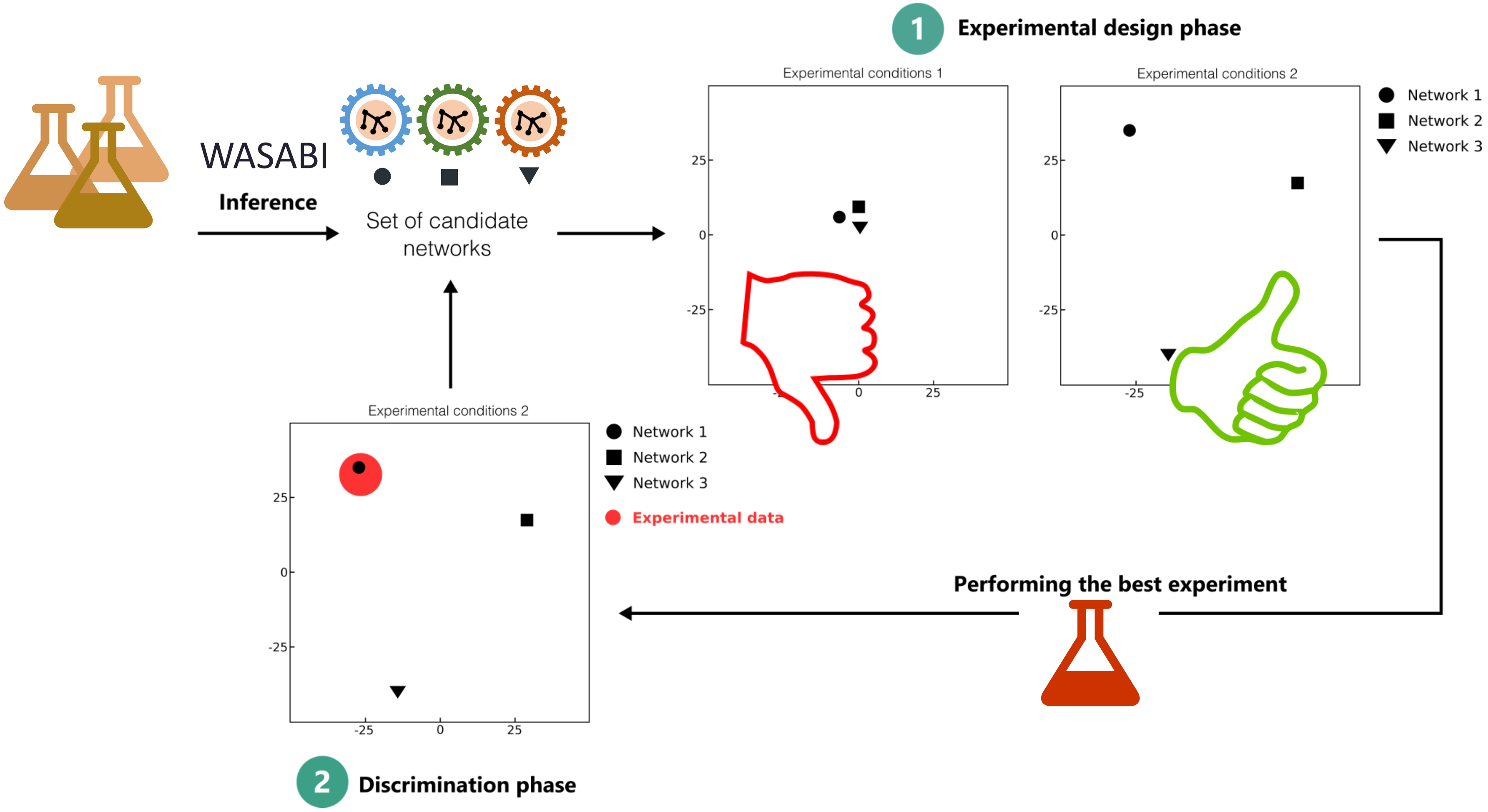
$$\begin{array}{c|cccccc} & N1 & N2 & N3 & N4 & N5 & N6 \\ \hline N1 & 0 & d_{2,1} & d_{3,1} & d_{4,1} & d_{5,1} & d_{6,1} \\ N2 & d_{1,2} & 0 & d_{3,2} & d_{4,2} & d_{5,2} & d_{6,2} \\ N3 & d_{1,3} & d_{2,3} & 0 & d_{4,3} & d_{5,3} & d_{6,3} \\ N4 & d_{1,4} & d_{2,4} & d_{3,4} & 0 & d_{5,4} & d_{6,4} \\ N5 & d_{1,5} & d_{2,5} & d_{3,5} & d_{4,5} & 0 & d_{6,5} \\ N6 & d_{1,6} & d_{2,6} & d_{3,6} & d_{4,6} & d_{5,6} & 0 \end{array}$$

Global distance matrix for a set of n genes and time point t

$\longrightarrow$ MDS $\longrightarrow$



DESIGN OF EXPERIMENT: plan *in silico* experiment



1 Experimental design phase

WASABI

Inference

Set of candidate networks

Experimental conditions 1

Experimental conditions 2

● Network 1
■ Network 2
▼ Network 3

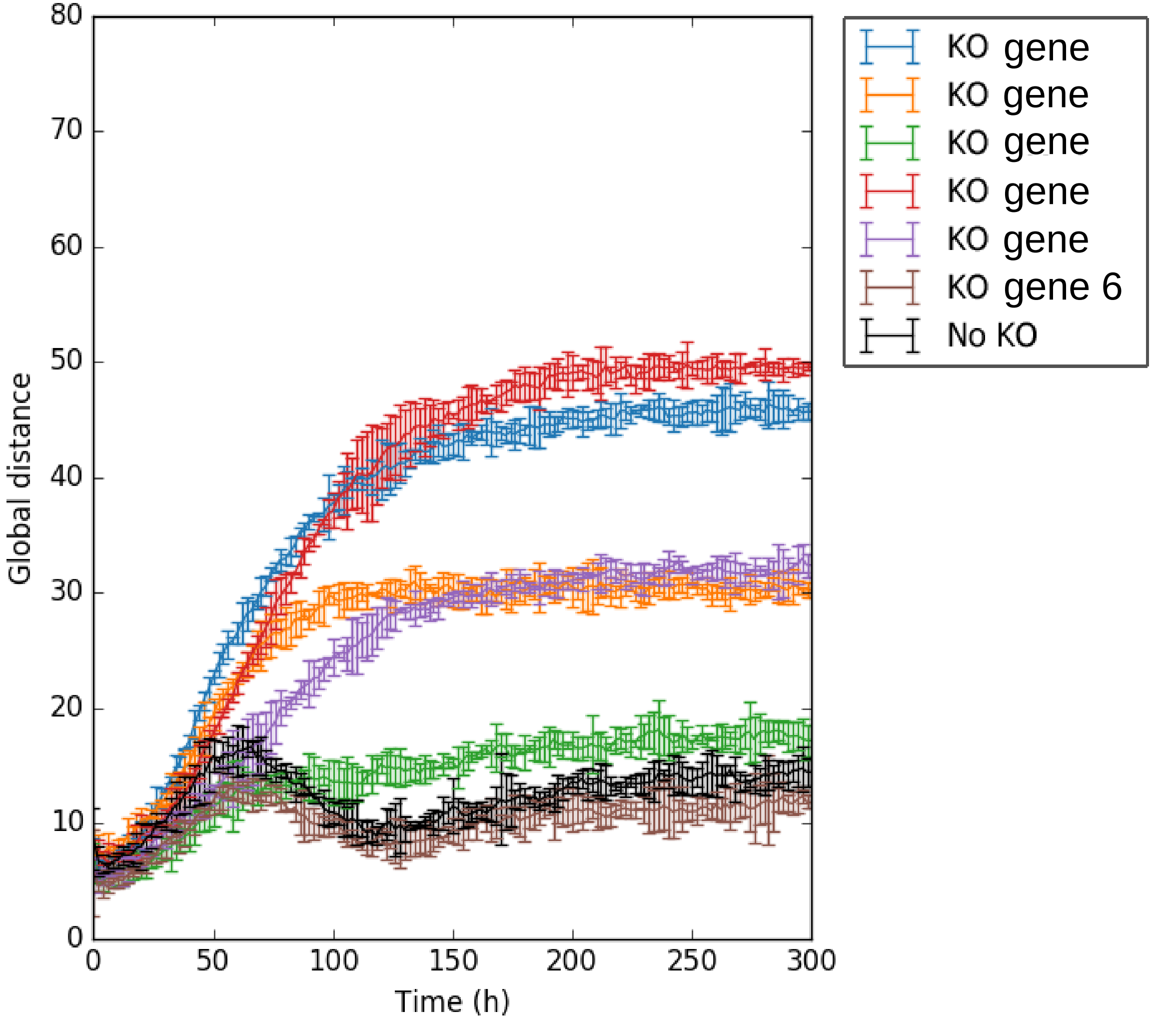
Performing the best experiment

Experimental conditions 2

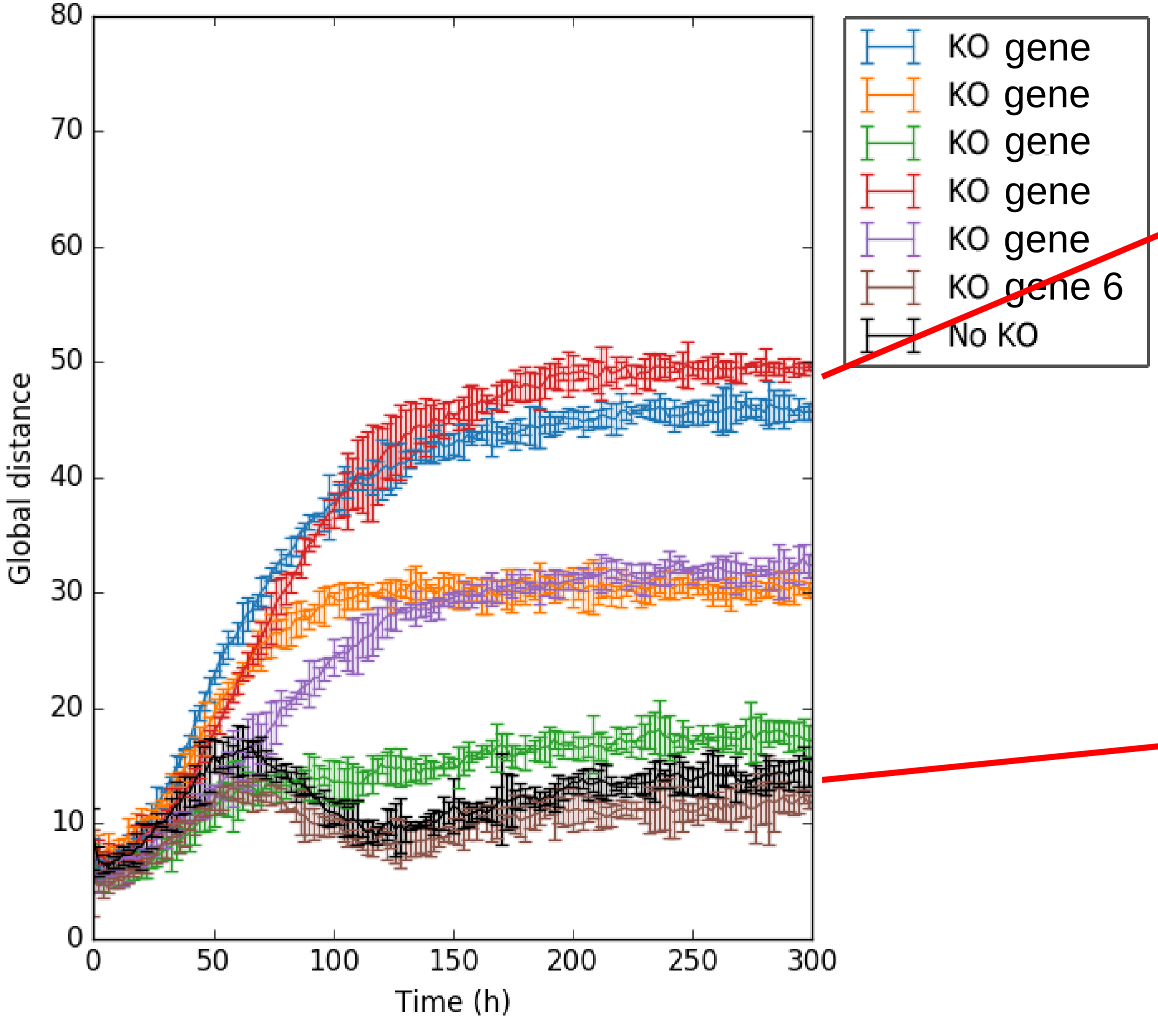
● Network 1
■ Network 2
▼ Network 3
● **Experimental data**

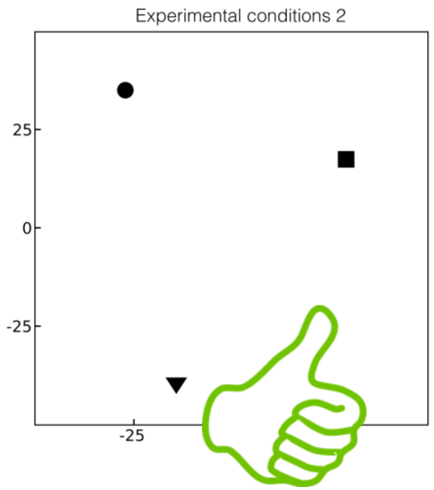
2 Discrimination phase

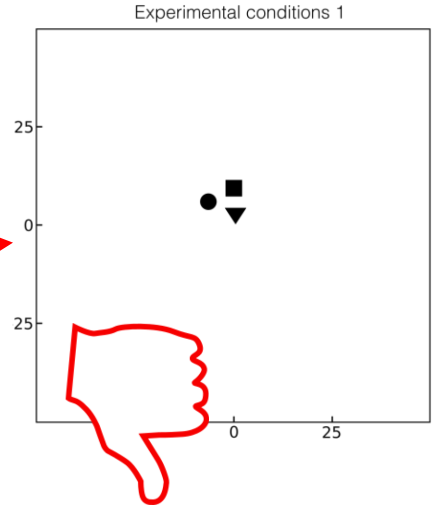
Knock-outs of specific genes **maximize the global distance** between the candidate networks



Knock-outs of specific genes maximize the global distance between the candidate networks



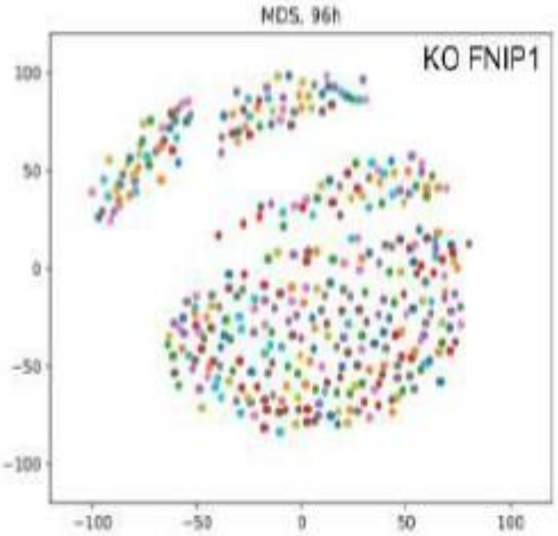


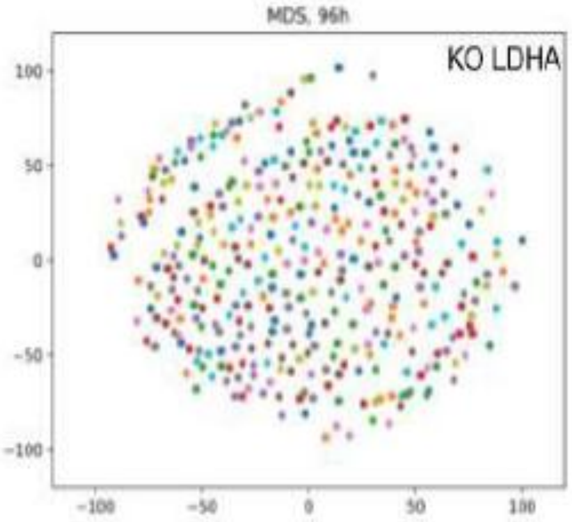


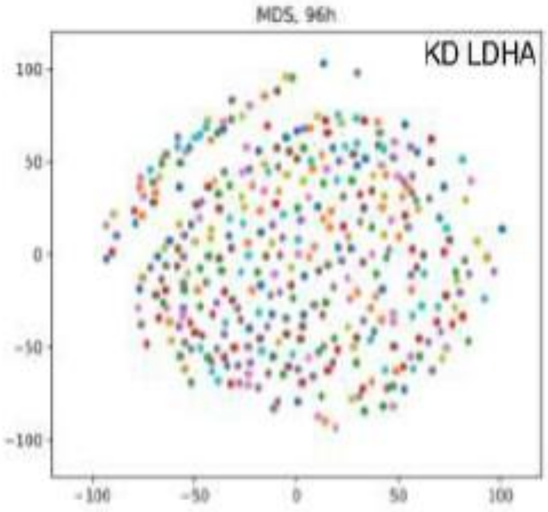
Best experimental conditions

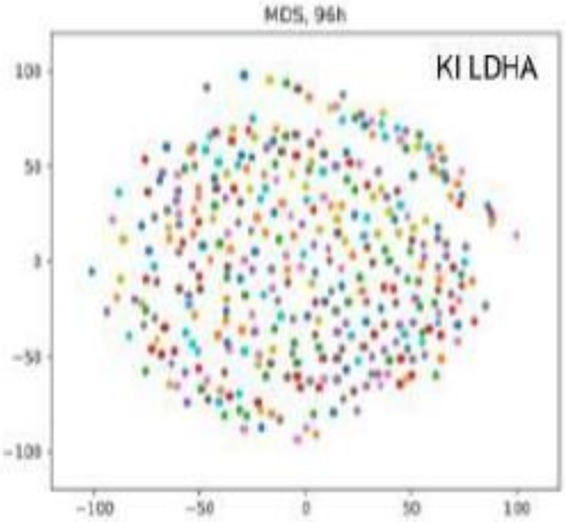
Type of experiment : KO gene 4

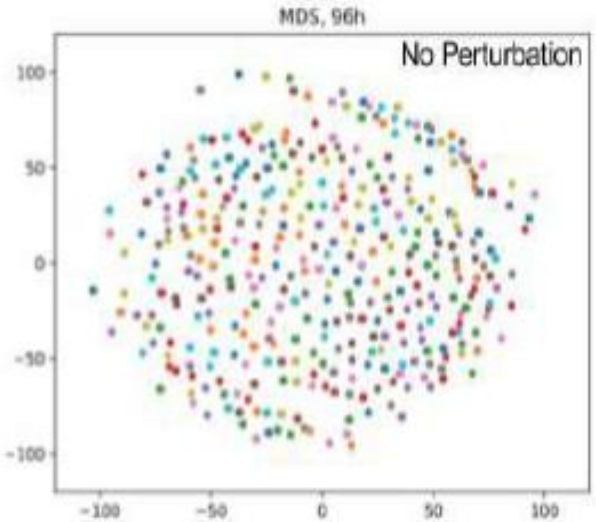
DOE APPLICATION ON IN-VITRO 364 GRN CANDIDATES



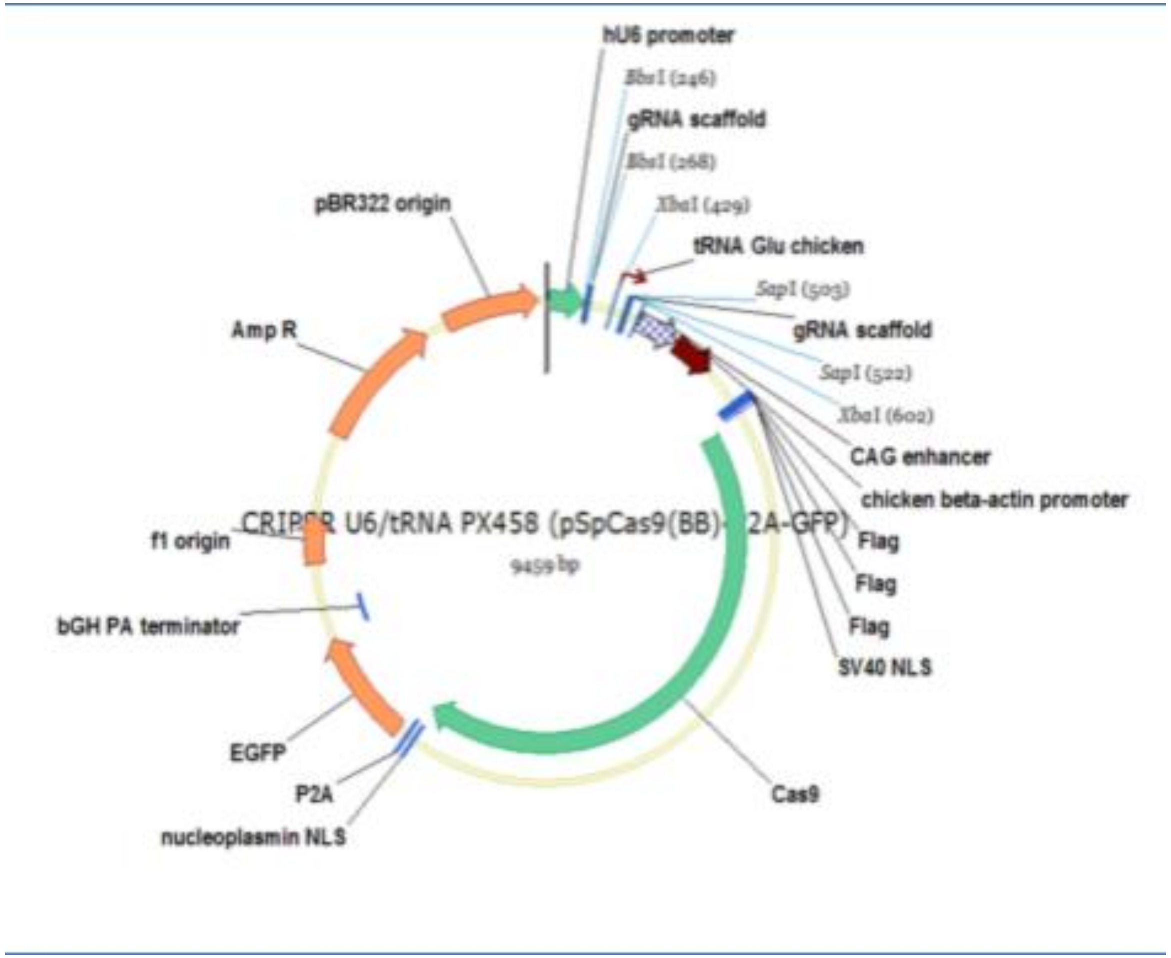


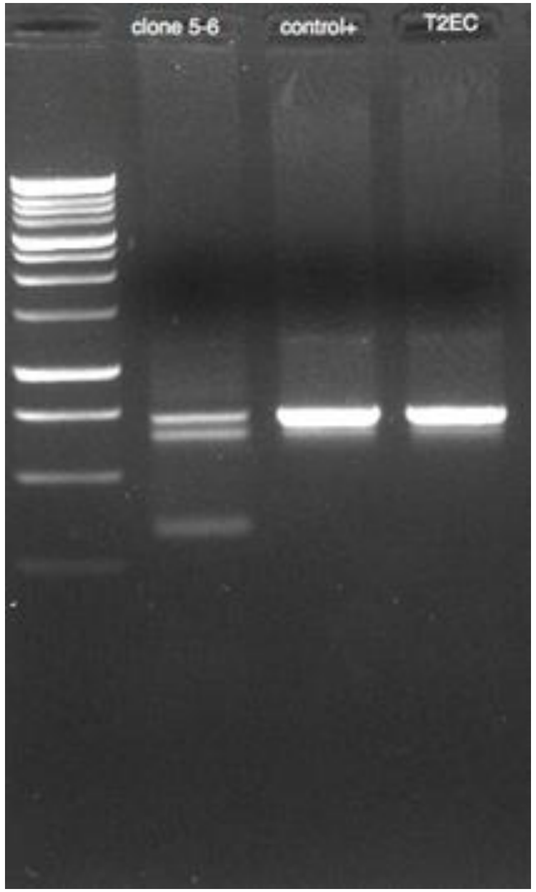




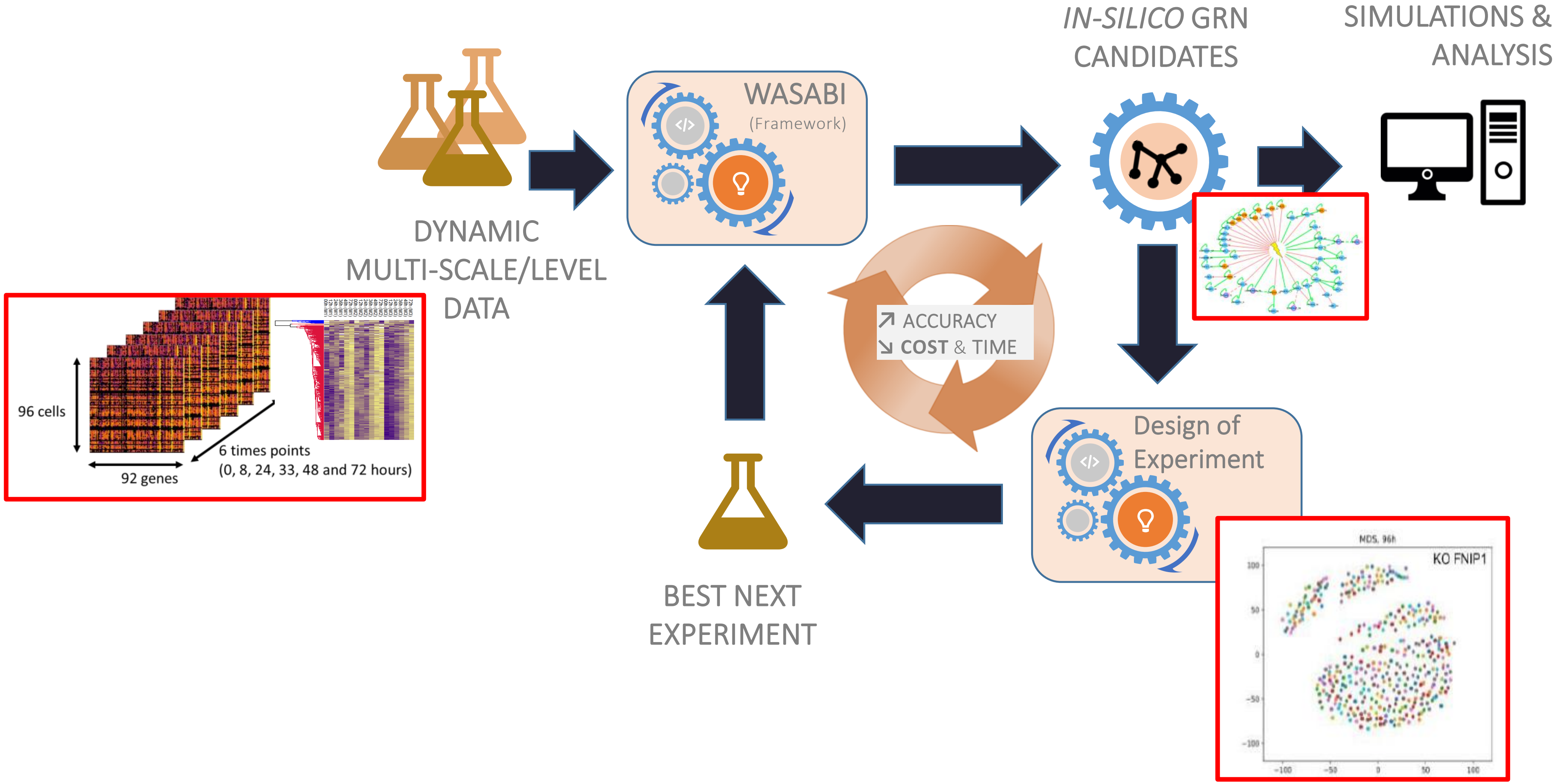


On going: CRSIPR-Cas9 KO of FNIP-1





GRN INFERENCE FRAMEWORK OVERVIEW



DYNAMIC
MULTI-SCALE/LEVEL
DATA

WASABI
(Framework)

IN-SILICO GRN
CANDIDATES

SIMULATIONS &
ANALYSIS

Design of
Experiment

↗ ACCURACY
↘ **COST** & TIME

BEST NEXT
EXPERIMENT

Special thank's

Thibault Espinasse (ICJ)
Sandrine Gonin-Giraud (LBMC)
Anissa Guillemin (LBMC)
Ulysse Herbach (LBMC/ICJ/Dracula)
Alice Hugues (LBMC)
Patrick Mayeux (Institut Cochin)
Angélique Richard (LBMC)
Elodie Vallin (LBMC)
Souad Zreika (LBMC)

BIOSYL
CCIN2P3
Institut des Systèmes Complexes
Complex Systems Institute
Rhône-Alpes
IXXI
AGENCE NATIONALE DE LA RECHERCHE
ANR

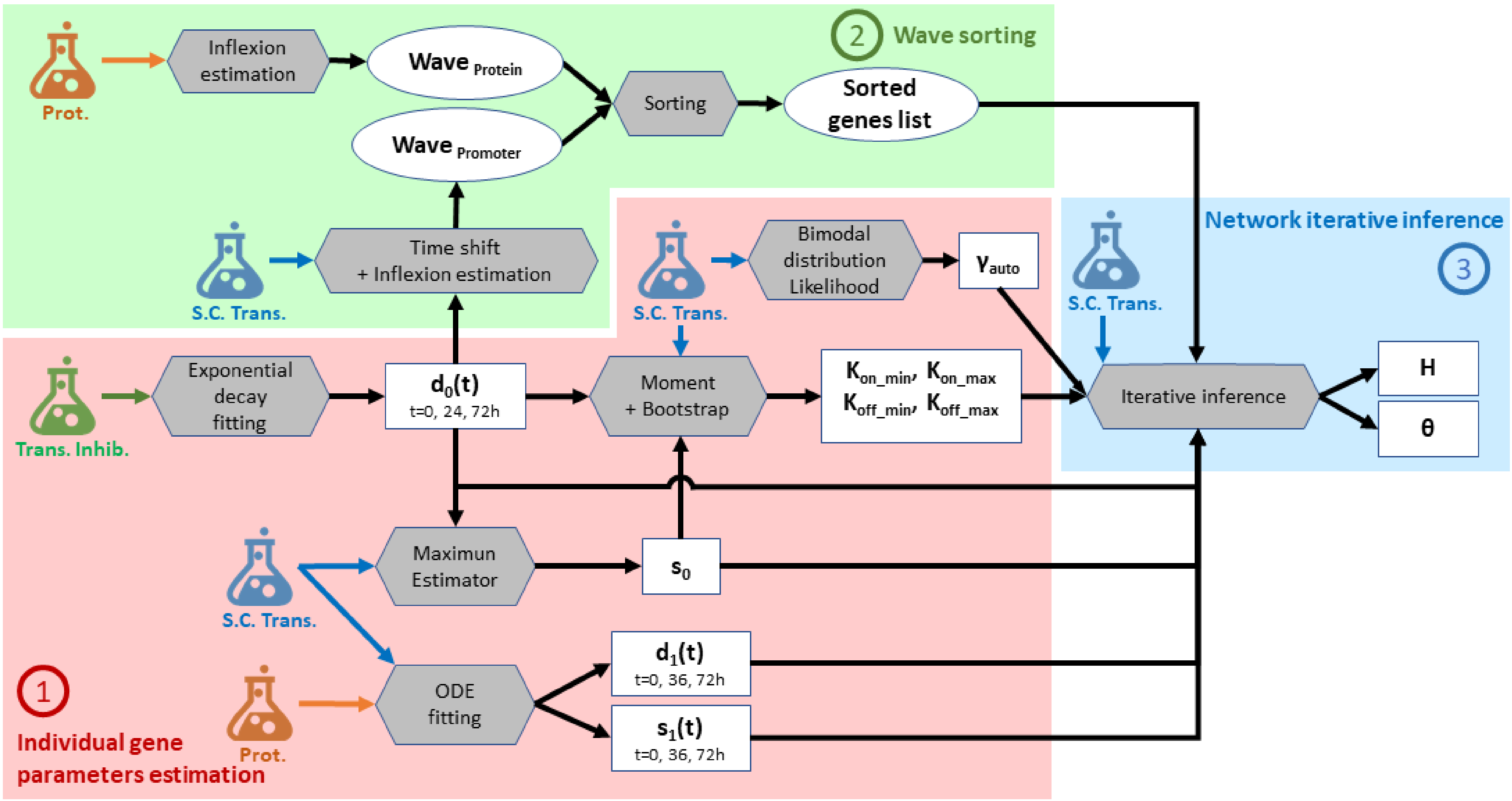
THANKS FOR LISTENING !

CONTACT

06 08 51 15 89
www.vidium-solutions.com
a.bonnaffoux@vidium-solutions.com

VIDIUM
SYSTEMS BIOLOGY SOLUTIONS

WASABI 3 STEPS FRAMEWORK



② Wave sorting

Prot.

Inflexion estimation

Wave $_{\text{Protein}}$

Wave $_{\text{Promoter}}$

Sorting

Sorted genes list

S.C. Trans.

Time shift + Inflexion estimation

S.C. Trans.

Bimodal distribution Likelihood

γ_{auto}

Network iterative inference

③

S.C. Trans.

Trans. Inhib.

Exponential decay fitting

$d_0(t)$

t=0, 24, 72h

Moment + Bootstrap

K_{on_min}, K_{on_max}

K_{off_min}, K_{off_max}

Iterative inference

H

θ

S.C. Trans.

Maximun Estimator

s_0

① Individual gene parameters estimation

Prot.

ODE fitting

$d_1(t)$

t=0, 36, 72h

$s_1(t)$

t=0, 36, 72h