Large data sets and complex models: A view from Systems Biology

Bärbel Finkenstädt, Department of Statistics, University of Warwick

OxWaSP workshop October 2015

WARWICK
THE UNIVERSITY OF WARWICK

Joint work with

Mathematical and Statistical Modelling:

Sylvia Calderazzo, Simone Tiberi, Kirsty Hey, Hiroshi Momiji, Dafyd Jenkins, George Minas, David Rand (University of Warwick)

Experimental Collaboration:

PRL expression in pituitary gland (Julien Davis, Mike White, University of Manchester)

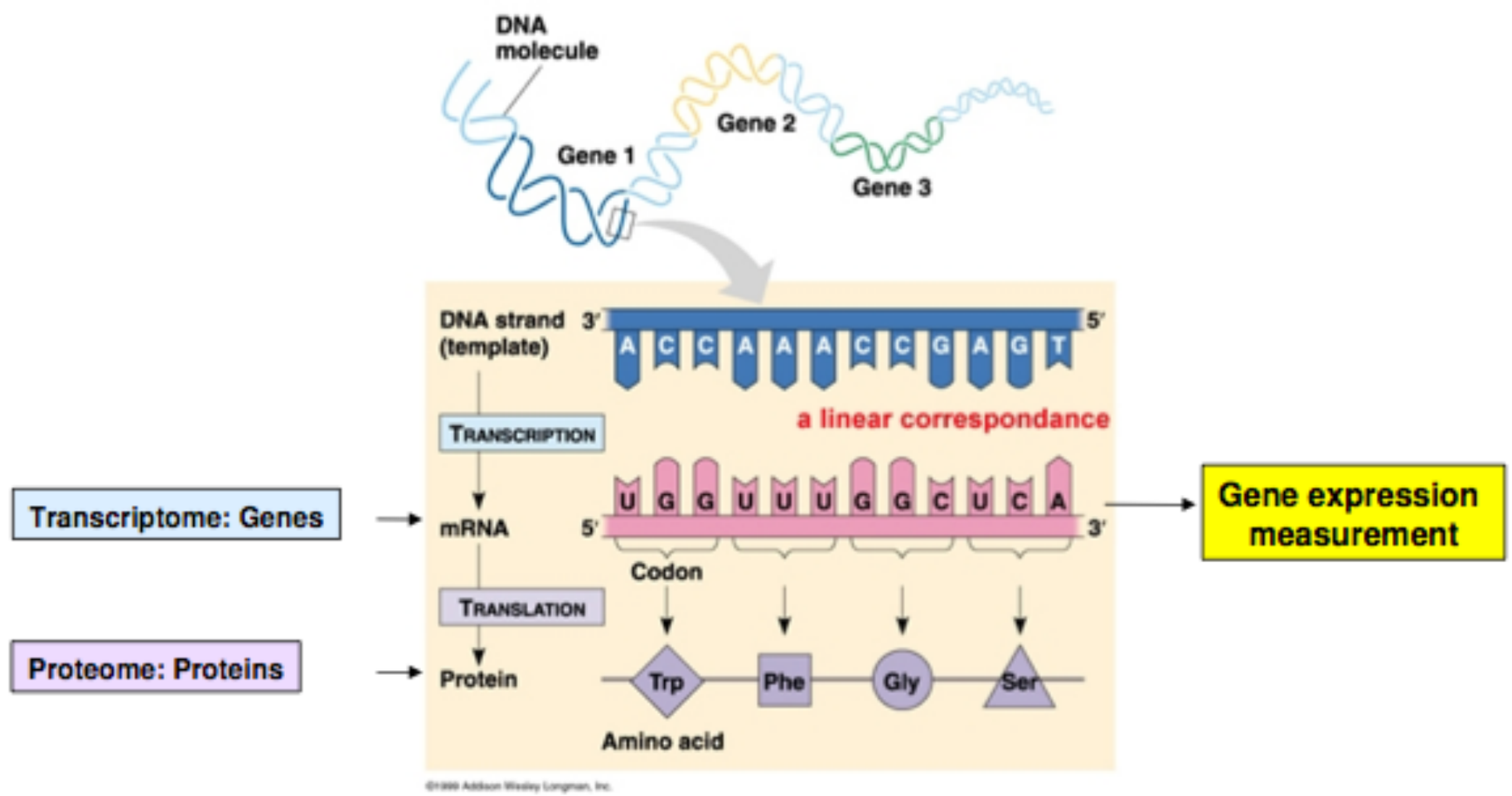
Chronotherapy and Cancer Research Unit (Francis Lévi and group), Warwick Medical School and INSERM, Paris

Molecular Neurobiology of circadian timing (Michael Hastings and group), MRC Laboratory of Molecular Biology, Cambridge

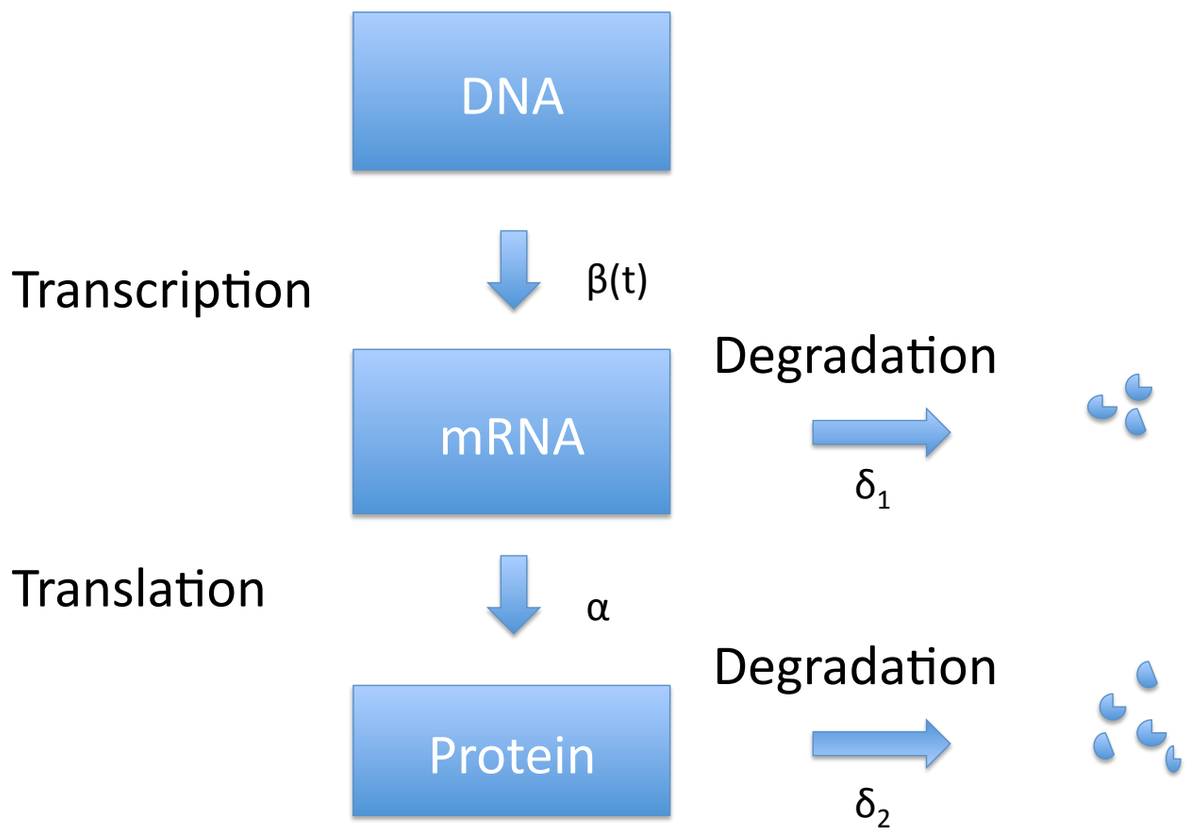
PRESTA Consortium (University of Warwick)

Funding: BBSRC, EPSRC, MRC, EU (BIOSIM), Wellcome

Gene Expression



Standard Model of Gene Expression (Single Gene)



Standard Model of Gene Expression (Single Gene, ODE Version)

$$\frac{dm(t)}{dt} = \beta(t) - \delta_m m(t),$$

$$\frac{dp(t)}{dt} = \alpha m(t) - \delta_p p(t)$$

where

$m(t)$: concentration of mRNA

$p(t)$: concentration of protein

$\beta(t)$: transcription rate of mRNA

δ_m, δ_p : degradation rate of mRNA, protein

Standard Model of gene expression (Stochastic version, single gene)

$$X(t) = \begin{pmatrix} X_m(t) \\ X_p(t) \end{pmatrix}$$

4 reactions (transcription, degradation mRNA, translation, degradation protein)

$$v_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, v_2 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, v_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, v_4 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$$

A reaction of type j changes $X(t)$ to $X(t) + v_j$.
Each reaction occurs at a rate $w_j(X(t))$.

Event	Effect	Transition Rate
Transcription	$(X_m, X_p) \to (X_m + 1, X_p)$	$w_1 = \beta(t)$
Degradation of mRNA	$(X_m, X_p) \to (X_m - 1, X_p)$	$w_2 = \delta_m X_m(t)$
Translation	$(X_m, X_p) \to (X_m, X_p + 1)$	$w_3 = \alpha X_m(t)$
Degradation of protein	$(X_m, X_p) \to (X_m, X_p - 1)$	$w_4 = \delta_p X_p(t)$

Table 1: Summary of reactions in the standard model of gene expression

Reaction networks constitute continuous time Markov jump processes and thus satisfy the Chapman-Kolmogorov equation for which one can obtain the forward form known as the master equation (ME) describing the evolution of the probability $P(X_m = n_1, X_p = n_2; t)$.

Although an exact numerical simulation algorithm is provided (Gillespie,1977) the ME is rarely tractable and hence an explicit formula for the exact likelihood is not available for parameter inference.

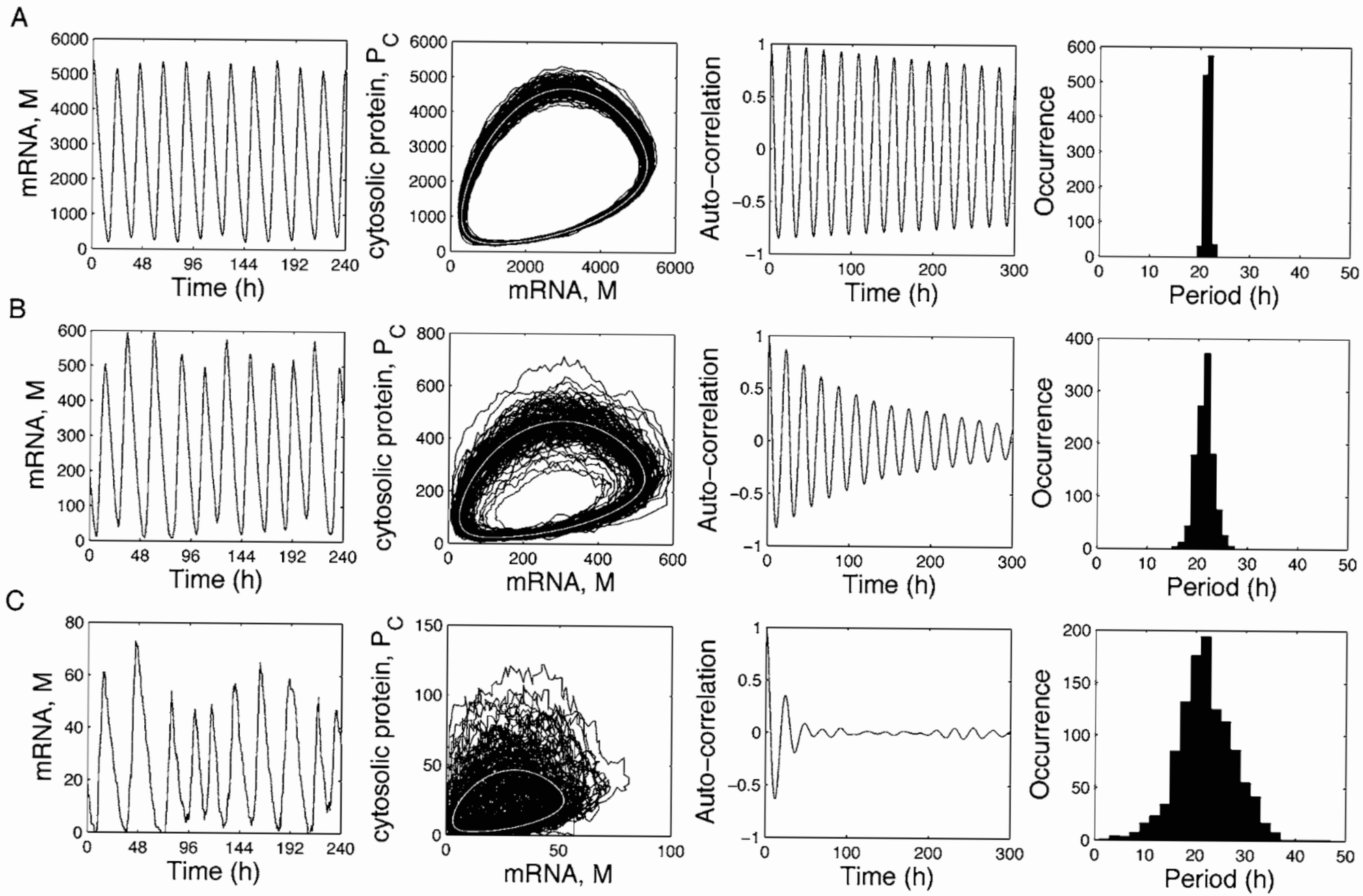
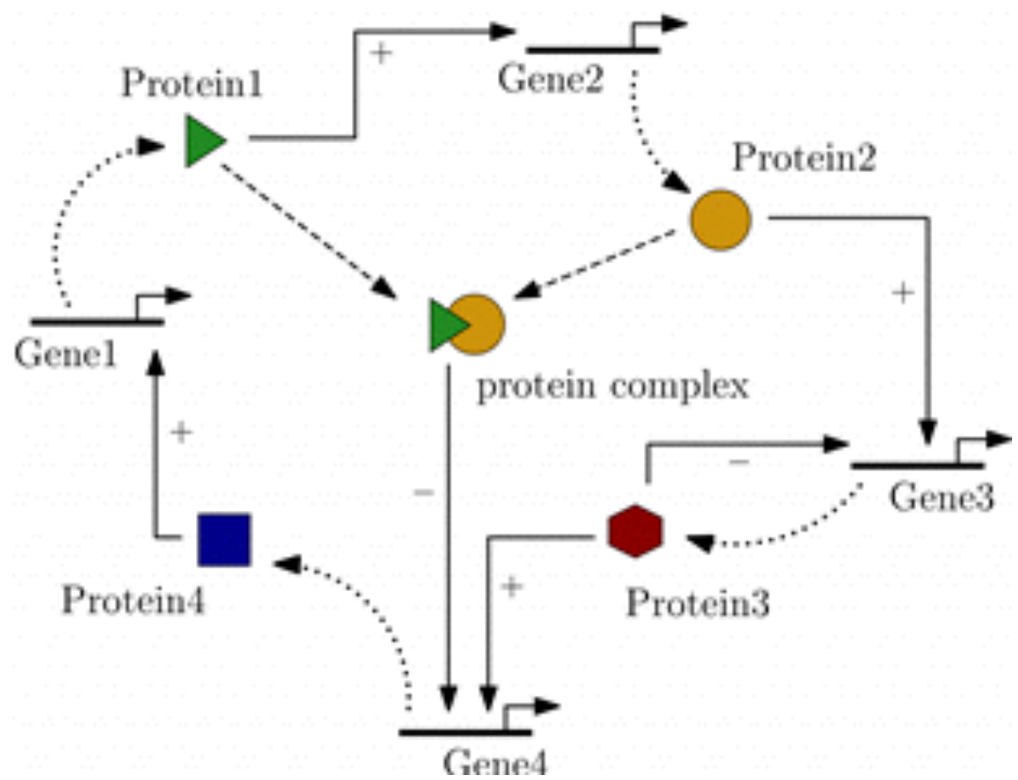
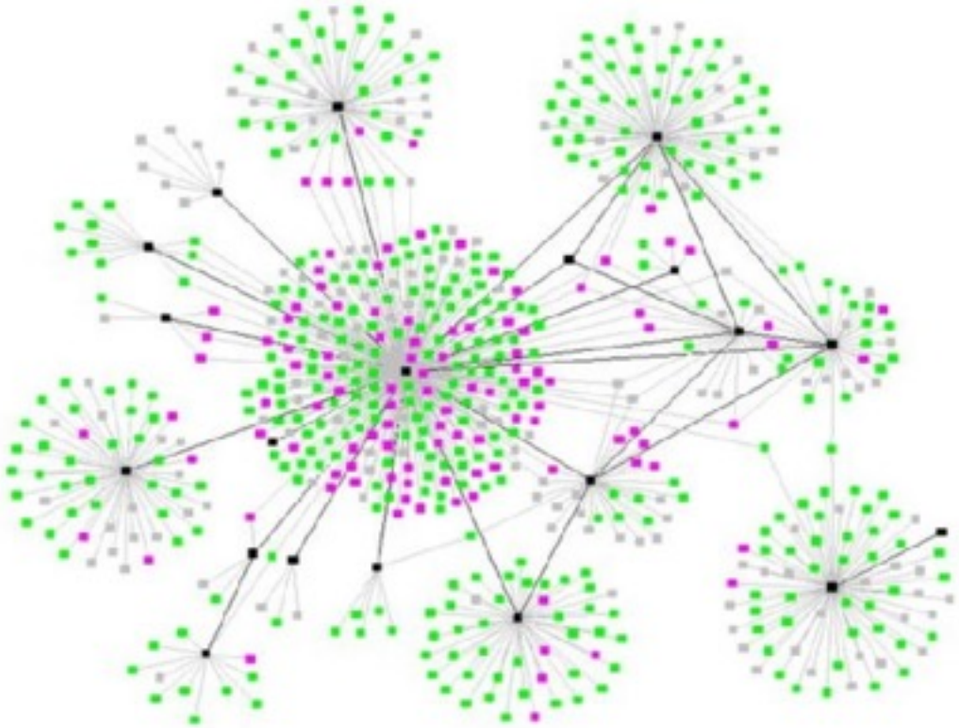
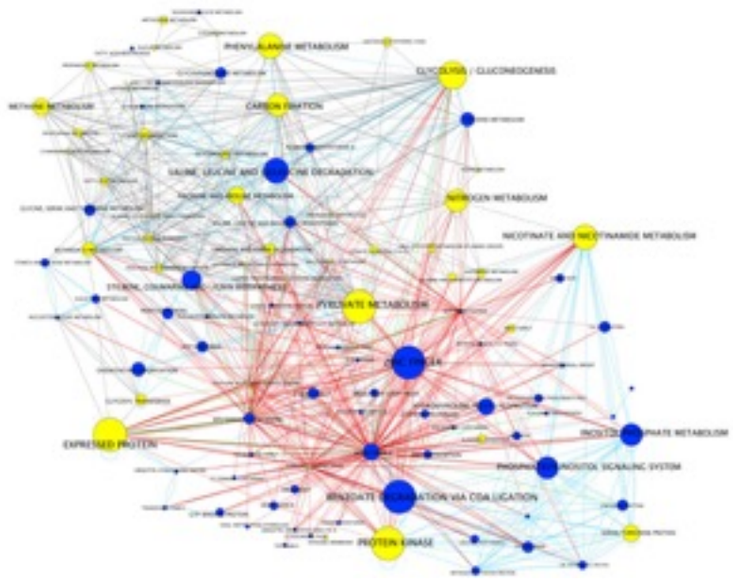


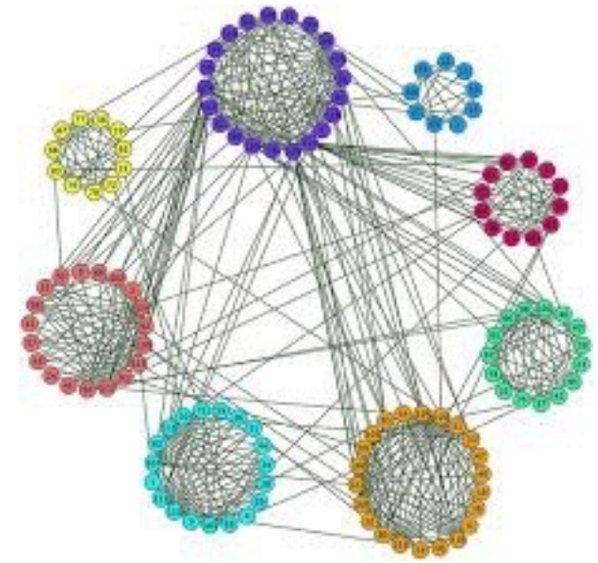
FIG. 2. Oscillations obtained by numerical simulation of the stochastic model for circadian rhythms with the Gillespie algorithm. The panels show oscillations of mRNA concentration, M (left column), limit cycles (second column), autocorrelation function (third column), and the period distribution (fourth column), for (A) $\Omega=1000$, (B) $\Omega=100$, and (C) $\Omega=10$. The autocorrelation and period histograms have been calculated on a time series of 25 000 h, i.e., more than 1000 cycles. The white curve on the phase plane corresponds to the deterministic limit cycle. The deterministic oscillations have a period of 22 h. Parameter values are: $v_s=1.6$ nM h^{-1}, $K_I=1$ nM, $n=4$, $v_m=0.505$ nM h^{-1}, $K_m=0.5$ nM, $k_s=0.5$ h^{-1}, $v_d=1.4$ nM h^{-1}, $K_d=0.13$ nM, $k_1=0.5$ nM h^{-1}, $k_2=0.6$ nM h^{-1}.

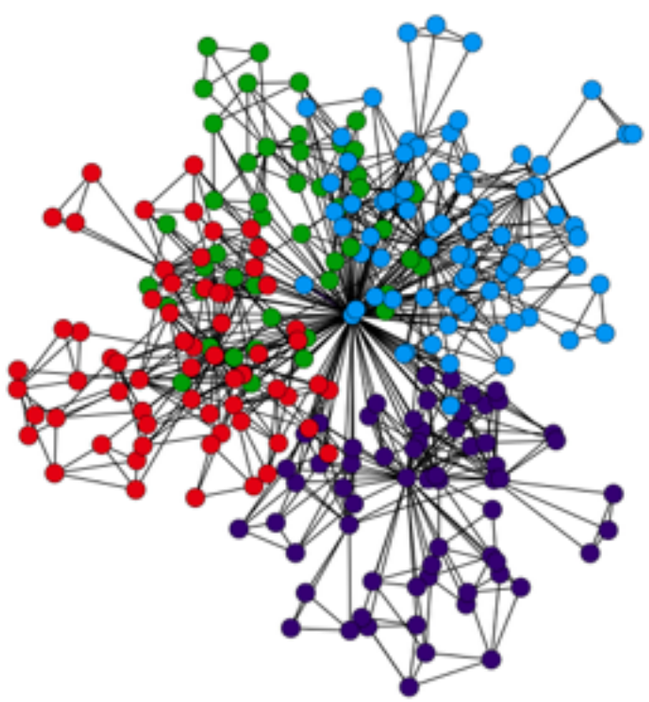
Gene Networks

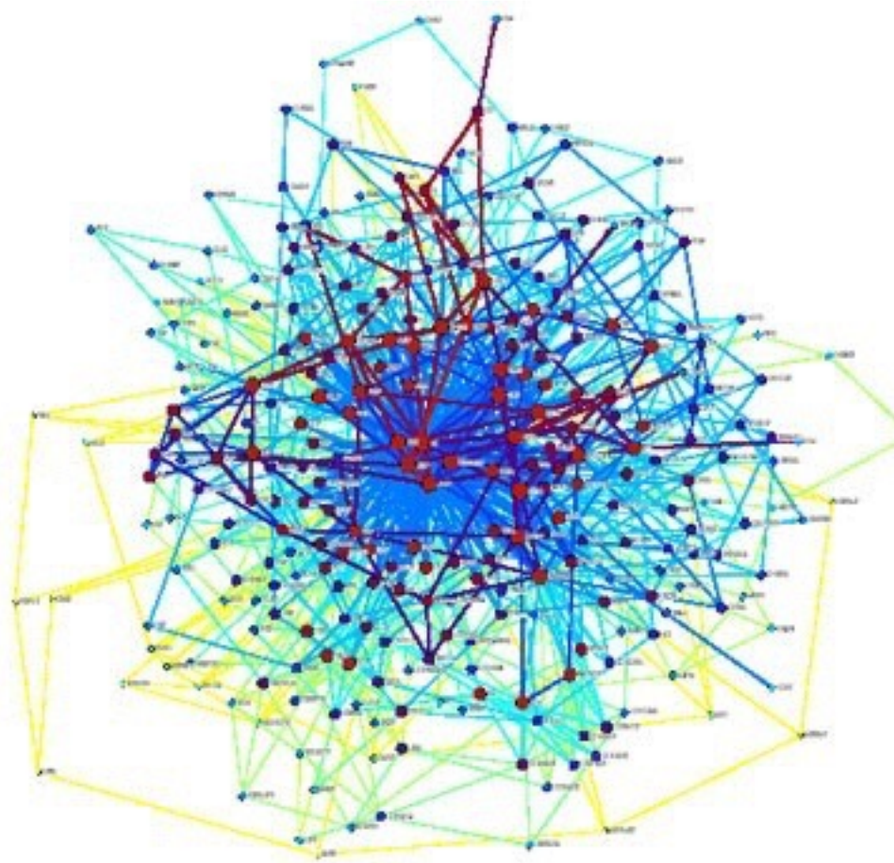





PHENYLALANINE METABOLISM
GLYCOLYSIS / GLUCONEOGENESIS
CARBON FIXATION
NITROGEN METABOLISM
NICOTINATE AND NICOTINAMIDE METABOLISM
PYRUVATE METABOLISM
EXPRESSED PROTEIN
PHOSPHOINOSITOL SIGNALING SYSTEM
PROTEIN KINASE







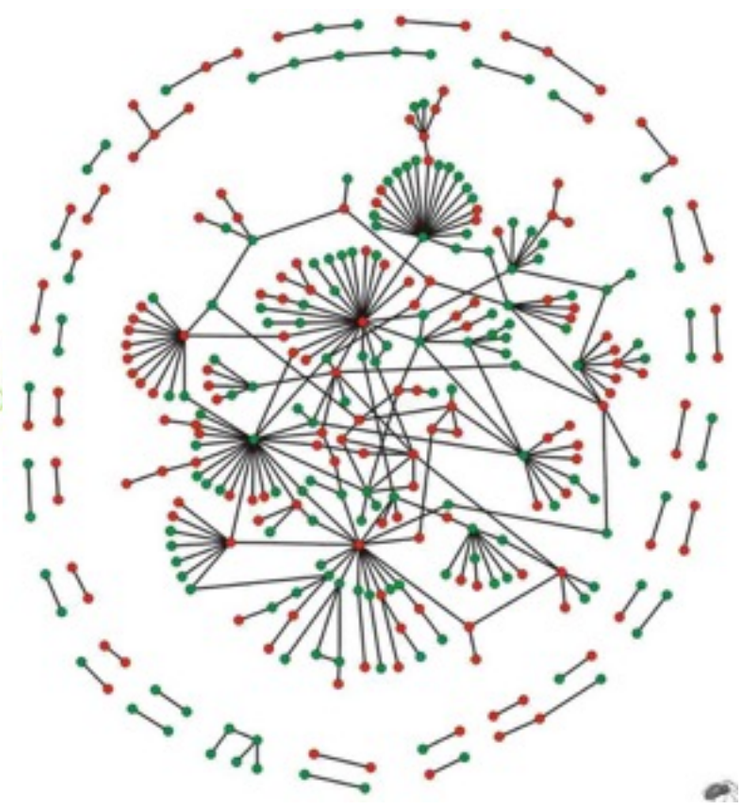
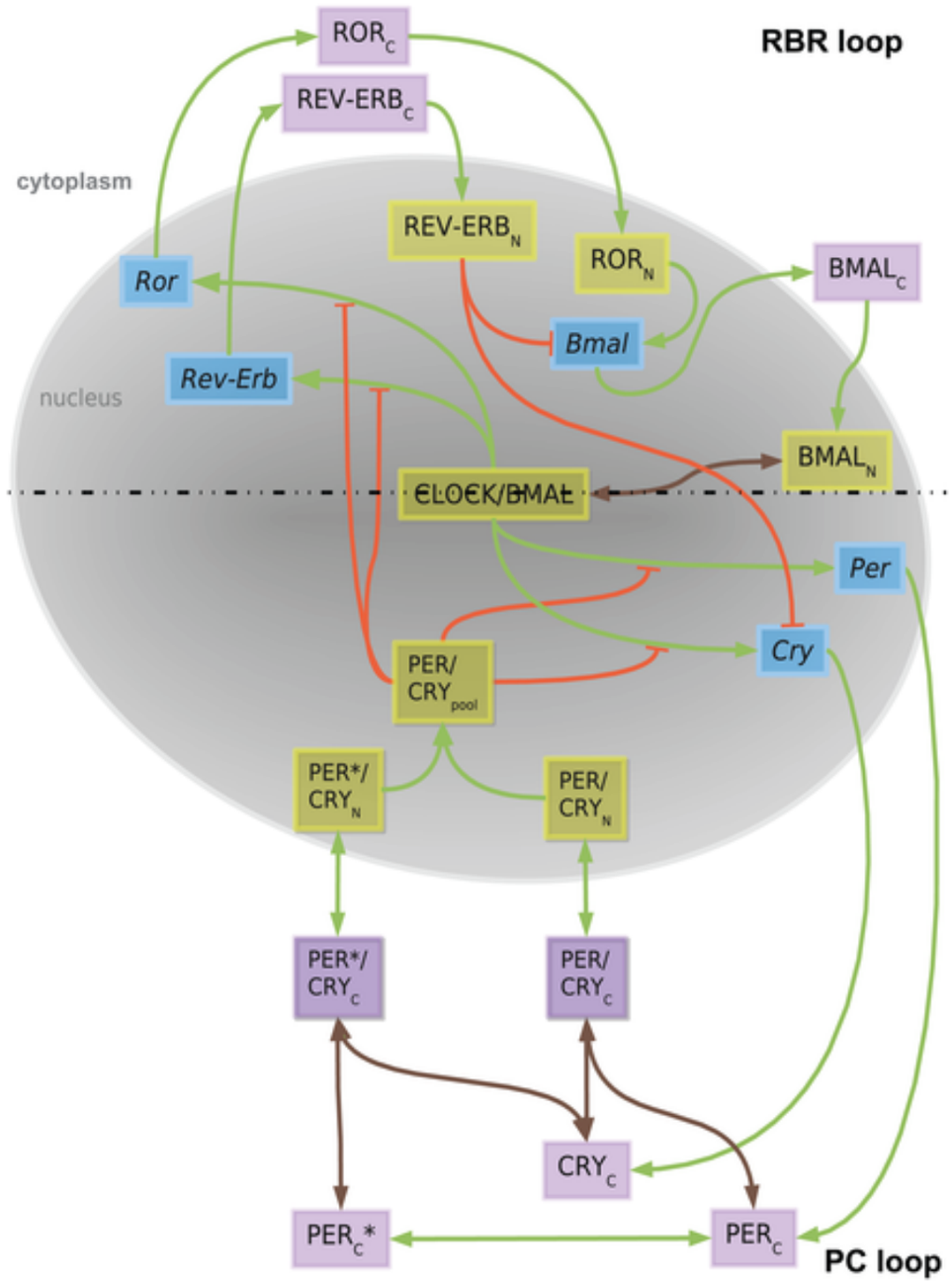


Figure 1. A model for the mammalian circadian clock.



Relógio A, Westermark PO, Wallach T, Schellenberg K, Kramer A, et al. (2011) Tuning the Mammalian Circadian Clock: Robust Synergy of Two Loops. PLoS Comput Biol 7(12): e1002309. doi:10.1371/journal.pcbi.1002309
http://127.0.0.1:8081/ploscompbiol/article?id=info:doi/10.1371/journal.pcbi.1002309

CLOCK/BMAL

$$\frac{dx1}{dt} = kf_{x1}x7 - kd_{x1}x1 - d_{x1}x1 \quad (1)$$

Rev-Erb

$$\frac{dy3}{dt} = V_{3\max}\frac{1+g\left(\frac{x1}{k_{t3}}\right)^{v}}{1+\left(\frac{PC}{k_{i3}}\right)^{w}\left(\frac{x1}{k_{t3}}\right)^{v}+\left(\frac{x1}{k_{t3}}\right)^{v}} - d_{y3}y3 \quad (2)$$

Ror

$$\frac{dy4}{dt} = V_{4\max}\frac{1+h\left(\frac{x1}{k_{t4}}\right)^{p}}{1+\left(\frac{PC}{k_{i4}}\right)^{q}\left(\frac{x1}{k_{t4}}\right)^{p}+\left(\frac{x1}{k_{t4}}\right)^{p}} - d_{y4}y4 \quad (3)$$

REV-ERB$_C$

$$\frac{dz6}{dt} = k_{p3}(y3+y3_0) - ki_{z6}z6 - d_{z6}z6 \quad (4)$$

ROR$_C$

$$\frac{dz7}{dt} = k_{p4}(y4+y4_0) - ki_{z7}z7 - d_{z7}z7 \quad (5)$$

REV-ERB$_N$

$$\frac{dx5}{dt} = ki_{z6}z6 - d_{x5}x5 \quad (6)$$

ROR$_N$

$$\frac{dx6}{dt} = ki_{z7}z7 - d_{x6}x6 \quad (7)$$

Bmal

$$\frac{dy5}{dt} = V_{5\max}\frac{1+i\left(\frac{x6}{k_{t5}}\right)^{n}}{1+\left(\frac{x5}{k_{i5}}\right)^{m}+\left(\frac{x6}{k_{t5}}\right)^{n}} - d_{y5}y5 \quad (8)$$

BMAL$_C$

$$\frac{dz8}{dt} = k_{p5}(y5+y5_0) - ki_{z8}z8 - d_{z8}z8$$

BMAL$_N$

$$\frac{dx7}{dt} = ki_{z8}z8 + kd_{x1}x1 - kf_{x1}x7 - d_{x7}x7$$

Per

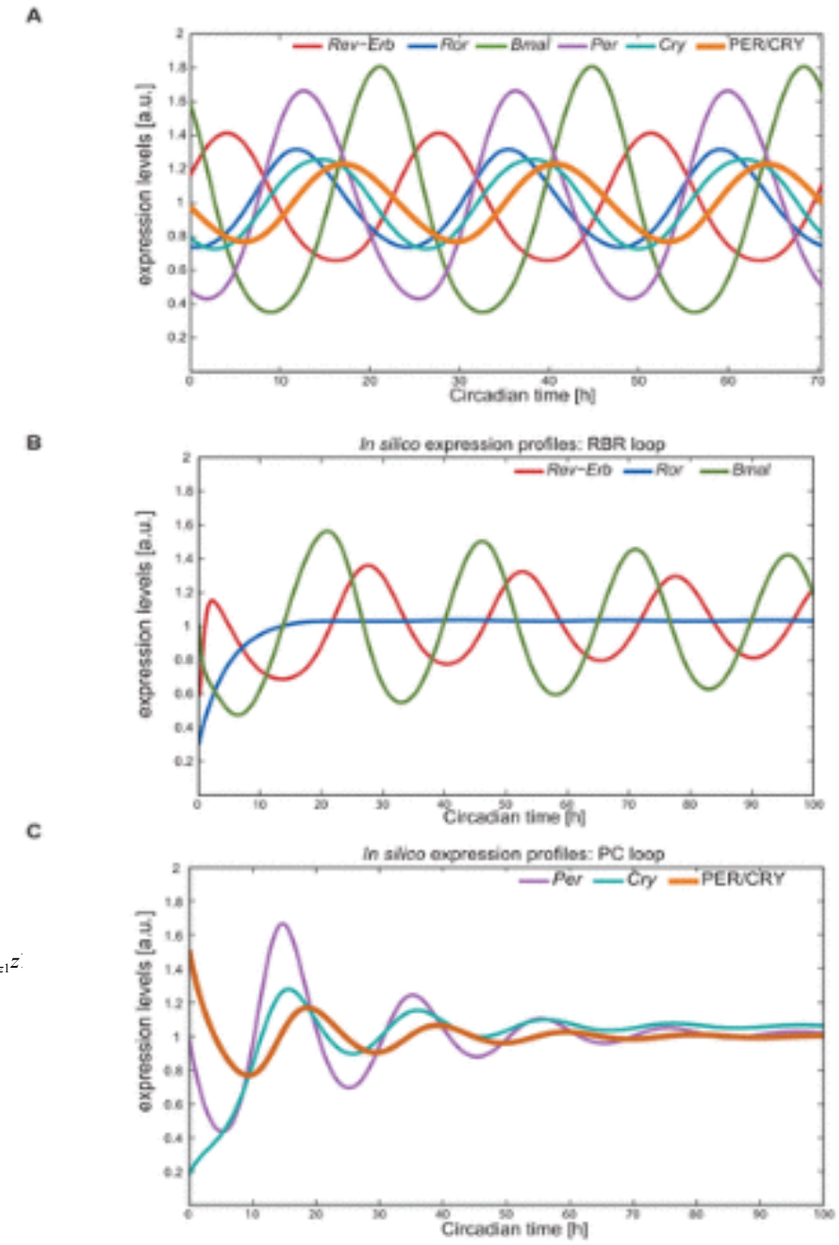
$$\frac{dy1}{dt} = V_{1\max}\frac{1+a\left(\frac{x1}{k_{t1}}\right)^{b}}{1+\left(\frac{PC}{k_{i1}}\right)^{c}\left(\frac{x1}{k_{t1}}\right)^{b}+\left(\frac{x1}{k_{t1}}\right)^{b}} - d_{y1}y1$$

Cry

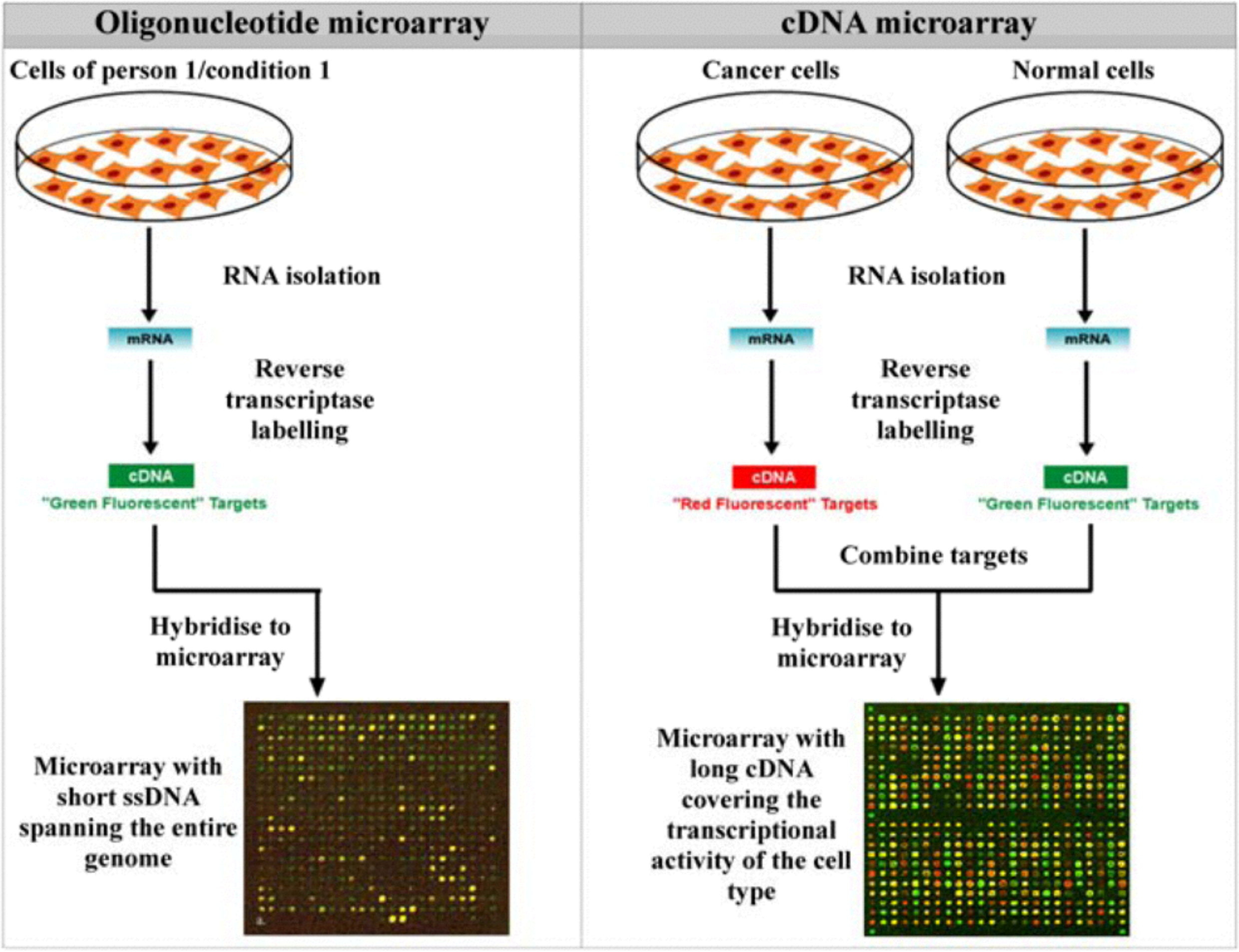
$$\frac{dy2}{dt} = V_{2\max}\frac{1+d\left(\frac{x1}{k_{t2}}\right)^{e}}{1+\left(\frac{PC}{k_{i2}}\right)^{f}\left(\frac{x1}{k_{t2}}\right)^{e}+\left(\frac{x1}{k_{t2}}\right)^{e}}\frac{1}{1+\left(\frac{x5}{k_{i21}}\right)^{f1}} - d_{y2}y2$$

CRY$_C$

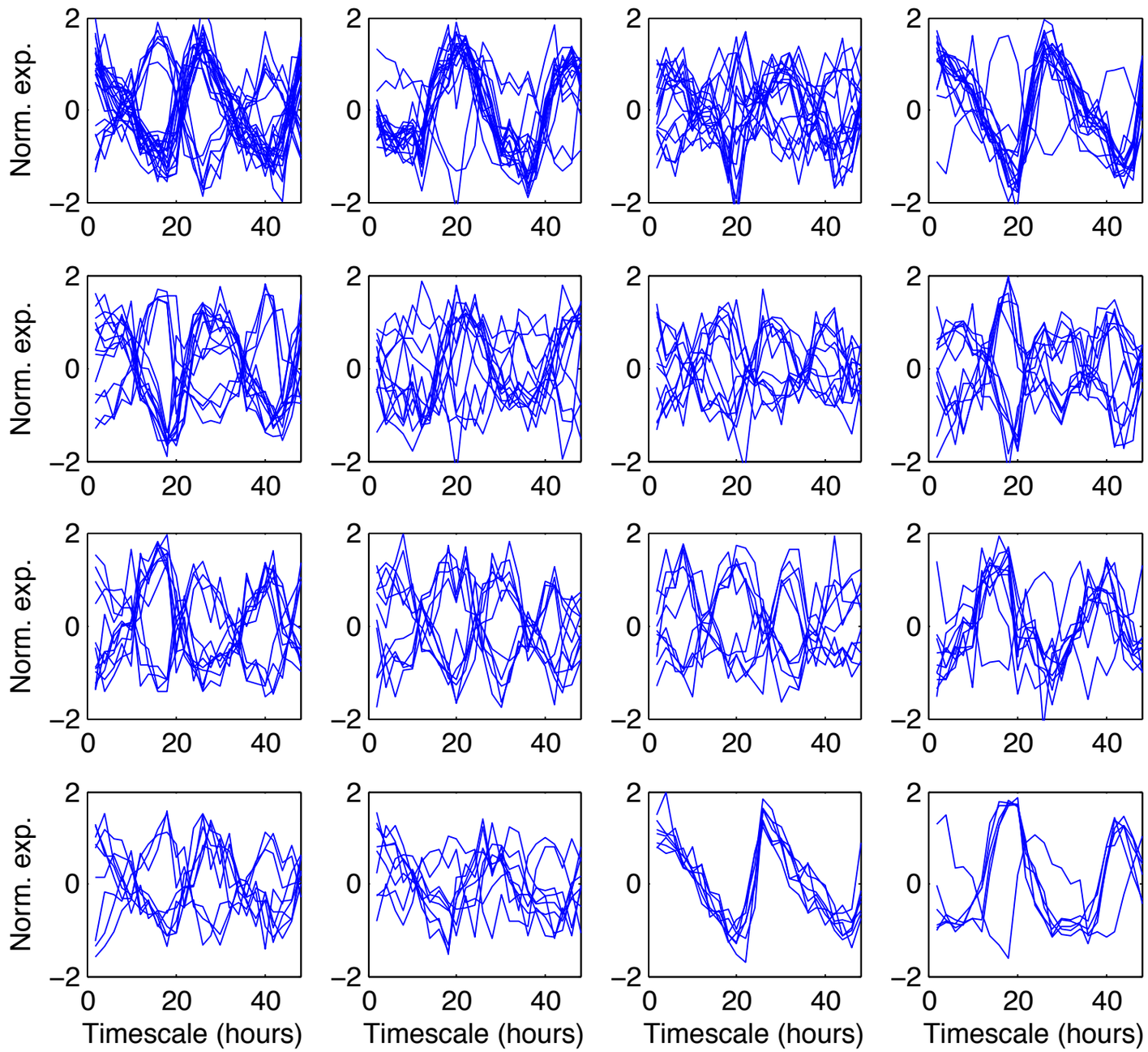
$$\frac{dz1}{dt} = k_{p2}(y2+y2_0) + kd_{z4}z4 + kd_{z5}z5 - kf_{z5}z1z2 - kf_{z4}z1z3 - d_{z1}z$$



Gene Expression Data


Oligonucleotide microarray
Cells of person 1/condition 1
RNA isolation
mRNA
Reverse transcriptase labelling
cDNA
"Green Fluorescent" Targets
Hybridise to microarray
Microarray with short ssDNA spanning the entire genome
cDNA microarray
Cancer cells
Normal cells
RNA isolation
mRNA
mRNA
Reverse transcriptase labelling
cDNA
cDNA
"Red Fluorescent" Targets
"Green Fluorescent" Targets
Combine targets
Hybridise to microarray
Microarray with long cDNA covering the transcriptional activity of the cell type

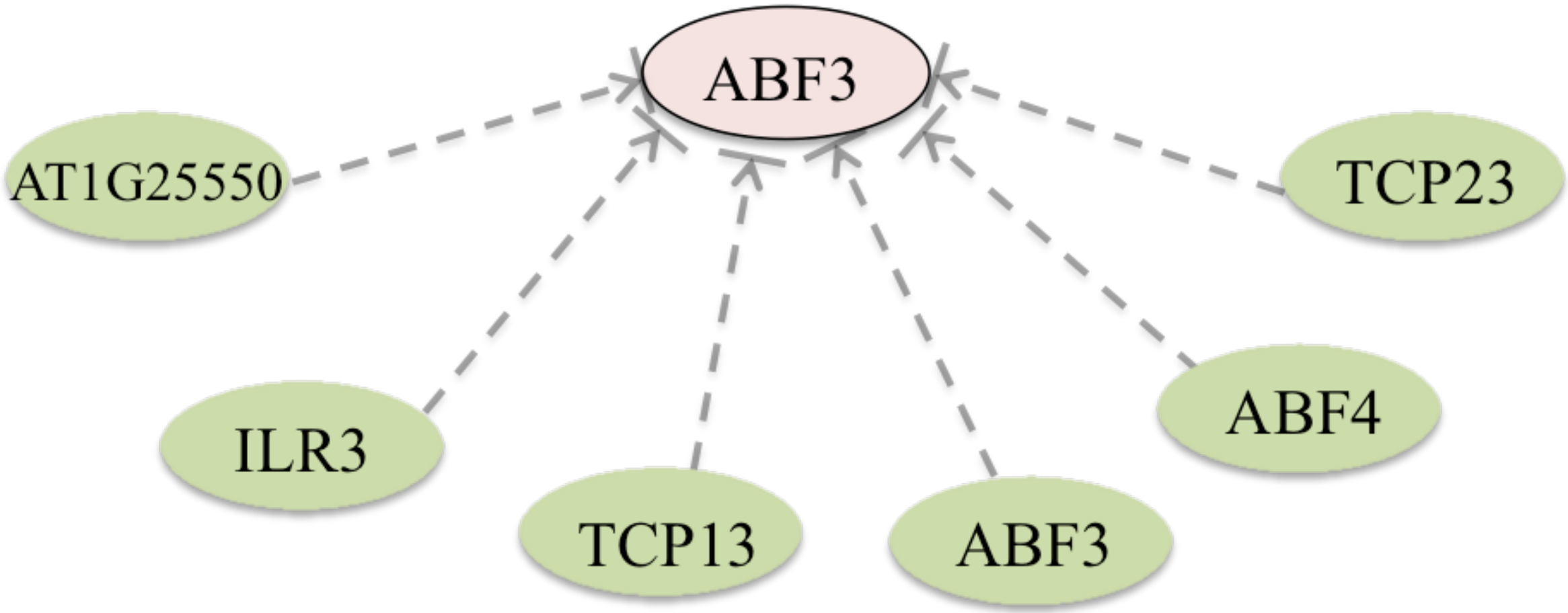
Plant Microarray time series (PRESTA project)

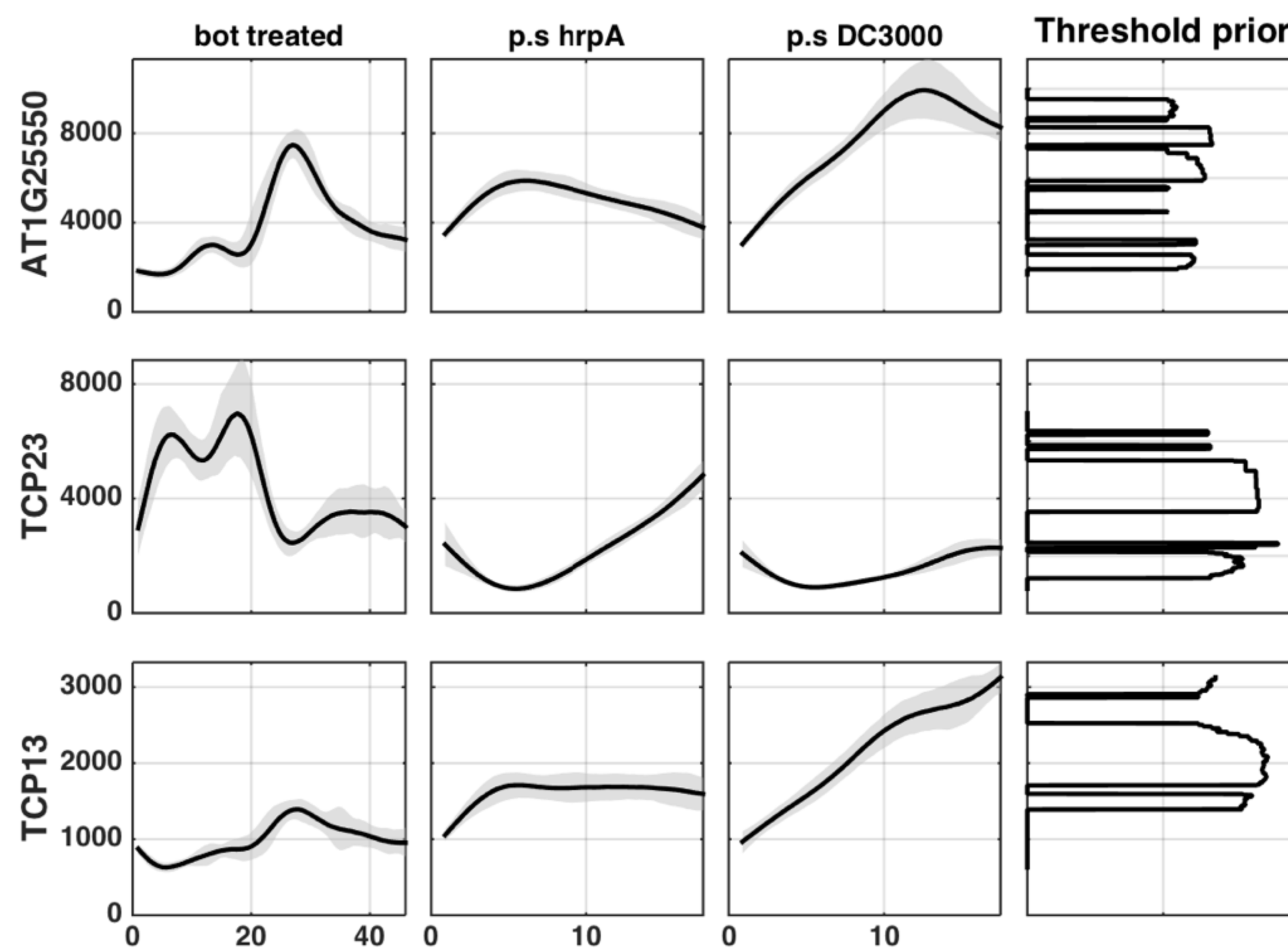


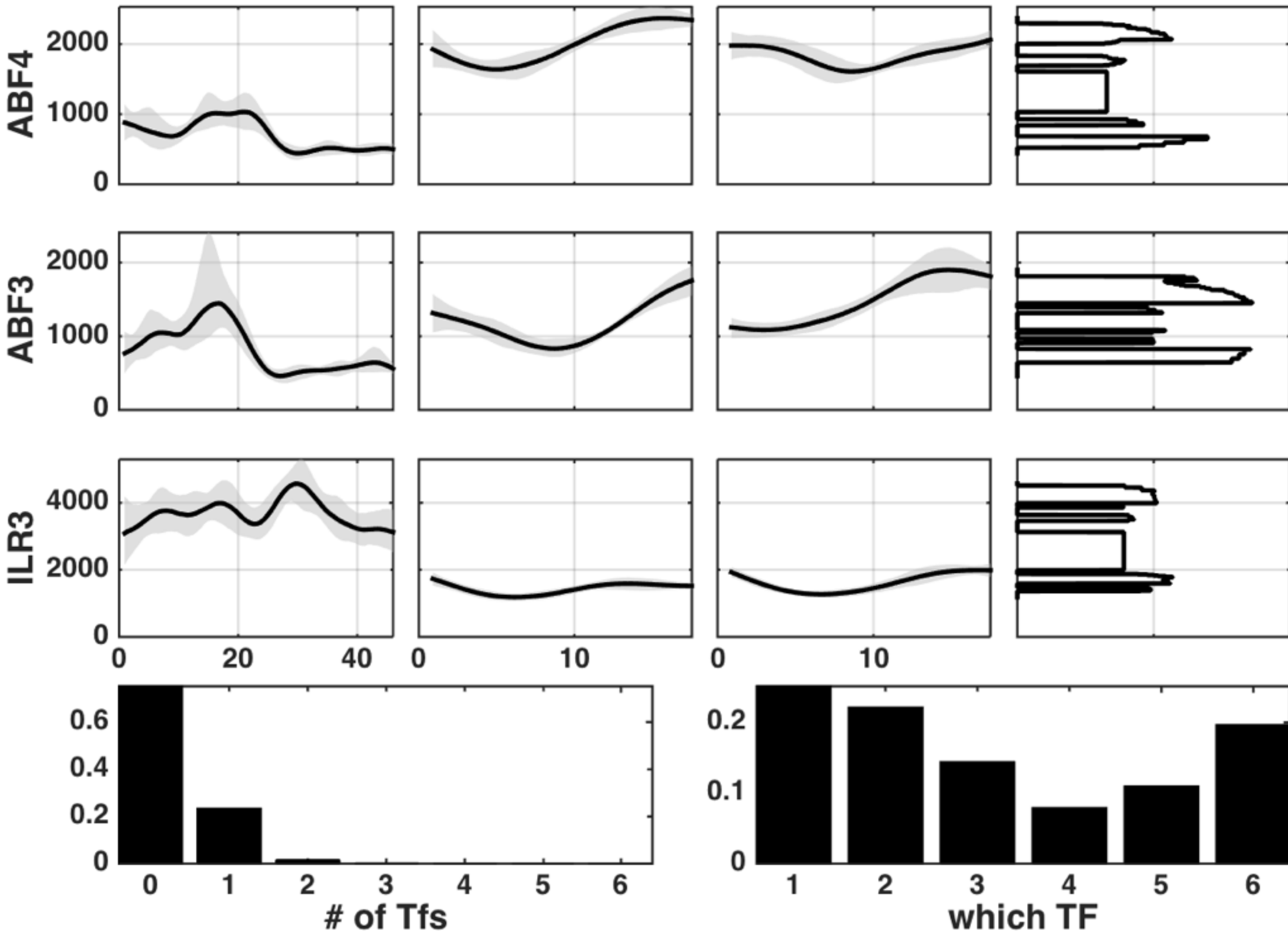
Searching for transcription factors (TFs) of gene regulation from microarray data

Scenario:

Have (replicate) time series microarray gene expression data across various experiments and a set of candidate parents from Y1H experiments


ABF3
AT1G25550
TCP23
ILR3
ABF4
TCP13
ABF3


bot treated
p.s hrpA
p.s DC3000
Threshold prior
AT1G25550
TCP23
TCP13
8000
4000
0
8000
4000
0
3000
2000
1000
0
0
20
40
0
10
0
10


ABF4
2000
1000
0
ABF3
2000
1000
0
ILR3
4000
2000
0
0
20
40
0
10
0
10
0.6
0.4
0.2
0
0
1
2
3
4
5
6
# of Tfs
0.2
0.1
0
1
2
3
4
5
6
which TF

Modeling:

Transcription β(t) of a gene (*Child gene*) is regulated by transcriptional activators and/or inhibitors. These are protein products, called Transcription Factors (TFs), of other genes (Parent genes).

Wish to draw inference about regulation of mRNA transcription of a child gene by an unknown subset

$$\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_\Gamma\}$$

of the candidate parents

$$\mathcal{G} = \{g_1, g_2, \ldots, g_G\}$$

Switches occur at times where the expression, Pγ (t) of a parent $\gamma \in \Gamma$ crosses a threshold level due to an increase or decrease.

Define activation function

$$\alpha_{\Gamma}(t) = (\alpha_{\gamma_1}(t), \alpha_{\gamma_2}(t), \ldots, \alpha_{\gamma_\Gamma}(t))$$

The set of switch times of the child's transcription is the set of time-points $\{s_1, s_2, \ldots, s_k\}$

at which at least one parent crosses its threshold.

The total time interval split into subintervals $[0, s_1], (s_1, s_2], \ldots, (s_k, L]$ where the transcription rates of the subintervals are the same if the expression of all parent genes remains at the same state

For a given activation function $\alpha_{\Gamma}(t)$ we then have a set of parental states, $b_j, j = 1, 2, \ldots, \kappa$ observed in the union I_{b_j} of (at least one of) the above subintervals with the corresponding transcription rates $\tau_{b_j}, j = 1, 2, \ldots, \kappa$

$$\frac{dM}{dt} = \begin{cases} \tau_{b_1} - \delta_M M(t), & \text{for } t \in I_{b_1} \\ \tau_{b_2} - \delta_M M(t), & \text{for } t \in I_{b_2} \\ \vdots & \vdots \\ \tau_{b_\kappa} - \delta_M M(t), & \text{for } t \in I_{b_\kappa}. \end{cases}$$

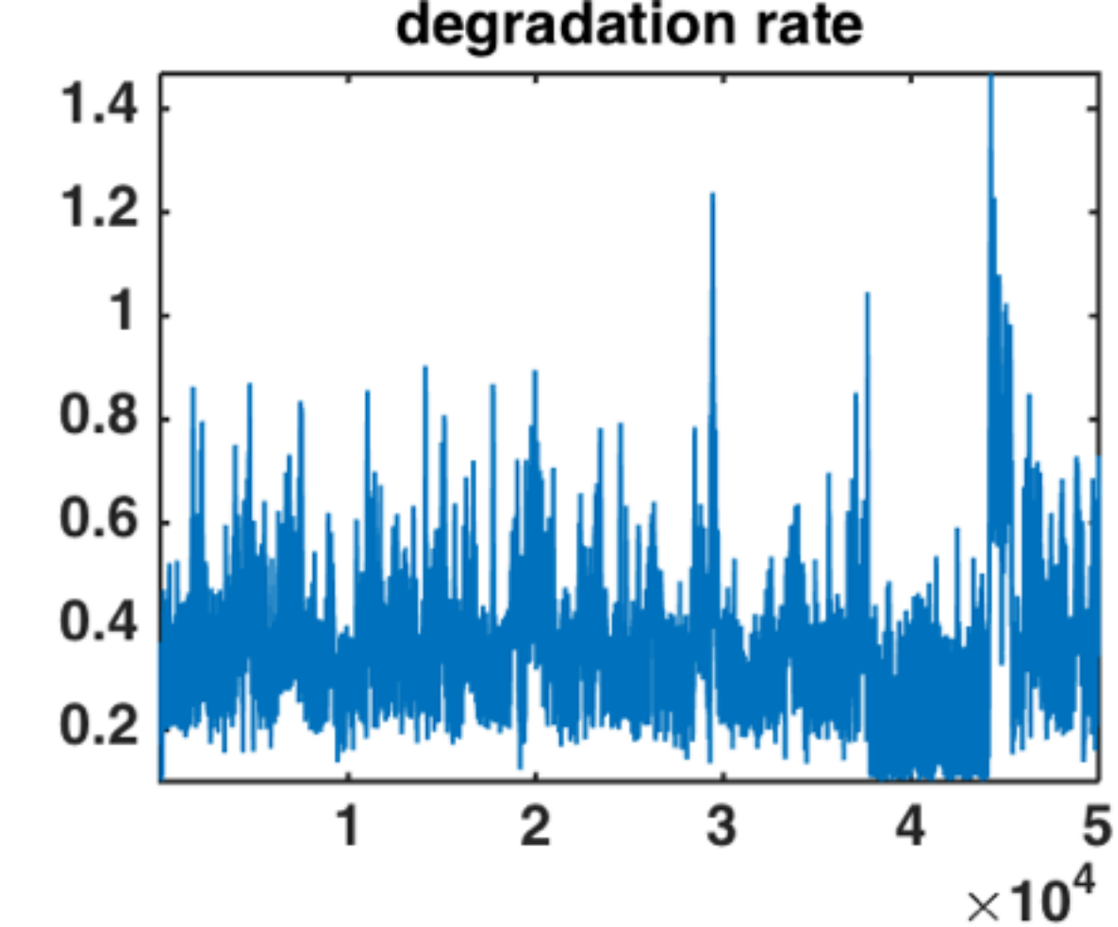
Statistical Implementation:

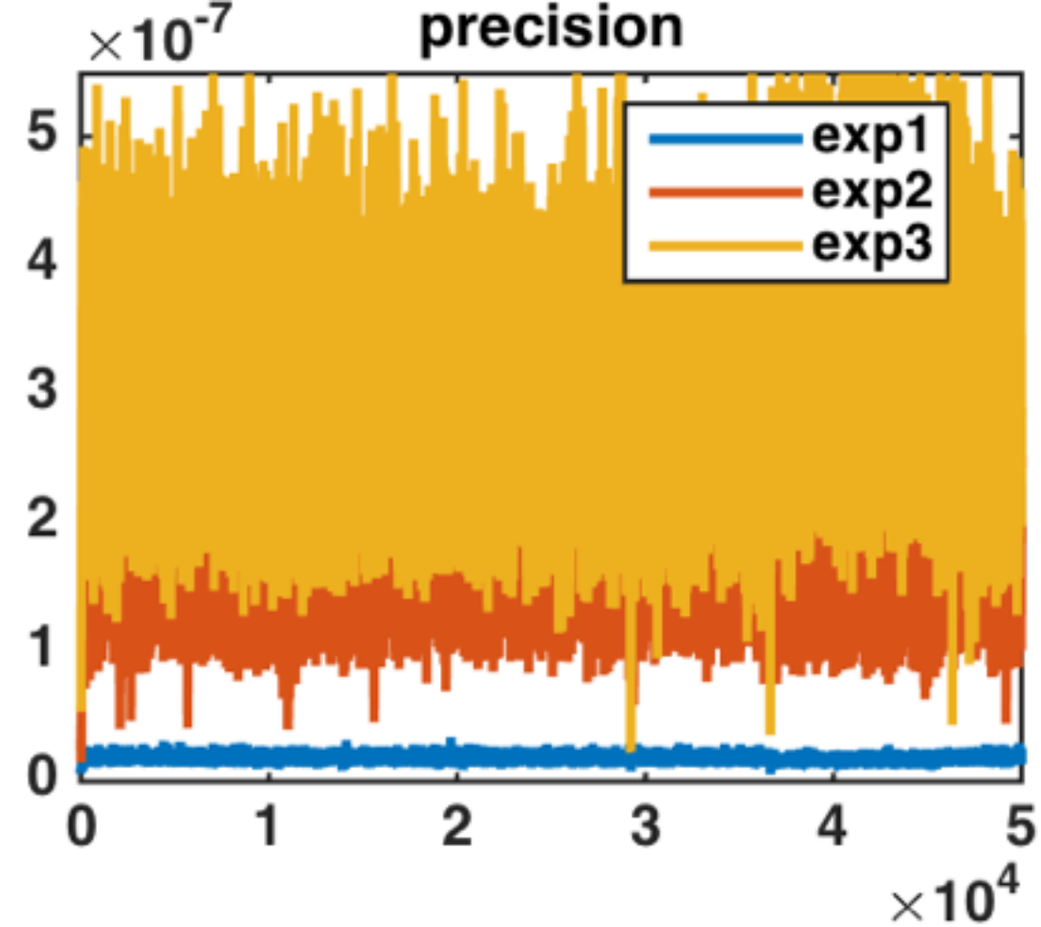
Bayesian approach: RJMCMC to infer plausible models for parents

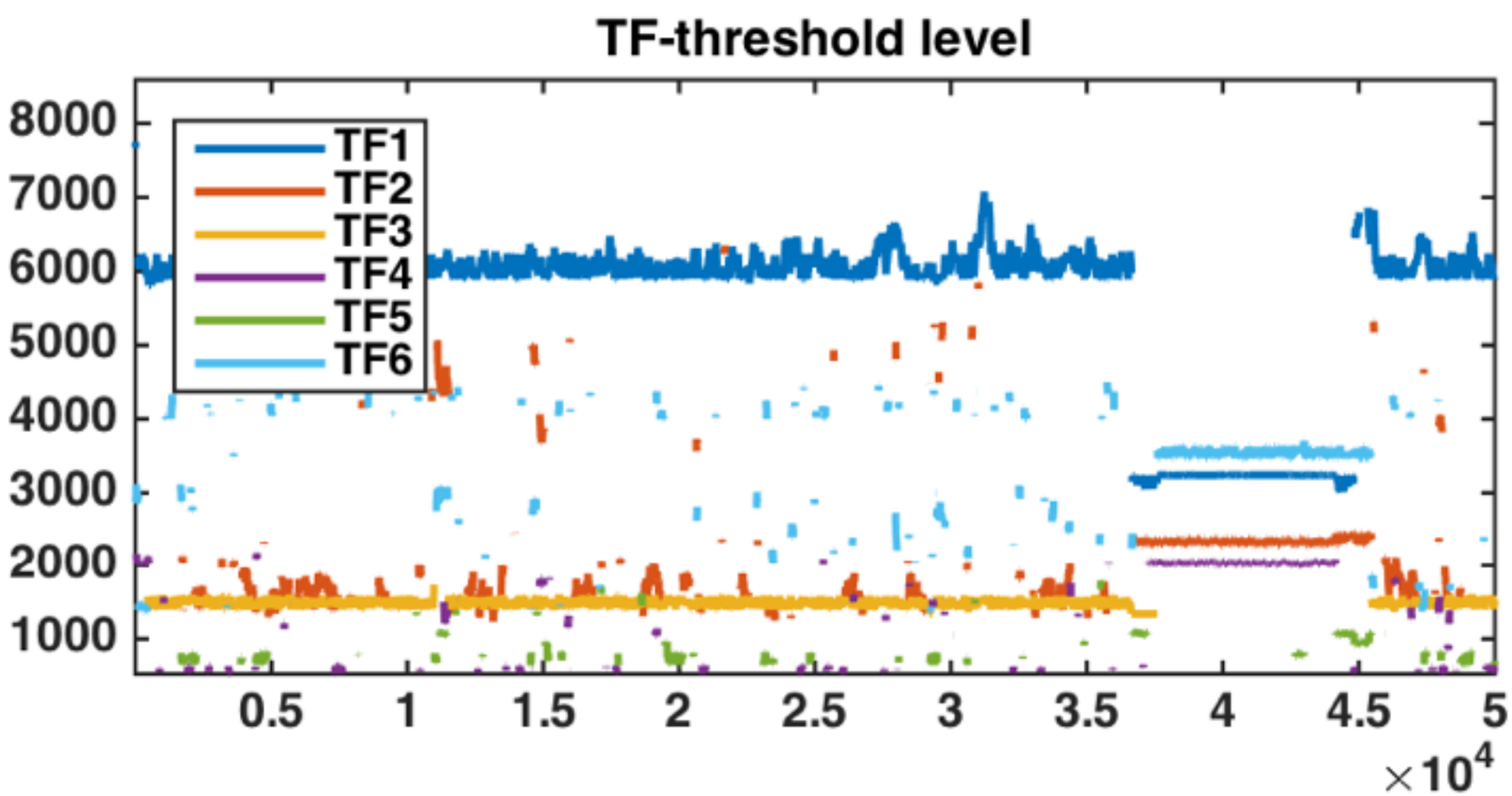
Likelihood based on normal error with inhomogeneous variance

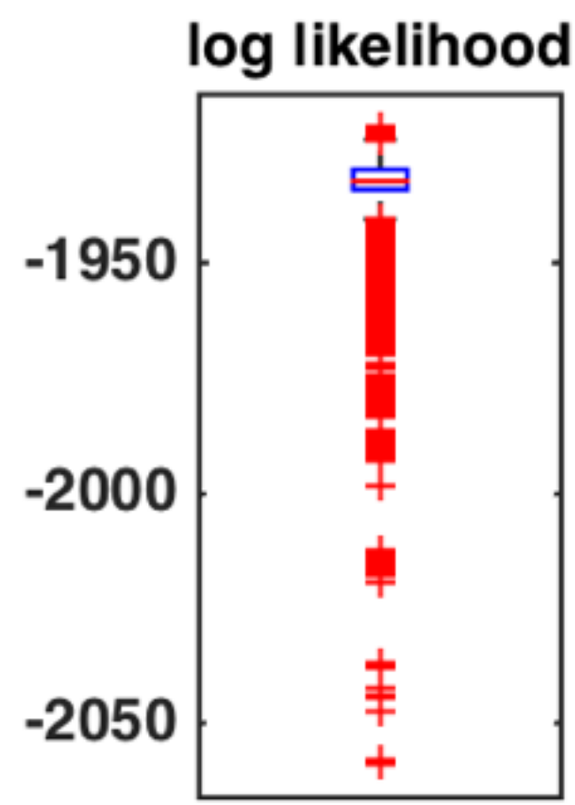
Weighted Least Squares solution to piecewise linear ODE makes algorithm fast

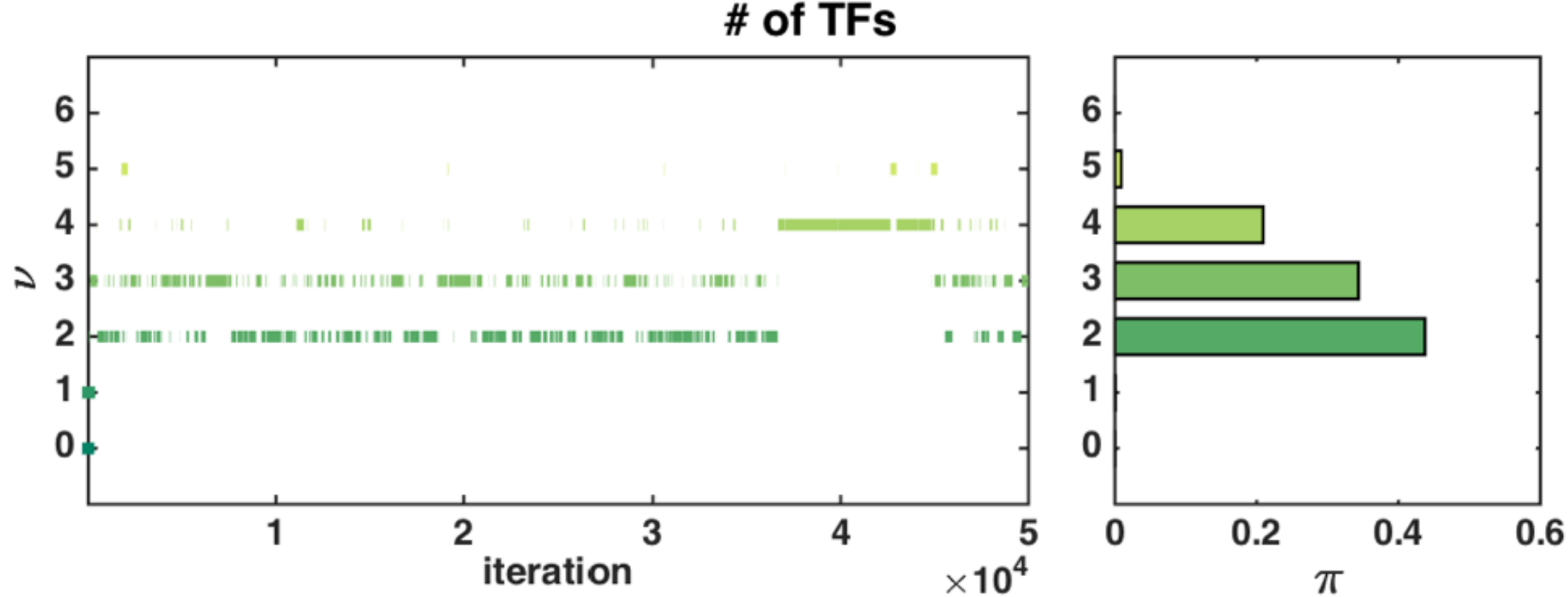
Delayed RNA as proxy for unobserved TF's

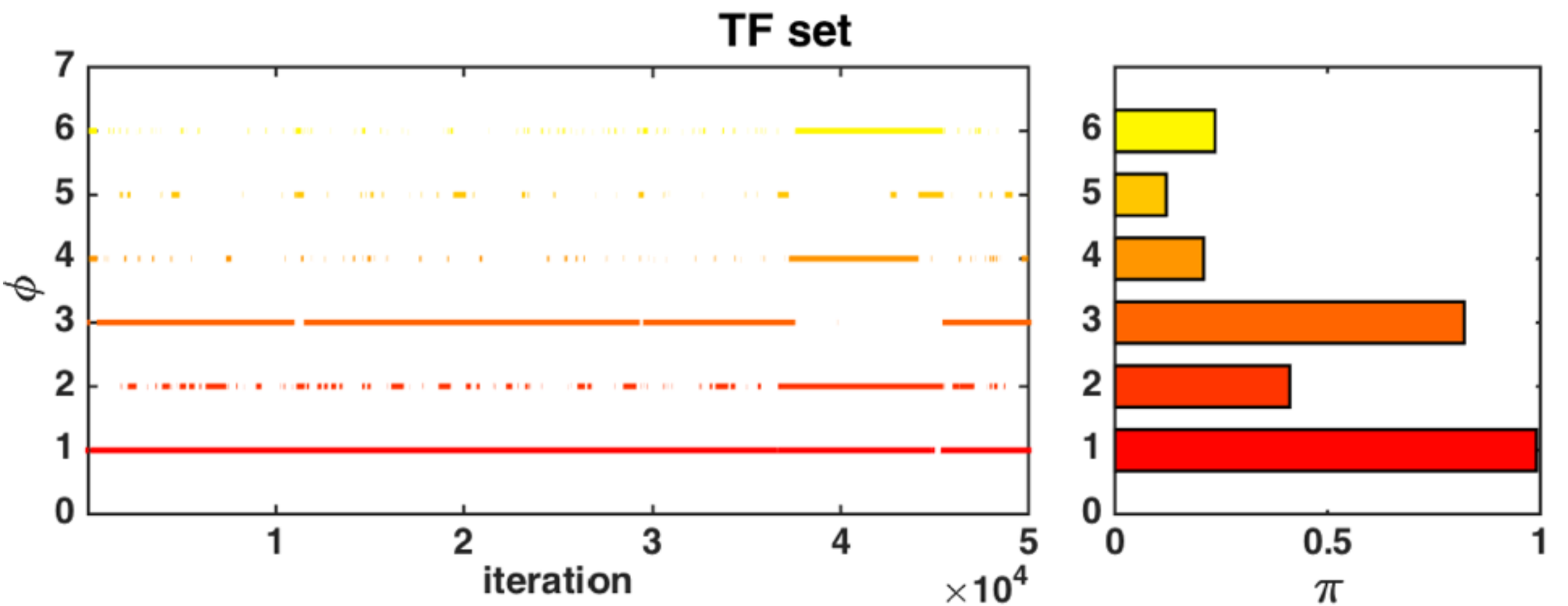

degradation rate
1.4
1.2
1
0.8
0.6
0.4
0.2
1
2
3
4
5
×10^4

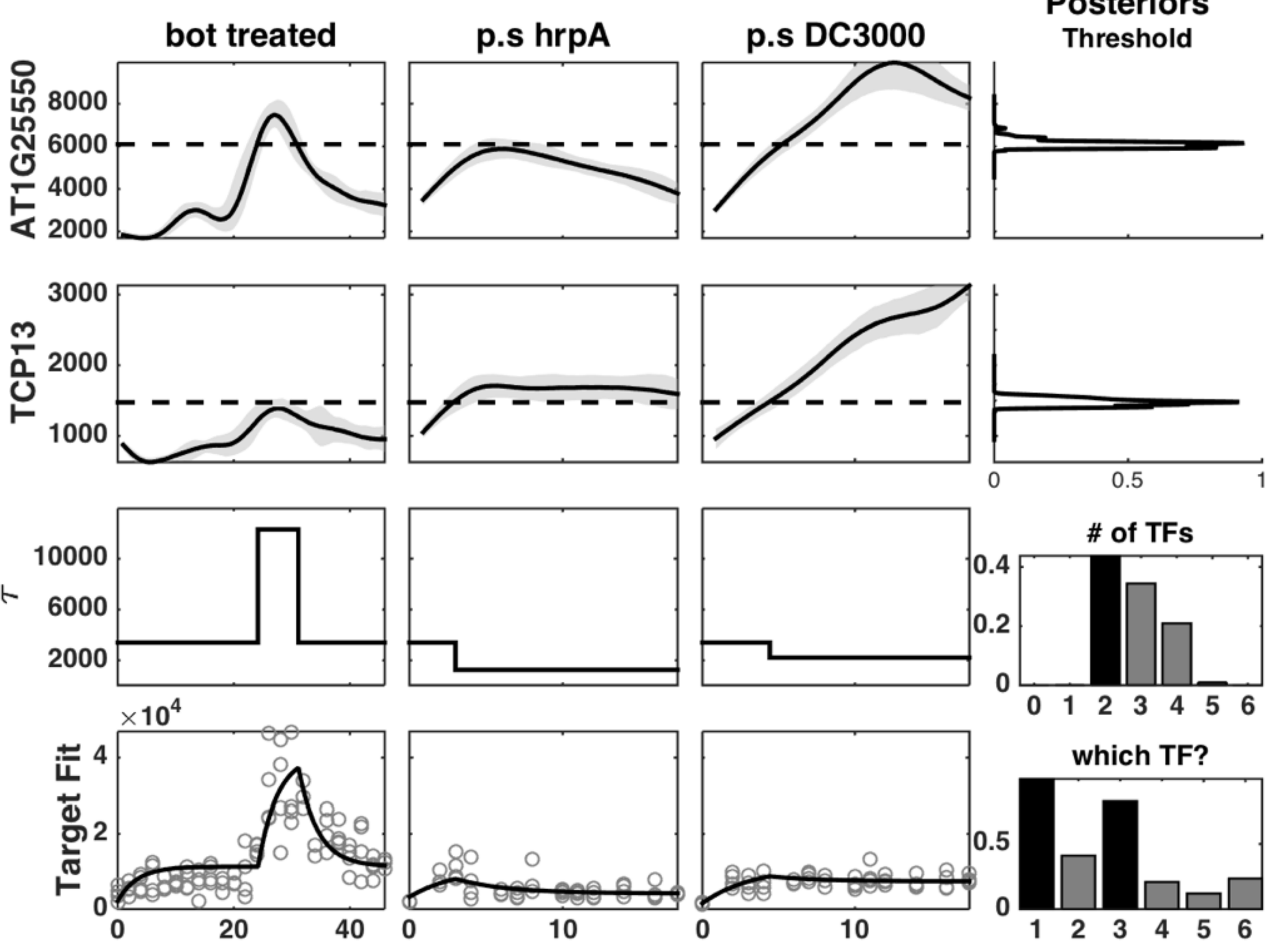

precision
×10^-7
5
4
3
2
1
0
exp1
exp2
exp3
0
1
2
3
4
5
×10^4


TF-threshold level
8000
7000
6000
5000
4000
3000
2000
1000
TF1
TF2
TF3
TF4
TF5
TF6
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
×10^4

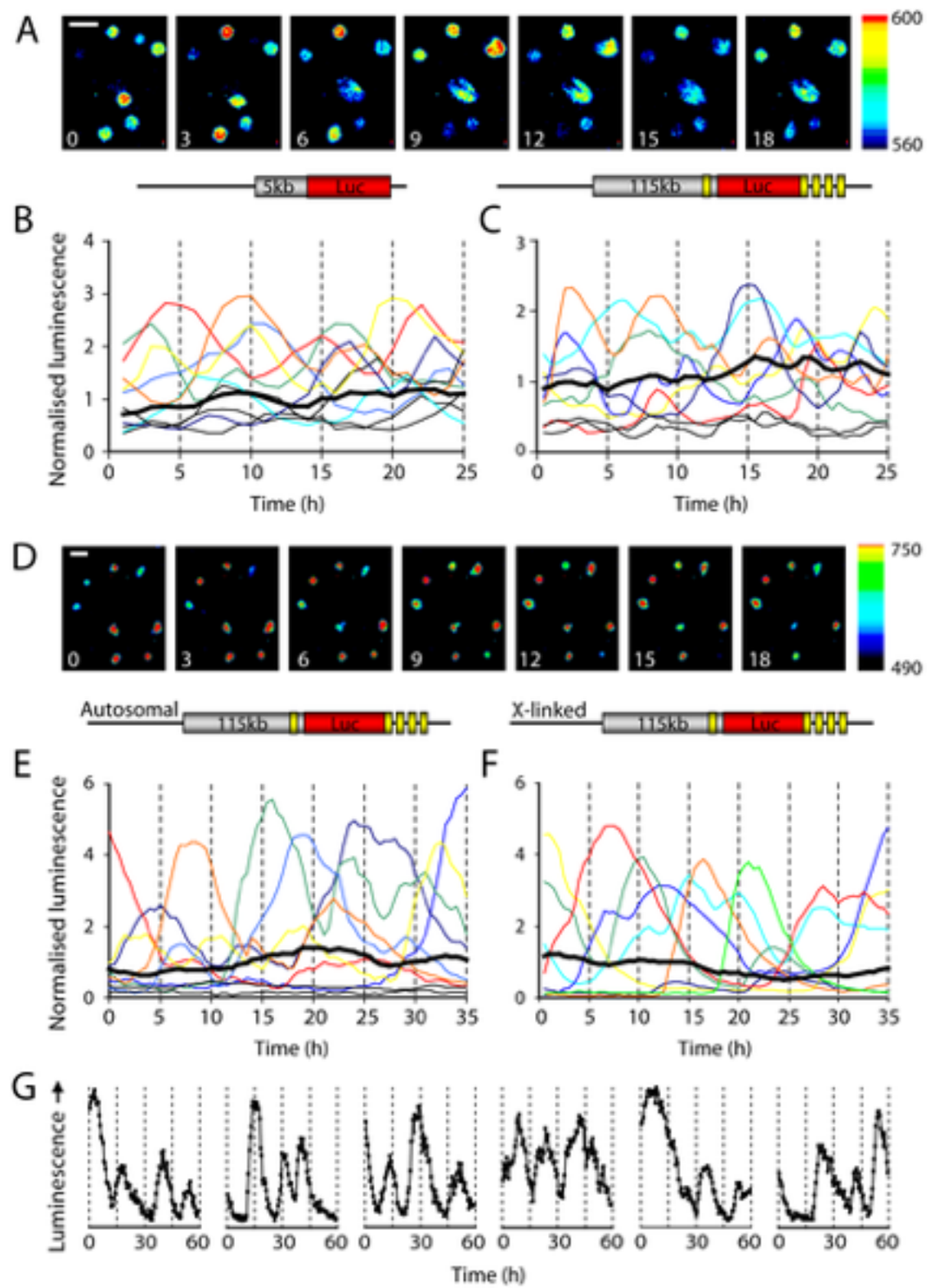

log likelihood
-1950
-2000
-2050


# of TFs
ν
iteration
×10^4
π

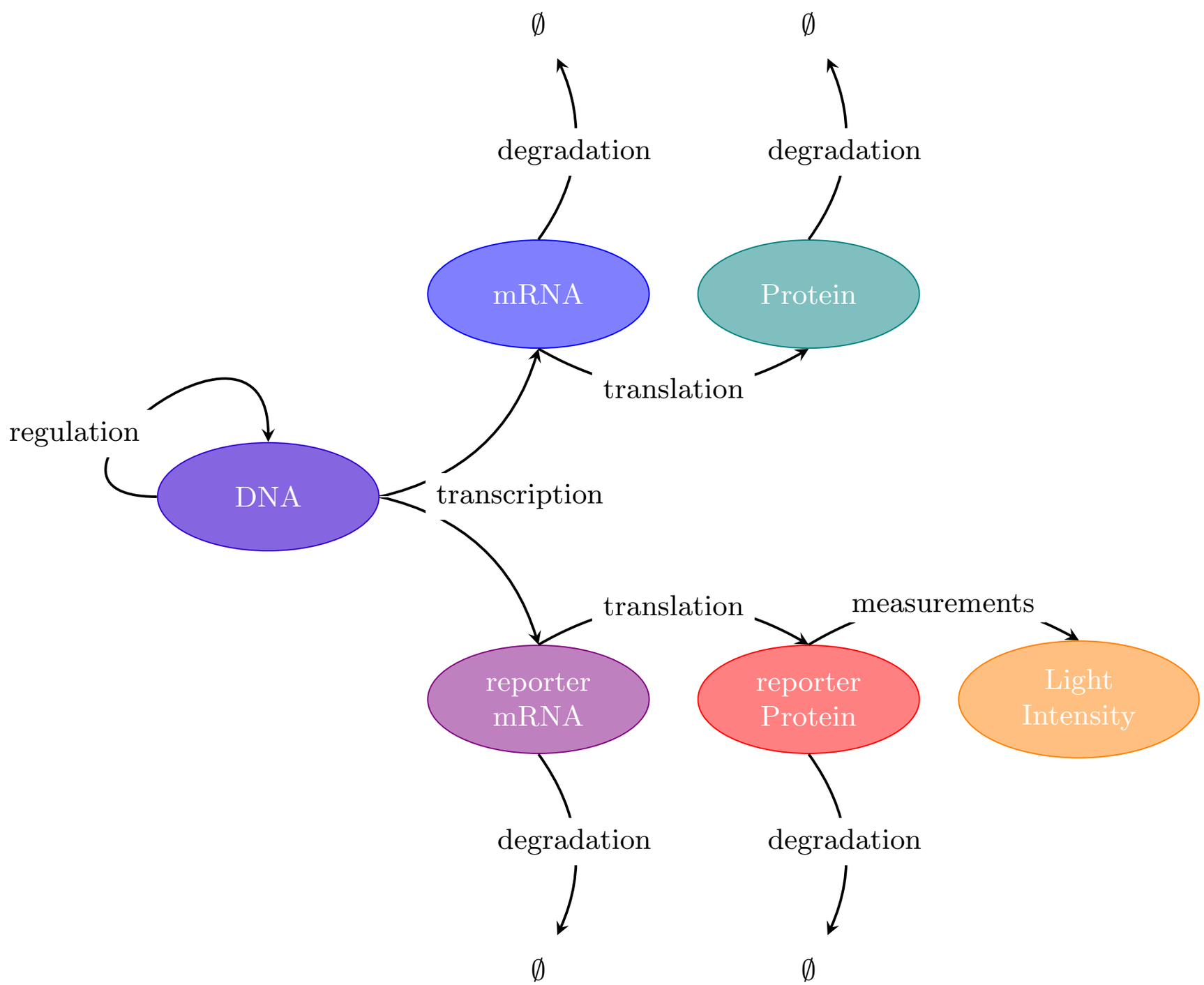

TF set
ϕ
iteration
×10^4
π


bot treated
p.s hrpA
p.s DC3000
Posteriors
Threshold
AT1G25550
8000
6000
4000
2000
TCP13
3000
2000
1000
0
0.5
1
τ
10000
6000
2000
# of TFs
0.4
0.2
0
0 1 2 3 4 5 6
Target Fit
×10^4
4
2
0
0 20 40
0 10
0 10
which TF?
0.5
0
1 2 3 4 5 6

Single cell imaging data



Reporter Gene constructs

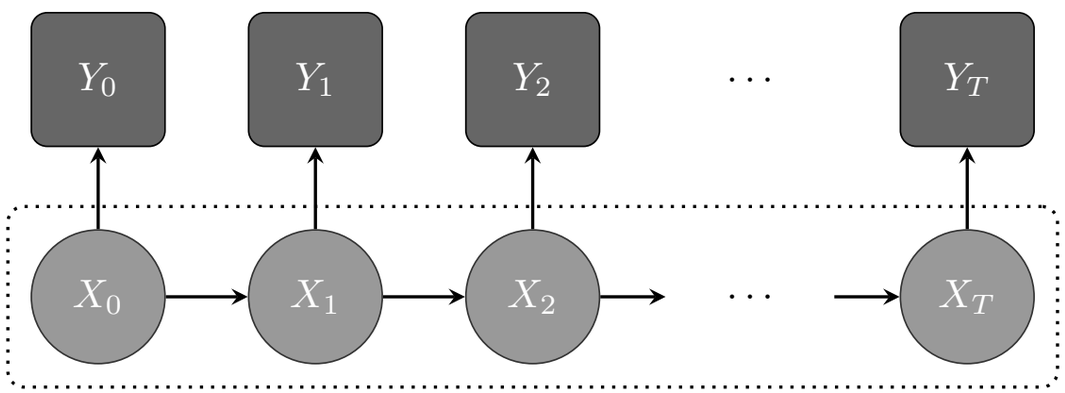


Modelling single cell data

The class of state space models provides a unifying framework for modelling SRNs

$$Y_t \sim g(y_t | x_t, \theta)$$

$$X_{t+1} \sim h(x_{t+1} | \theta)$$



h is the transition density of the approximating SRN

g is the density of the measurement process

Approximations

- ODE (neglects intrinsic noise)
- Chemical Langevin Equation (usually intractable)
- Linear Noise Approximation (linearization of the ME, tractable)
- Other approximations ?

Parameter Inference

The data likelihood is given by the marginal density

$$f(y|\theta) = \int_x f(y, x|\theta) dx$$

where the integrand can be factorized as

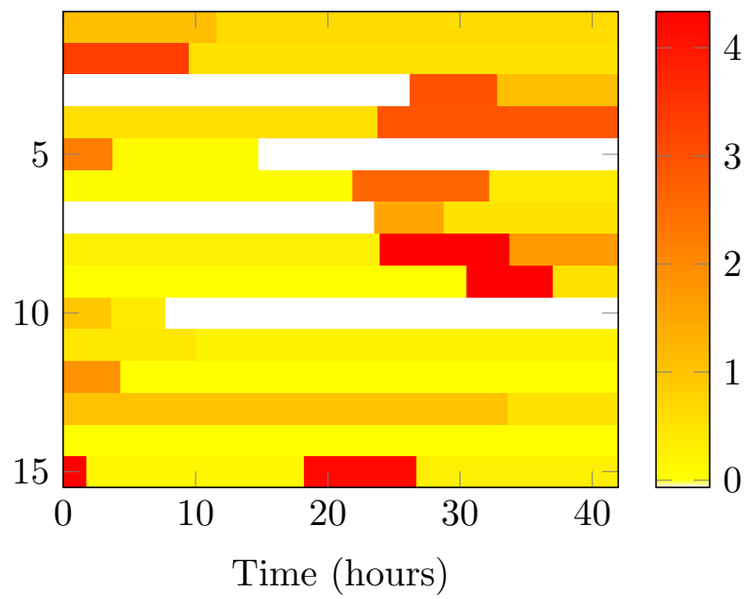
$$f(y, x|\theta) = h(x_0|\theta) g(y_0|x_0, \theta) \prod_{t=1}^{T} h(x_t|x_{t-1}, \theta) g(y_t|x_t, \theta)$$

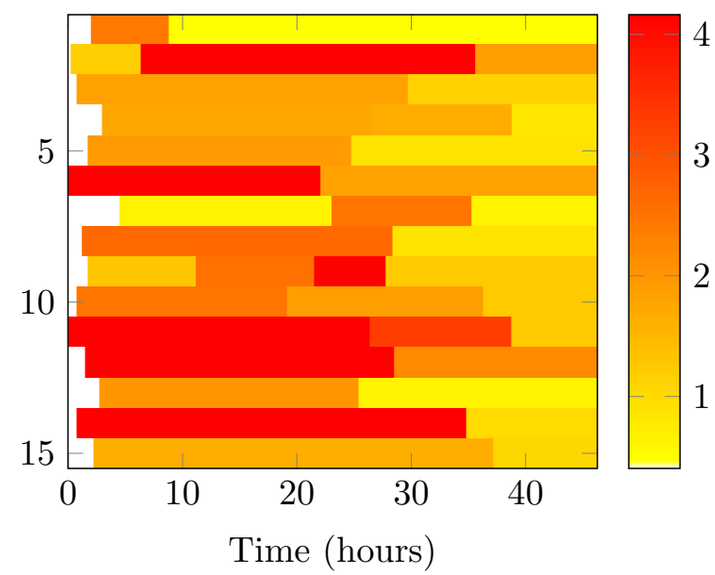
Parameter Inference

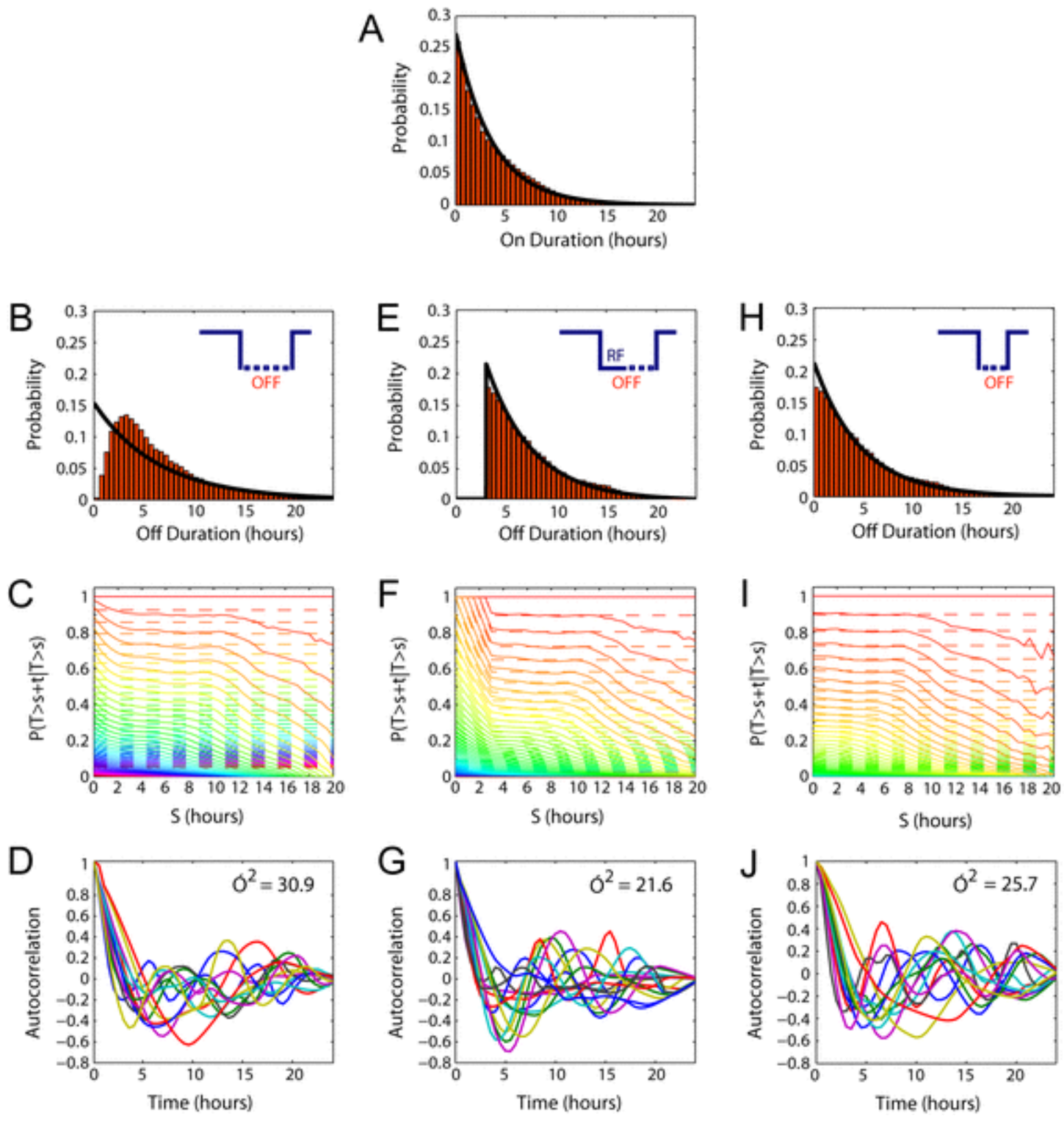
Under the LNA (with Gaussian measurement error):
Interval can be evaluated explicitly using Kalman methodology.
Inference can be achieved by sampling from the posterior $f(\theta|y)$

Under BDA (for example) : use 2-step Gibbs sampler:

1. Sample the parameter vector from $f(\theta|x, y)$
2. Sample the latent states x from the filtering density $f(x|y, \theta)$


5
10
15
0
10
20
30
40
4
3
2
1
0
Time (hours)


5
10
15
0
10
20
30
40
4
3
2
1
Time (hours)


A
Probability
On Duration (hours)
B
E
H
OFF
RF
Off Duration (hours)
C
F
I
P(T>s+t|T>s)
S (hours)
D
G
J
Autocorrelation
Time (hours)
$\sigma^2 = 30.9$
$\sigma^2 = 21.6$
$\sigma^2 = 25.7$

Challenges for statisticians in Systems Biology and Systems Medicine

- Complex and large networks of genes and their products

- Data from many sources:
  - Replicates and if so what type?
  - Many cells (Bayesian Hierarchical Modeling)
  - Various labs (Prior Distributions)
  - Destructive Sampling
  - ….

- Different Types of Experiments imply different modeling approaches
  - Information about intrinsic noise?
  - Extrinsic noise?
  - Destructive sampling?
  - Reporter gene constructs? Which reporter gene?
  - Type of camera used for imaging

- Not all state variables are observed

Can use distributed delay functions as a proxy ?

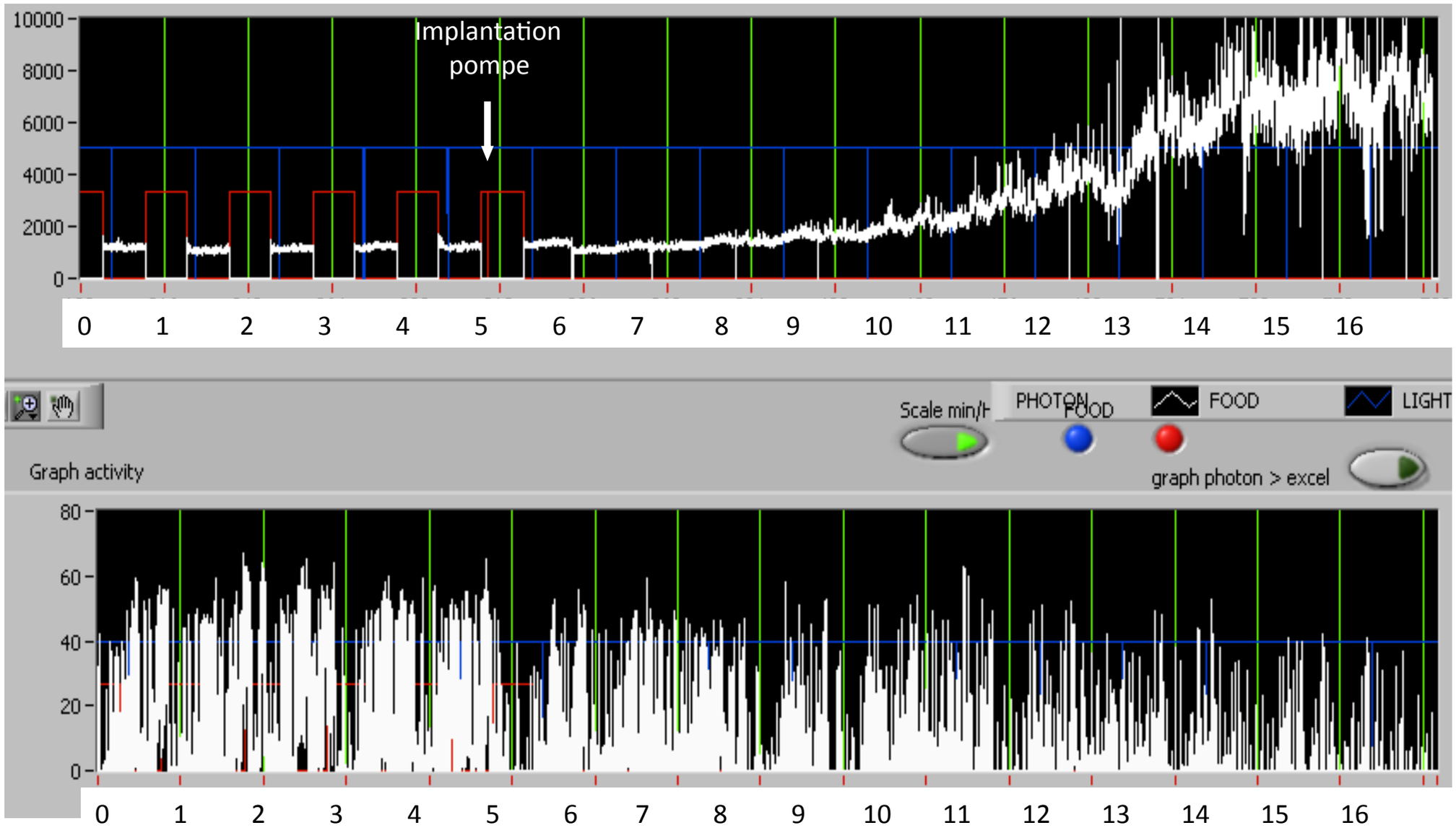
- Parameter Identifiability
- Spatio-temporal modelling and inference
- Collaboration between experimentalists and mathematician/statisticians
-
-

Research Questions

- Mammalian Circadian Oscillator under diseases and treatments. Exploitation for Chronotherapy
- Development of signal processing tools for circadian time series to be used in forecasting systems for patient care at home
- Development of Spatio-temporal models to investigate how cells synchronize in the SCN

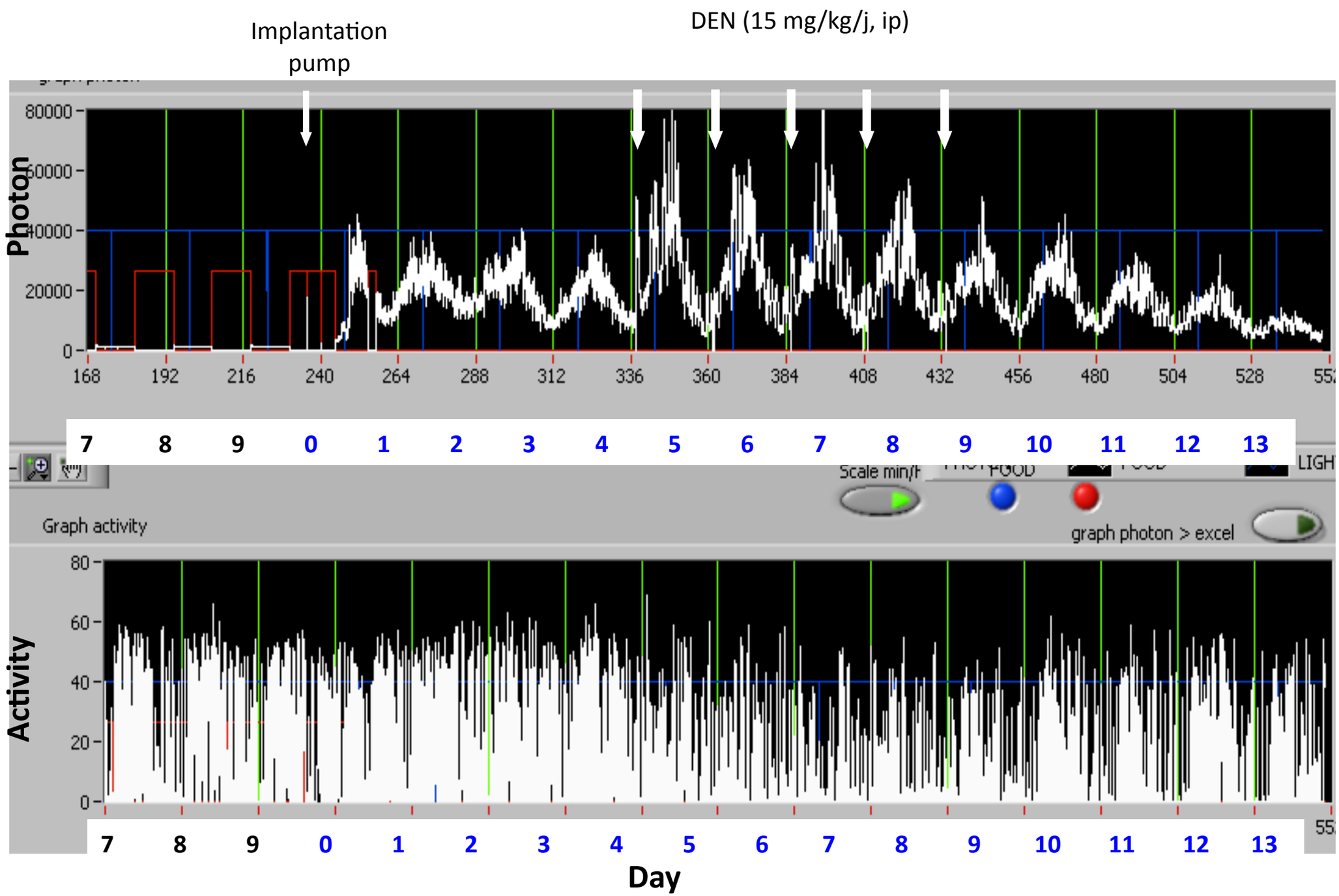
#B2 (-shBmal1)

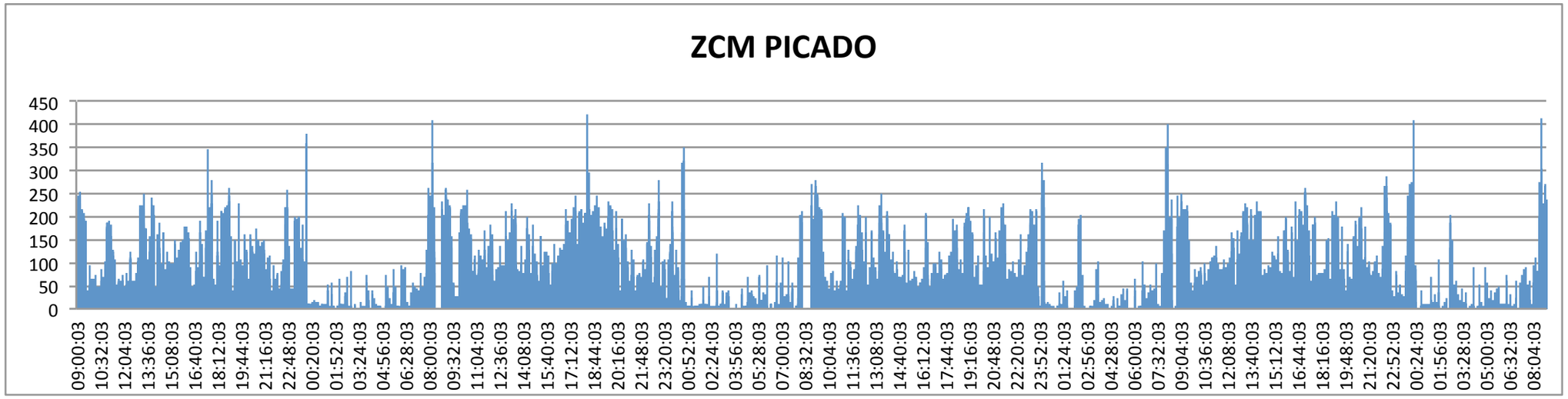
Photon

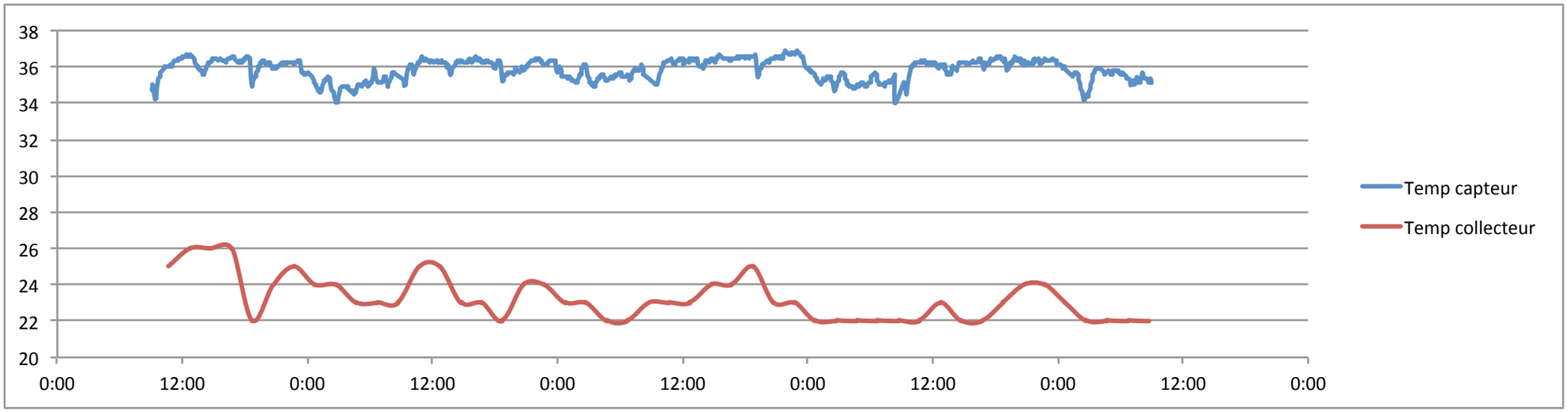


Jour après inoculation tumorale

CT490 : KI/KI Per2::luc male mouse (B1)




ZCM PICADO
450
400
350
300
250
200
150
100
50
0
09:00:03
10:32:03
12:04:03
13:36:03
15:08:03
16:40:03
18:12:03
19:44:03
21:16:03
22:48:03
00:20:03
01:52:03
03:24:03
04:56:03
06:28:03
08:00:03
09:32:03
11:04:03
12:36:03
14:08:03
15:40:03
17:12:03
18:44:03
20:16:03
21:48:03
23:20:03
00:52:03
02:24:03
03:56:03
05:28:03
07:00:03
08:32:03
10:04:03
11:36:03
13:08:03
14:40:03
16:12:03
17:44:03
19:16:03
20:48:03
22:20:03
23:52:03
01:24:03
02:56:03
04:28:03
06:00:03
07:32:03
09:04:03
10:36:03
12:08:03
13:40:03
15:12:03
16:44:03
18:16:03
19:48:03
21:20:03
22:52:03
00:24:03
01:56:03
03:28:03
05:00:03
06:32:03
08:04:03


38
36
34
32
30
28
26
24
22
20
0:00
12:00
0:00
12:00
0:00
12:00
0:00
12:00
0:00
12:00
0:00
Temp capteur
Temp collecteur

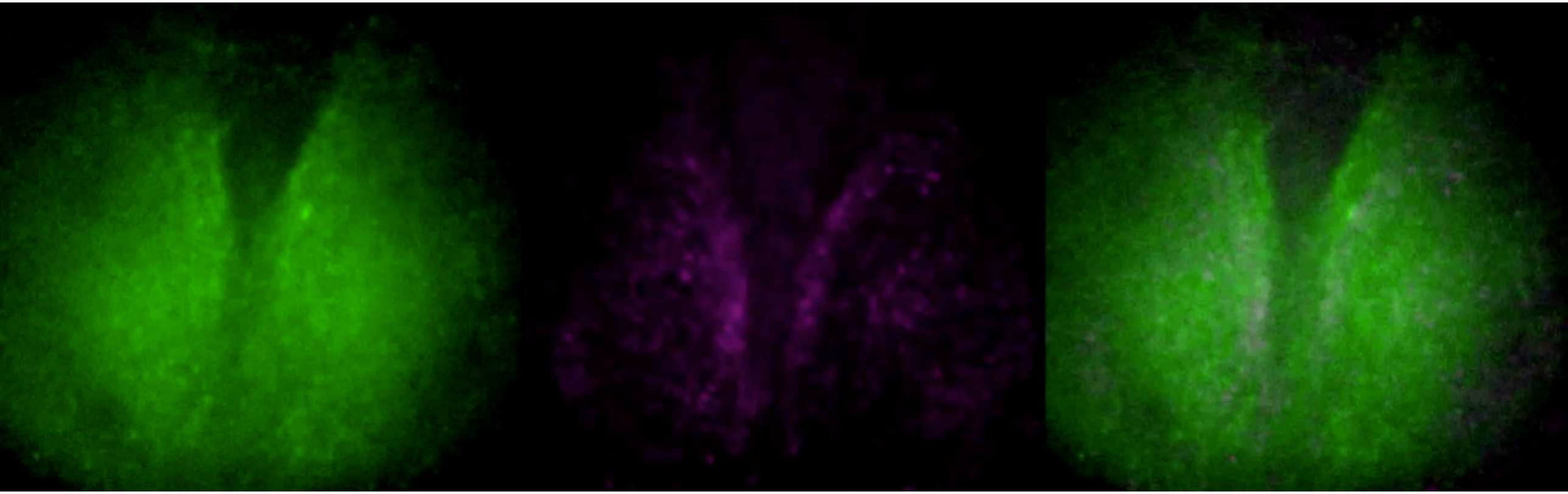
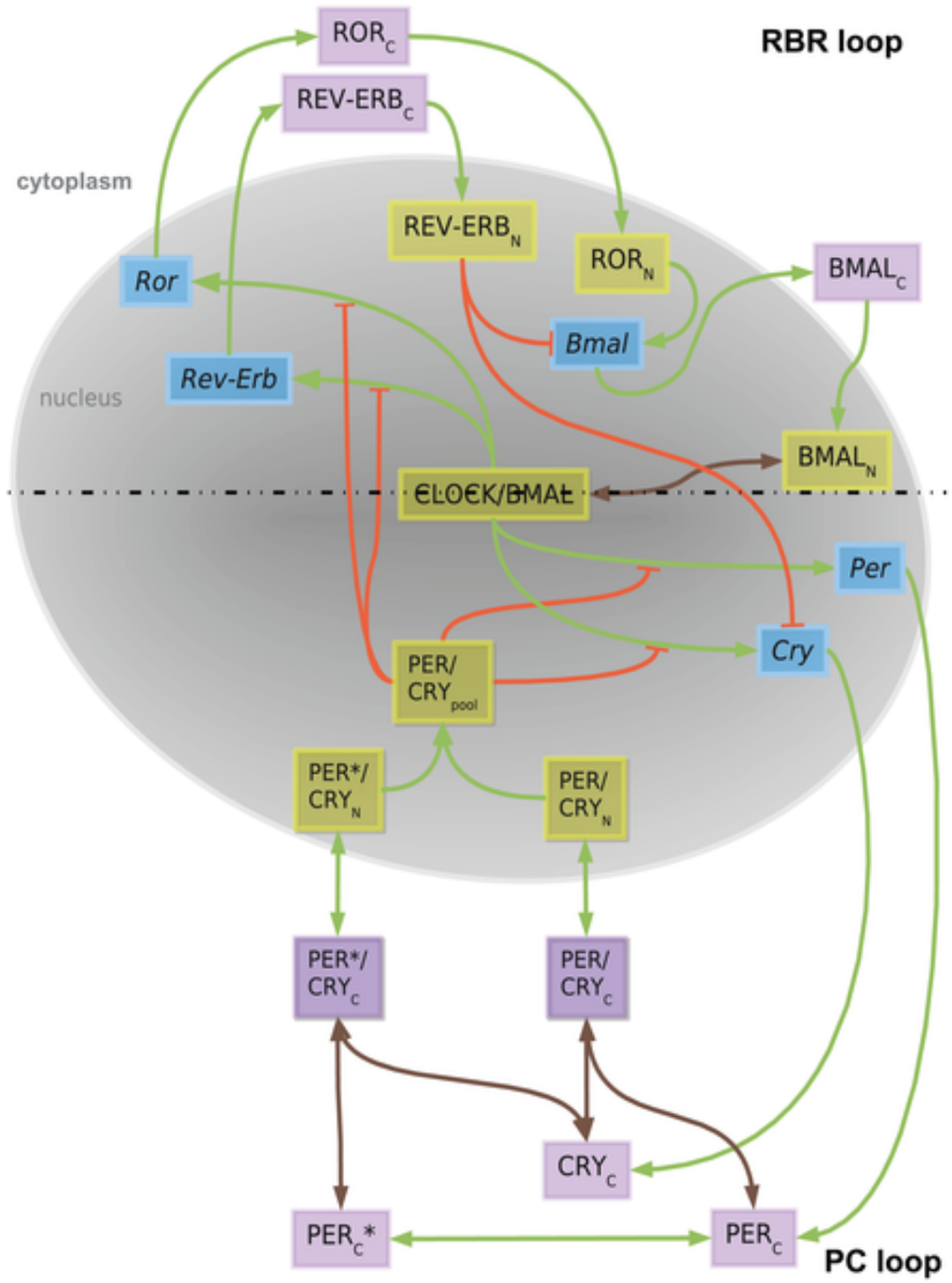


Figure 1. A model for the mammalian circadian clock.



Relógio A, Westermark PO, Wallach T, Schellenberg K, Kramer A, et al. (2011) Tuning the Mammalian Circadian Clock: Robust Synergy of Two Loops. PLoS Comput Biol 7(12): e1002309. doi:10.1371/journal.pcbi.1002309
http://127.0.0.1:8081/ploscompbiol/article?id=info:doi/10.1371/journal.pcbi.1002309