

SOFTWARE

Open Access



# ReTrOS: a MATLAB toolbox for reconstructing transcriptional activity from gene and protein expression data

Giorgos Minas<sup>1,2\*†</sup> , Hiroshi Momiji<sup>1†</sup>, Dafyd J. Jenkins<sup>1,4†</sup>, Maria J. Costa<sup>1,2†</sup>, David A. Rand<sup>1,3</sup> and Bärbel Finkenstädt<sup>2\*</sup>

## Abstract

**Background:** Given the development of high-throughput experimental techniques, an increasing number of whole genome transcription profiling time series data sets, with good temporal resolution, are becoming available to researchers. The ReTrOS toolbox (Reconstructing Transcription Open Software) provides MATLAB-based implementations of two related methods, namely ReTrOS–Smooth and ReTrOS–Switch, for reconstructing the temporal transcriptional activity profile of a gene from given mRNA expression time series or protein reporter time series. The methods are based on fitting a differential equation model incorporating the processes of transcription, translation and degradation.

**Results:** The toolbox provides a framework for model fitting along with statistical analyses of the model with a graphical interface and model visualisation. We highlight several applications of the toolbox, including the reconstruction of the temporal cascade of transcriptional activity inferred from mRNA expression data and protein reporter data in the core circadian clock in *Arabidopsis thaliana*, and how such reconstructed transcription profiles can be used to study the effects of different cell lines and conditions.

**Conclusions:** The ReTrOS toolbox allows users to analyse gene and/or protein expression time series where, with appropriate formulation of prior information about a minimum of kinetic parameters, in particular rates of degradation, users are able to infer timings of changes in transcriptional activity. Data from any organism and obtained from a range of technologies can be used as input due to the flexible and generic nature of the model and implementation. The output from this software provides a useful analysis of time series data and can be incorporated into further modelling approaches or in hypothesis generation.

**Keywords:** Gene transcription, Time series, Transcriptional switches, Circadian timing

## Background

Analyzing the temporal dynamics of mRNA and protein expression is a key ingredient to the study of gene function within the cell. The widespread adoption of high-throughput transcriptomic and proteomic technologies, such as microarrays, fluorescent imaging, transcriptional reporter constructs and sequencing, has enabled the

generation of large numbers of high-resolution genome-scale time series data sets. The processing, analysis and summarising of such time series data has a number of theoretical and computational difficulties to overcome. If, for example, a protein reporter construct is used, what is the relationship between the observed reporter protein and the mRNA expression dynamics of the gene of interest? Moreover, allowing for the process of mRNA and protein degrading, what is the actual transcriptional activity?

Here, we present a software toolbox called ReTrOS (Reconstructing Transcription Open Software) which provides several approaches for processing and analysing both gene or protein expression time series data sets, with

\*Correspondence: Giorgos.Minas@warwick.ac.uk;

B.F.Finkenstadt@warwick.ac.uk

†Equal contributors

<sup>1</sup>Systems Biology Centre, University of Warwick, Coventry CV34 4HD, UK

<sup>2</sup>Department of Statistics, University of Warwick, Coventry CV34 4HD, UK

Full list of author information is available at the end of the article

an easy-to-use graphical interface for user interaction. The software is written in the cross-platform MATLAB® environment. The approach used in ReTrOS is based on a differential equation model [1] to account for transcription and degradation of mRNA molecules and translation and degradation of protein molecules. The model has been the basis to a number of novel methodologies to study the dynamics of gene expression [2–6]. The ReTrOS software currently provides two methods for the processing, reconstruction and analysis of gene transcription dynamics:

- *ReTrOS-Smooth*: based upon the algorithm introduced in Harper et al. (2011) [2]. ReTrOS-Smooth outputs a smooth reconstruction of transcription activity from a non-parametric representation of the transcriptional process. The algorithm is extended to incorporate both variability of measurement error and uncertainty about model parameters with credibility envelopes simulated through a bootstrap procedure.
- *ReTrOS-Switch*: based upon the algorithm introduced in Jenkins et al. (2013) [3]. ReTrOS-Switch outputs a reconstruction of transcriptional activity using a temporal “switch” model where transcription rates are subject to temporal jumps at unknown time points. A Bayesian inference algorithm using reversible jump Markov Chain Monte Carlo is provided for modeling mRNA data as in [3], and has, for the purpose of this article, been extended to model protein dynamics.

This article introduces and describes the underlying model and the two algorithms implemented in ReTrOS. We briefly discuss the data requirements and compare the applicability of the two methods. Following the methods overview, we present several cases of applying ReTrOS concluding with some final remarks summarising the software and its uses.

## Implementation

### mRNA and protein expression dynamics

The basic model underlying ReTrOS is an ordinary differential equation (ODE) model introduced in [1]:

$$\frac{dM}{dt} = \tau(t) - \delta_M M(t) \tag{1}$$

$$\frac{dP}{dt} = \alpha M(t) - \delta_P P(t) \tag{2}$$

where  $M(t)$  is the amount of mRNA transcript at time  $t$ ,  $P(t)$  is the amount of protein at time  $t$ ,  $\tau(t)$  is the rate of transcription/mRNA synthesis,  $\delta_M$  is the rate of mRNA transcript decay,  $\alpha$  is the rate of translation/protein synthesis and  $\delta_P$  is the rate of protein decay.

Let  $y(t_i)$ ,  $i = 1, \dots, n$ , be the discretely observed (not necessarily at equidistant time points) imaging protein time series. Assuming that the mean of the data are proportional with unknown factor  $\kappa$  to the concentration of the reporter protein, we have

$$y(t_i) = \tilde{P}(t_i) + \epsilon(t_i), \text{ where } \tilde{P}(t) = \kappa P(t), \tag{3}$$

and  $\epsilon(t_i)$ ,  $i = 1, \dots, n$  are independent normally distributed random variables each with mean zero and unknown variance  $\sigma^2(t_i)$ . If the observations are mRNA expression levels (e.g. measured using microarrays technology), instead of (Eq. 3) we have

$$y(t_i) = \tilde{M}(t_i) + \epsilon(t_i), \text{ where } \tilde{M}(t) = \kappa M(t), \tag{4}$$

with  $\epsilon(t_i)$ ,  $i = 1, \dots, n$  again independent normally distributed random variables each with mean zero and unknown variance  $\sigma^2(t_i)$ .

The main difference between ReTrOS-Smooth and ReTrOS-Switch lies in the formulation of the transcription function  $\tau(t)$  as follows.

### ReTrOS-smooth: reconstruction of a smooth transcription function

Here we use a kernel smoothing function that we found to be more flexible outperforming the spline approach particularly when the data changes rapidly over short time-scales. The user is assumed to have prior information about the values of the degradation rates and has to provide these through appropriate distributions. The algorithm extends the transcription reconstruction algorithm originally introduced in [2] by incorporating both variability of measurement error and model parameter uncertainty through a bootstrap procedure.

Rewriting (Eq. 2) in terms of  $\tilde{P}(t)$  we first note that, for a given solution path  $\tilde{P}(t)$  and degradation rate  $\delta_P$  we can compute the mRNA path by

$$\tilde{M}(t) := \kappa \alpha M(t) = \frac{d\tilde{P}(t)}{dt} + \delta_P \tilde{P}(t). \tag{5}$$

In ReTrOS-Smooth, the continuous function  $\tilde{P}(t)$  is replaced using kernel smoothing [7] with a bandwidth optimized by leave-one-out cross-validation. One can then recover the transcriptional dynamics by inserting the estimated  $\tilde{M}(t)$  into

$$\tilde{\tau}(t) := \kappa \alpha \tau(t) = \frac{d\tilde{M}(t)}{dt} + \delta_M \tilde{M}(t). \tag{6}$$

If the scaling factors  $\kappa$  and  $\alpha$  are unknown, the profile of the reconstructed time paths representing the mRNA and transcriptional dynamics are computed at an arbitrary level. In ReTrOS we set  $\kappa \alpha = 1$  as default parameter values.

This two-step back-calculation is applied if the observed time courses represent protein data arising, for example,

from reporter protein or fluorescently tagged functional protein. If the data  $y(t_i)$  represent mRNA levels, the procedure involves only the one step of the back-calculation in (Eq. 6) and instead of (Eq. 3),  $y(t_i)$  are assumed to be related to the unknown mRNA expression levels,  $M(t)$ , through (Eq. 4). A prior distribution for the mRNA degradation rate,  $\delta_M$ , is provided by the user here. The next paragraph details the steps of the ReTrOS-Smooth implementation algorithm.

**Variability of the reconstructed profile** The variability of the profiles for  $\tilde{\tau}(t)$  depends on the variance of the fitted kernel function given the sample size and variance of the residuals. ReTrOS-Smooth incorporates uncertainty about knowledge of the degradation rate parameters. Let  $\hat{\epsilon}_{t_i} = \hat{y}(t_i) - y(t_i)$  denote the residuals between the data and the fitted smooth kernel function at time  $t_i$ . The influence of both sources of variability can be estimated in a straightforward way via bootstrap simulation methods [8]:

- (1) Compute  $\hat{y}(t)$  using kernel smoothing with a bandwidth optimized by leave-one-out cross-validation. In order to estimate a model for the variance we apply kernel estimation at the same bandwidth to fit a smooth function  $\hat{\sigma}^2(t)$  to the squared residuals  $\hat{\epsilon}_i^2$ .
- (2) Obtain a resampled data profile of same sample size  $y^*(t_i) = \hat{y}(t_i) + e(t_i)$ ,  $t_i = 1, \dots, n$  where  $e(t_i) \sim N(0, \hat{\sigma}^2(t_i))$ , i.e. using the variance function evaluated at  $t_i$ , and such that  $e(t_i)$  has the same sign as  $\hat{\epsilon}_{t_i}$ .
- (3) Find  $\hat{y}^*(t)$  using kernel smoothing at the same bandwidth as used in step 1.
- (4) Draw values for  $\delta_P$  and  $\delta_M$  from their prior distributions.
- (5) Calculate  $\tilde{M}$  and  $\tilde{\tau}$  using (Eq. 5) and (Eq. 6), respectively.

Steps (2) to (5) are repeated  $R$  times (by default set to 99 in ReTrOS-Smooth), after which point-wise mean estimates for  $\tilde{M}$  and  $\tilde{\tau}$  are plotted along with estimated 95% confidence envelopes for the reconstructed curves. The values of  $\delta_P$  and  $\delta_M$  in step (4) are drawn from a gamma distribution with mean and standard deviation provided by the user. If the time course represents mRNA levels then steps (1) to (5) are implemented analogously where only a prior distribution for  $\delta_M$  is required. An example output from the ReTrOS-Smooth method is shown in Fig. 1.

**ReTrOS-switch: reconstruction of a switching transcription profile**

This method is an implementation of the transcriptional switch model inference introduced in Finkenstädt et al. [1] and Jenkins et al. [3] where it was applied to mRNA time

series data. Here it is further extended to include protein dynamics through (Eq. 2). We assume that the transcriptional rate function  $\tau(t)$  in (Eq. 1) has the profile of a step function where the rate is constant between time points,  $s_i$ , where it changes due to unobserved transcriptional events (for example, activation or inhibition):

$$\tau(t) = \tau_i \text{ for } s_i < t \leq s_{i+1}, i = 1, \dots, k, t \in [0, L] \quad (7)$$

with  $L$  the total length of the time interval observed. Note that transcription might not be fully turned off and that there may be more than just two states. The posterior distributions for the switch-times  $s_1, \dots, s_k$  and number of switches  $k$  are estimated by the software. Extending this model to incorporate protein expression dynamics (Eq. 2) gives the solution:

$$P(t) = P(0)e^{-\delta_P t} + \alpha e^{-\delta_P t} \int_0^t e^{\delta_P u} M(u) du \quad (8)$$

with  $M(t)$  satisfying (Eq. 5) and  $\tau(t)$  as in (Eq. 7).

A fast estimation algorithm can be implemented due to the fact that the ODE system, for given values of the degradation parameters during an iteration of the estimation sampler, can be written as a linear model for functions of the parameters  $P(0), M(0)$  and all  $\tau$ . That is, (Eq. 8) can be written as a linear regression model, which for the protein is of the form:

$$P(t) = \beta_0 X_0^P + \beta_1 X_1^P + \beta_2 X_2^P + \dots + \beta_{k+2} X_{k+2}^P \quad (9)$$

where  $\beta_0 = P(0), \beta_1 = M(0), \beta_2 = \frac{\tau_0}{\delta_M}, \beta_3 = \frac{\tau_1 - \tau_0}{\delta_M}, \dots, \beta_{k+2} = \frac{\tau_k - \tau_{k-1}}{\delta_M}$  and the regressors are

$$\begin{aligned} X_0^P &= e^{-\delta_P t} \\ X_1^P &= \frac{\alpha}{\delta_P - \delta_M} (e^{-\delta_M t} - e^{-\delta_P t}) \\ X_2^P &= \alpha \left( \frac{1 - e^{-\delta_P t}}{\delta_P} - \frac{e^{-\delta_M t} - e^{-\delta_P t}}{\delta_P - \delta_M} \right) \\ &\vdots \\ X_{k+2}^P &= \alpha \left( \frac{1 - e^{-\delta_P(t-s_k)}}{\delta_P} - \frac{e^{-\delta_M(t-s_k)} - e^{-\delta_P(t-s_k)}}{\delta_P - \delta_M} \right) \end{aligned}$$

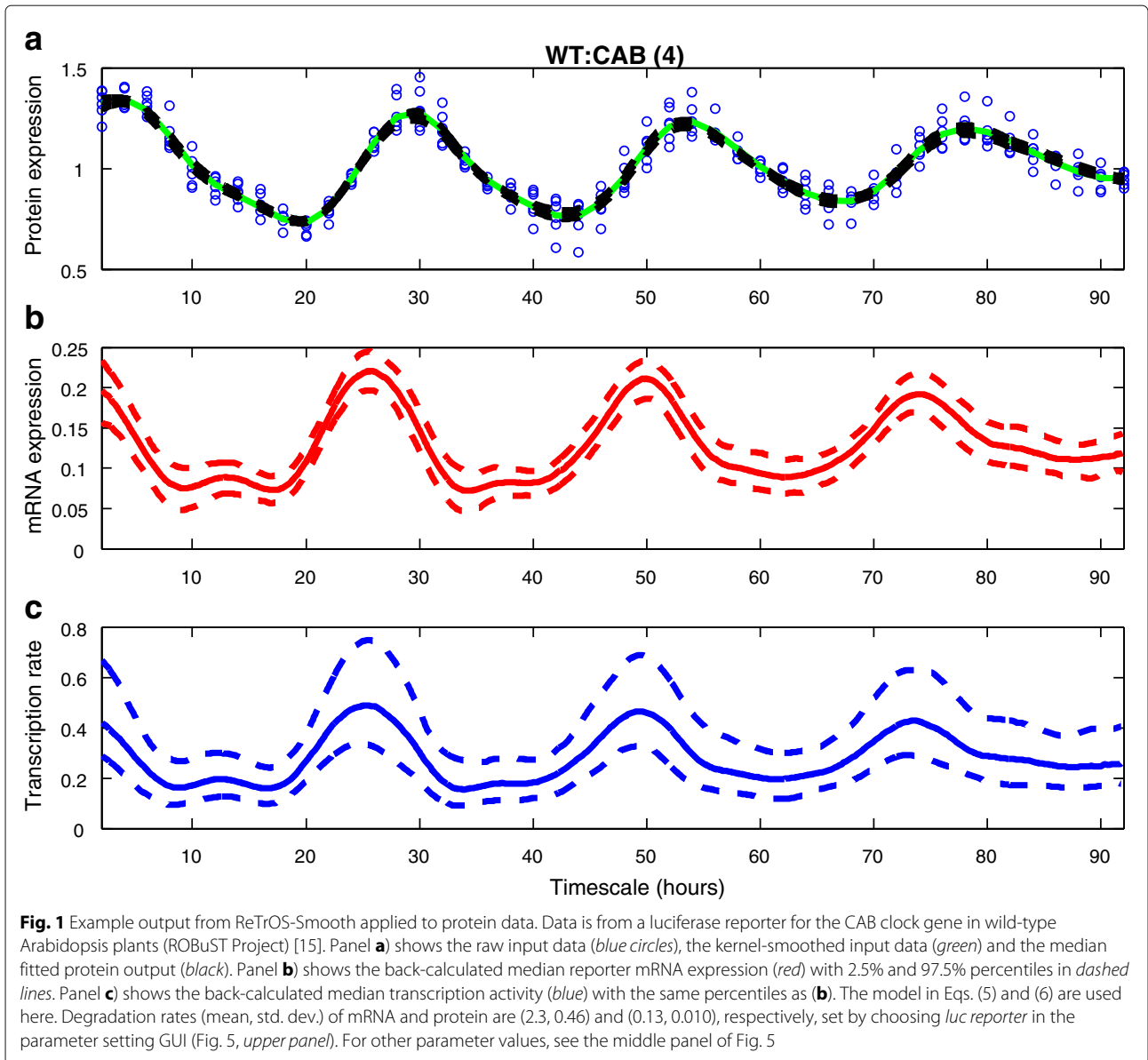
and, analogously, for the mRNA [3]

$$M(t) = \beta_1 X_1^M + \beta_2 X_2^M + \dots + \beta_{k+2} X_{k+2}^M \quad (10)$$

with

$$\begin{aligned} X_1^M &= e^{-\delta_M t}, X_2^M = 1 - e^{-\delta_M t}, \\ X_3^M &= 1 - e^{-\delta_M(t-s_1)}, \dots, X_{k+2}^M = 1 - e^{-\delta_M(t-s_k)}. \end{aligned}$$

**Parameter estimation and model output** Parameter estimation is performed using a Markov Chain Monte Carlo (MCMC) sampler following the methodological approach introduced in [3]. Like for ReTrOS-Smooth the user is required to provide prior means and standard deviations so that an informative Gamma distribution can be



formulated for the degradation rates,  $\delta_P$  and  $\delta_M$ , which are updated using a standard Metropolis-Hastings acceptance scheme. For protein reporter observations (Eq. 3) is used while (Eq. 4) is used for mRNA level observations, where the translation rate parameters,  $\alpha$  and  $\kappa$ , if unknown, are fixed to 1 as in the ReTrOS-Smooth model.

The number and positions of the switch points are sampled using the reversible jump methodology [9]. Estimates of the values for the  $\beta_0$  to  $\beta_{k+2}$  coefficients, representing  $P(0)$ ,  $M(0)$  and  $\tau_0, \dots, \tau_k$ , are obtained through linear regression of the model (Eq. 9) given  $\delta_P$  and  $\delta_M$ , protein expression data and the design matrix  $X^P$ . The regression can be performed using ordinary least squares or a parametric weighted least squares method (see, for example, [10] for details) with weights,

$w_{e(i)} = \hat{M}_{e(i)}$ , an estimate of the given expression at time  $t_i$ , derived by a smoothing spline kernel. If mRNA expression data are provided, the reduced model given in (Eq. 10) can be inferred using  $\delta_M$  and the  $X^M$  design matrix only.

Given the nature of the sampling methodology, we have to summarise the posterior distributions obtained from the simulated chains. Univariate parameter distributions are summarised in terms of percentiles, whilst the switch time distribution is summarised as a Gaussian mixture model, such that each estimated switch event has a mean and a standard deviation derived from the marginal switch time probability density (see [3] for further details). In addition to summarising the sampled distributions across all model sizes, the sampled switch time density for

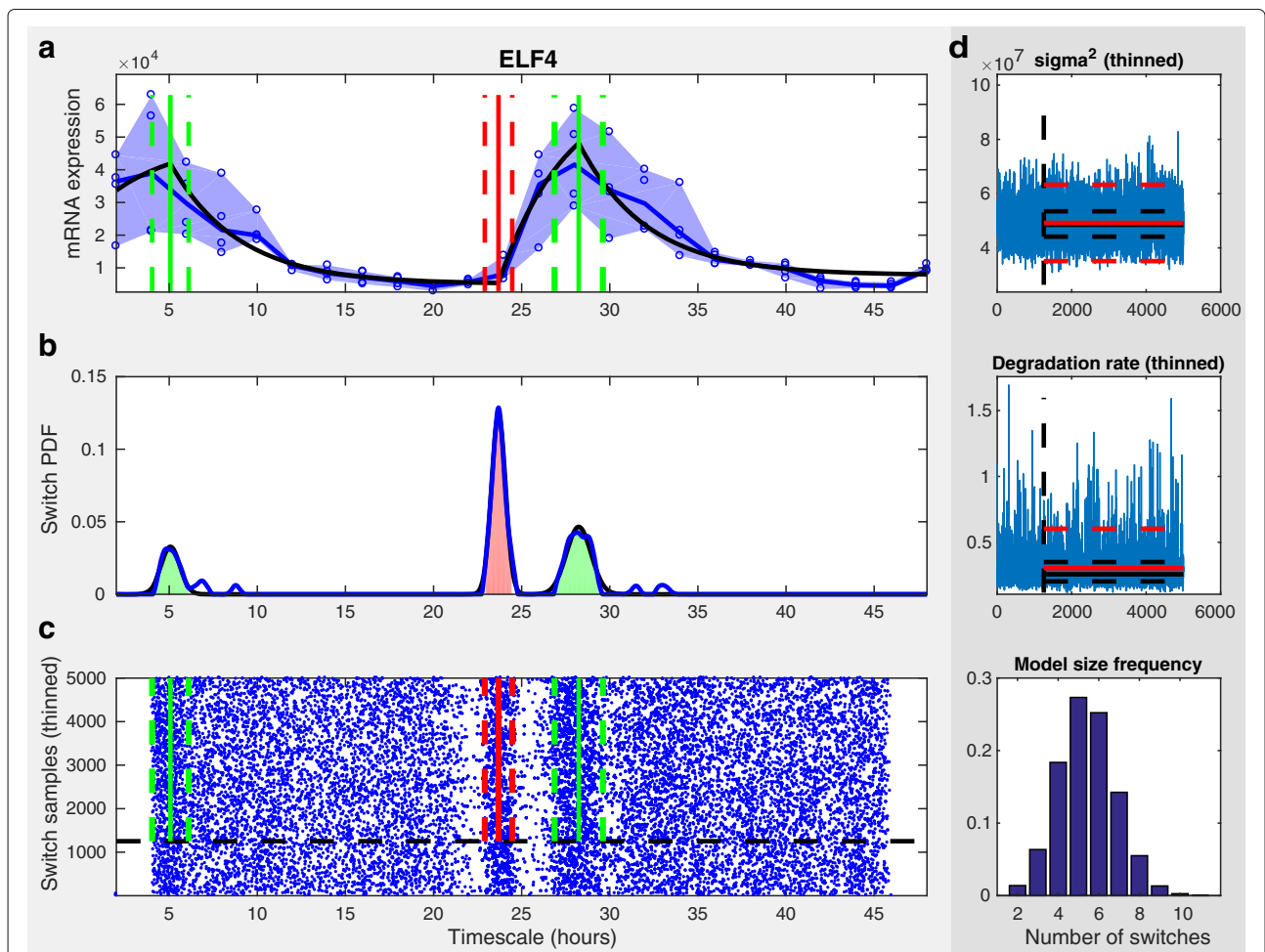
specific model sizes are also shown in order of sampled frequency. Example output summary from the ReTrOS-Switch method applied to microarray mRNA expression data is shown in Fig. 2 and to protein reporter (luciferase) data in Fig. 3.

**Running ReTrOS**

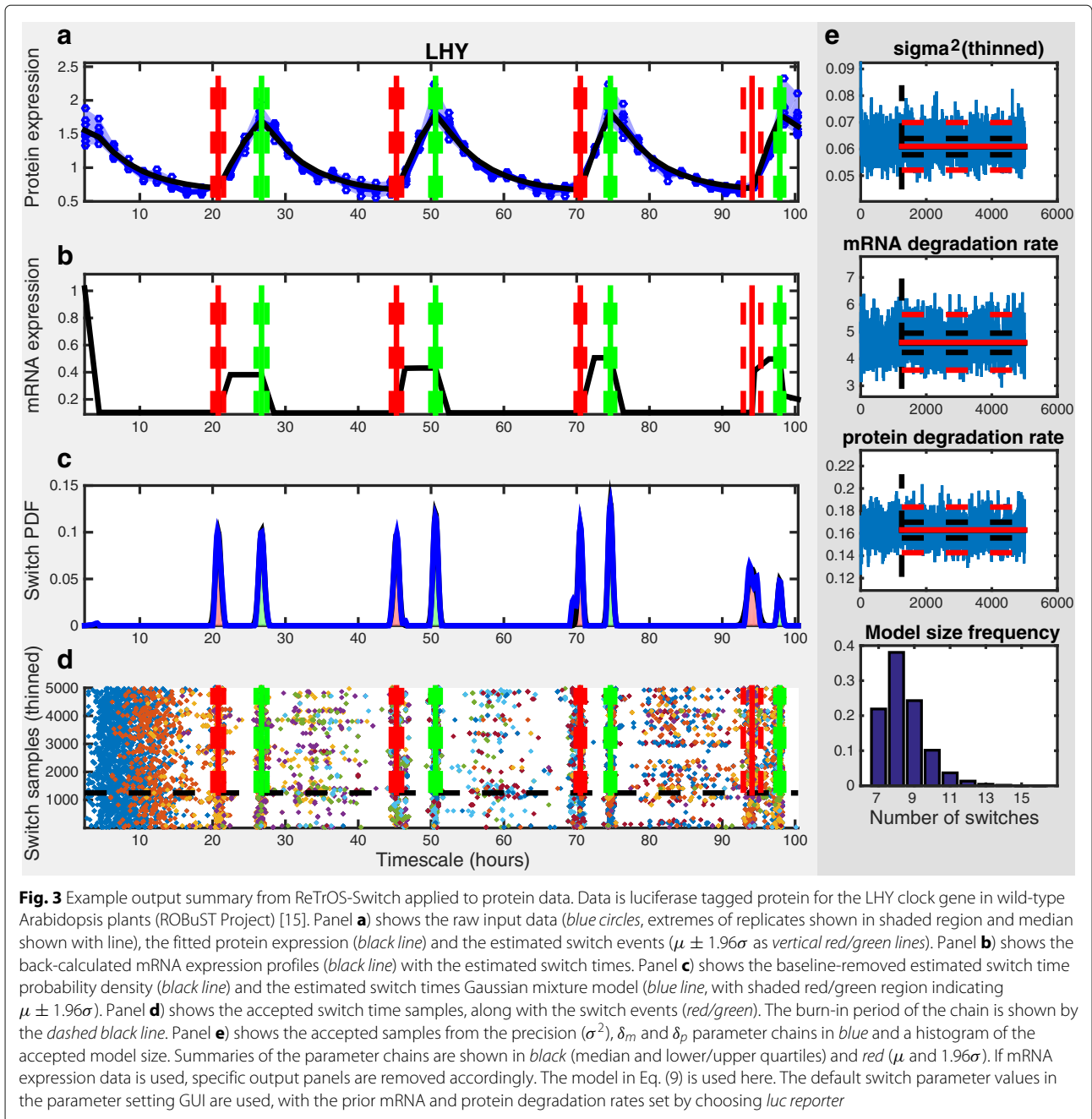
The ReTrOS toolbox provides both a graphical and a command-line interface for selecting input data files and specifying algorithm parameters. Users can either run the analysis directly from the graphical user interface, or alternatively from the command-line or batch scripts generated from the GUI.

**Input data format** The data format used is a simple table generated by standard spreadsheet software. A header row defines the observation time and each row contains the observed data for a single sample, gene or protein. The  $N$  columns in the table can be used to specify the name of the gene/protein or other annotation data and the remaining columns contain the observed values for the corresponding observation time in the header row (see Fig. 4). Replicate samples can be either treated as individual time series, or combined together.

**Algorithm parameters** A range of user-definable parameters are available for both methods. The user interface



**Fig. 2** Example output summary from ReTrOS-Switch applied to mRNA microarray data. Data is from a microarray time-series for the ELF4 clock gene in wild-type Arabidopsis plants (PRESTA project) [12]. Panel **a**) shows the raw input data (blue circles, extremes of replicates shown in shaded region and median shown with line), the fitted mRNA expression (black line) and the estimated switch events ( $\mu \pm 1.96\sigma$  as vertical red/green lines). Panel **b**) shows the baseline-removed estimated switch time probability density (black line) and the estimated switch times Gaussian mixture model (blue line, with shaded red/green region indicating  $\mu \pm 1.96\sigma$ ). Panel **c**) shows the accepted switch time samples, along with the switch events (red/green). The burn-in period of the chain is shown by the dashed black line. Panel **d**) shows the accepted samples from the precision ( $\sigma^2$ ),  $\delta_m$  and  $\delta_p$  parameter chains in blue and a histogram of the accepted model size. Summaries of the parameter chains are shown in black (median and lower/upper quartiles) and red ( $\mu$  and  $1.96\sigma$ ). The model in Eq. (10) is used here. The default switch parameter values in the parameter setting GUI are used, with the prior mRNA degradation rate (mean = 1.5, std. dev. = 0.18), set by choosing Arabidopsis (see Fig. 5, upper panel)

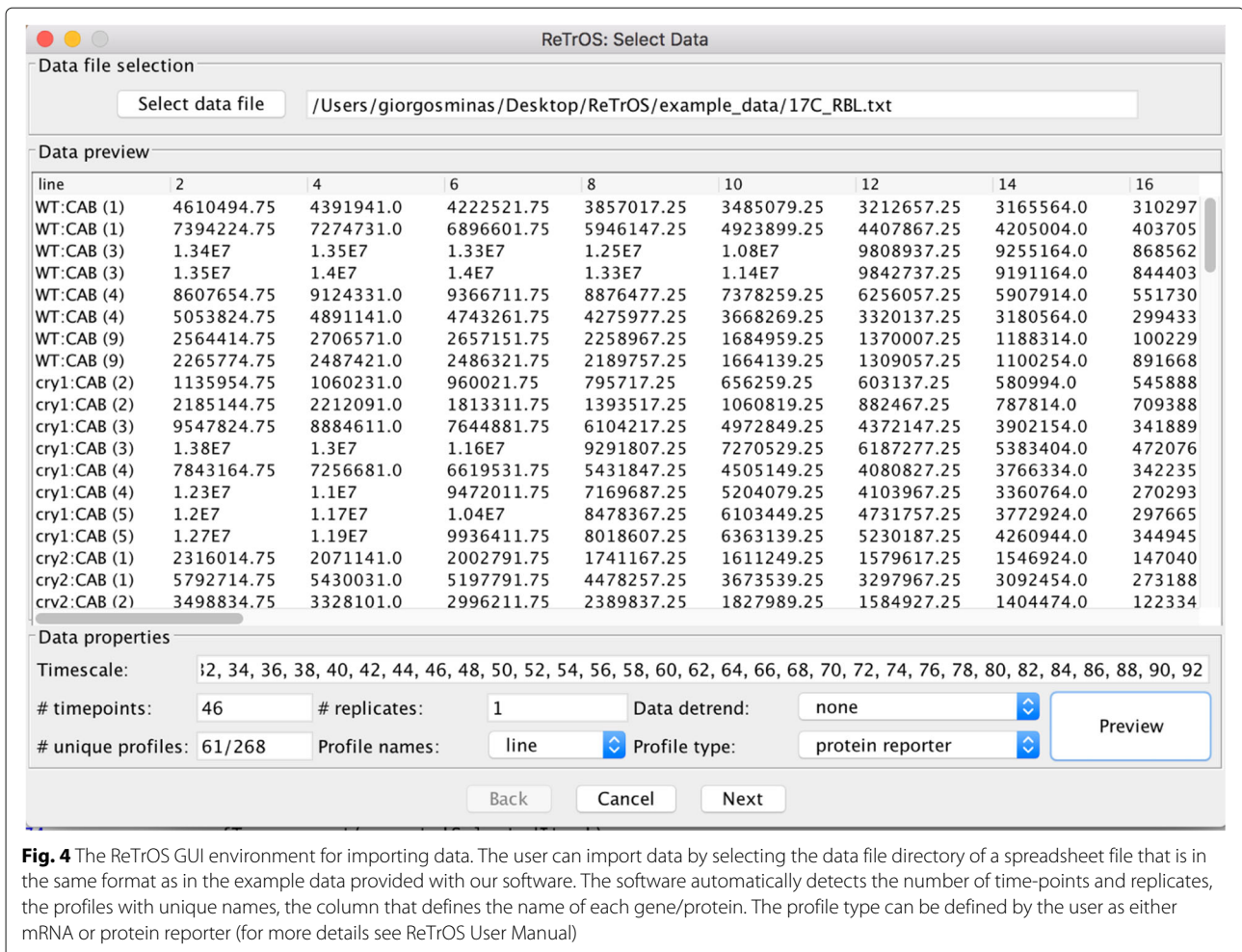


provides default values for required parameters (such as number of iterations or bootstrap samples to run) and each method provides default values for all other parameters. Most algorithmic parameters can be modified through the user interface and all parameters can be modified through the batch scripts (see Fig. 5).

**Algorithm comparison**

Whilst the two algorithms in ReTrOS use the same underlying transcription/translation model, there are a number of considerations as to which method may be

more appropriate to use. The ReTrOS-Smooth method does not require, but can handle, replicate data and due to the non-parametric approach has fewer restrictions on data requirements. Model fitting using the ReTrOS-Switch method is improved by the use of replicate data samples, but requires that the replicate data samples are reasonably synchronised (both in expression range and dynamics) and that changes in transcription activity are expected. The algorithmic complexity of the ReTrOS-Smooth method is lower than for the ReTrOS-Switch method, which also requires a large number of MCMC



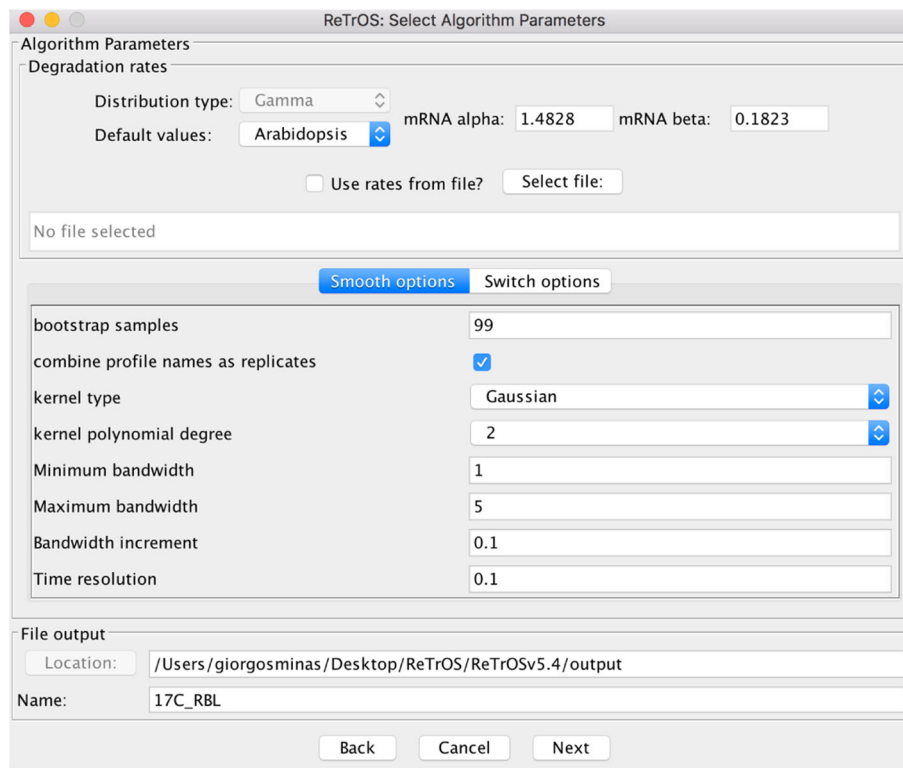
**Fig. 4** The ReTrOS GUI environment for importing data. The user can import data by selecting the data file directory of a spreadsheet file that is in the same format as in the example data provided with our software. The software automatically detects the number of time-points and replicates, the profiles with unique names, the column that defines the name of each gene/protein. The profile type can be defined by the user as either mRNA or protein reporter (for more details see ReTrOS User Manual)

iterations to estimate the parameter distributions accurately. As such, the computational time required for the ReTrOS-Smooth method is far smaller than for the ReTrOS-Switch method. ReTrOS-Smooth processes time series such as in Fig. 1 (4 replicates of time-series of 46 data-points) with the default parameter setting (see Fig. 5) in about 20 s (3.1-GHz Intel Core i7 processor), about 8 times faster on average than ReTrOS-Switch. The computations can be parallelised, which improves processing times particularly for large number of protein reporters or genes, and the current ReTrOS implementation provide the option to use workers of the parallel pool of the local machine. Whilst both methods require informative prior distributions of  $\delta_m$  and  $\delta_p$ , the mRNA and protein degradation rates, ReTrOS-Switch samples the parameter distributions through the Metropolis-Hastings acceptance scheme and provides an updated posterior estimate of the distributions, whereas ReTrOS-Smooth draws directly from the given distributions only. Finally, in addition to the continuous estimate of transcription activity generated by both algorithms, ReTrOS-Switch also generates

distributions over the timing and type of changes in transcriptional activity and therefore it is more appropriate when switching events are of interest.

## Results

The algorithms used in the ReTrOS toolbox have previously been applied to a number of different mRNA and protein reporter data sets, in addition to those in [1]. The study in [2] used the ReTrOS-Smooth to investigate the transcriptional dynamics of two protein-reporter systems, firefly luciferase (*luc*) and destabilized enhanced green fluorescent protein (EGFP), under identical promoter control in mammalian cells. The study in [3] used ReTrOS-Switch to explore correlations between discrete transcriptional ‘switch events’ and promoter structure and also obtained updated degradation rate estimates for 200 genes displaying a circadian rhythm in *Arabidopsis thaliana* leaf samples. Here we also present a new case study applying ReTrOS to recently published mRNA and protein-reporter time series of a selection of central *Arabidopsis thaliana* circadian clock-related genes. Circadian



**Fig. 5** The GUI environment for selecting parameter values of ReTrOS-Smooth. Default parameter values are available to the user. The rate parameter values can also be imported by selecting the directory of the corresponding file. A similar GUI environment is available for the ReTrOS-Switch method (for more details see ReTrOS User Manual)

clocks and rhythms are present in most living organisms and provide a regulatory method for many important processes. A common feature of circadian clocks is that they consist of both transcriptional and translational components. We use the ReTrOS software to explore the oscillatory nature of the mRNA expression and protein reporter time series data from the model plant *Arabidopsis thaliana*.

**Identifying temporal events in the *Arabidopsis thaliana* circadian ‘repressilator’ circuit**

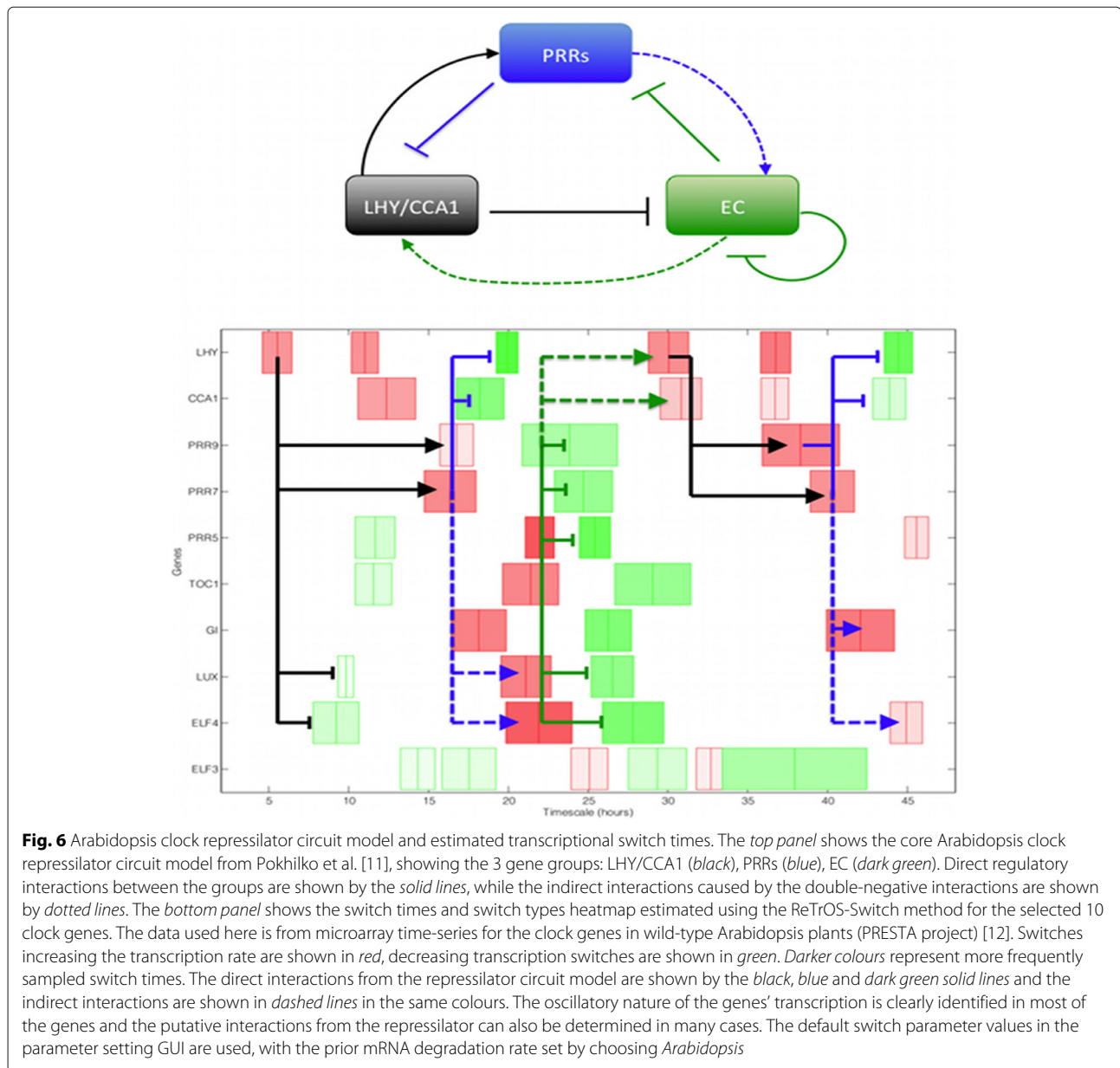
Using the simplified core ‘repressilator’ model (shown in Fig. 6 top) introduced by Pokhilko et al. [11], we aim to identify the direct and indirect regulatory interactions caused by the double-negative feedback in terms of transcriptional switch timing. We selected 10 core clock genes (Table 1) including several from each broad gene group, LHY/CCA1, PRRs and Evening Complex (EC), and analysed mRNA microarray data from the 2-day wild-type time series dataset [12] using ReTrOS-Switch (shown in Fig. 6). The temporal switch distributions (shown at  $\pm 1.96\sigma$  with darker colours representing more frequently sampled switches) clearly show the circadian nature of the transcriptional profiles: for example LHY has three

transcriptional switches, at approximately 6, 10 and 20 h, during the first 24-h observation period, followed by three switches of the same types at approximately the same times during the second 24-h period. However, not all of the analysed profiles displayed identifiable expression dynamics (specifically ELF3) and as such a clear transcriptional switch profile could not be obtained. The putative interactions of the repressilator model can be visually identified in many cases: for example the regulatory interactions by the LHY/CCA1 group (shown in black) can be identified with positive regulation of several PRR genes (increasing transcription switch event to increasing transcriptional switch event) and negative regulation of several EC genes (increasing transcriptional switch event to decreasing transcriptional switch event). The temporal flow of transcriptional regulation in waves can also be identified, with the LHY/CCA interactions followed by the PRR interactions which are followed by the EC interactions.

**Effects of light and temperature conditions on circadian marker genes**

We analysed protein luciferase reporter data from a 4-day time series dataset of the circadian-controlled CCR2 and

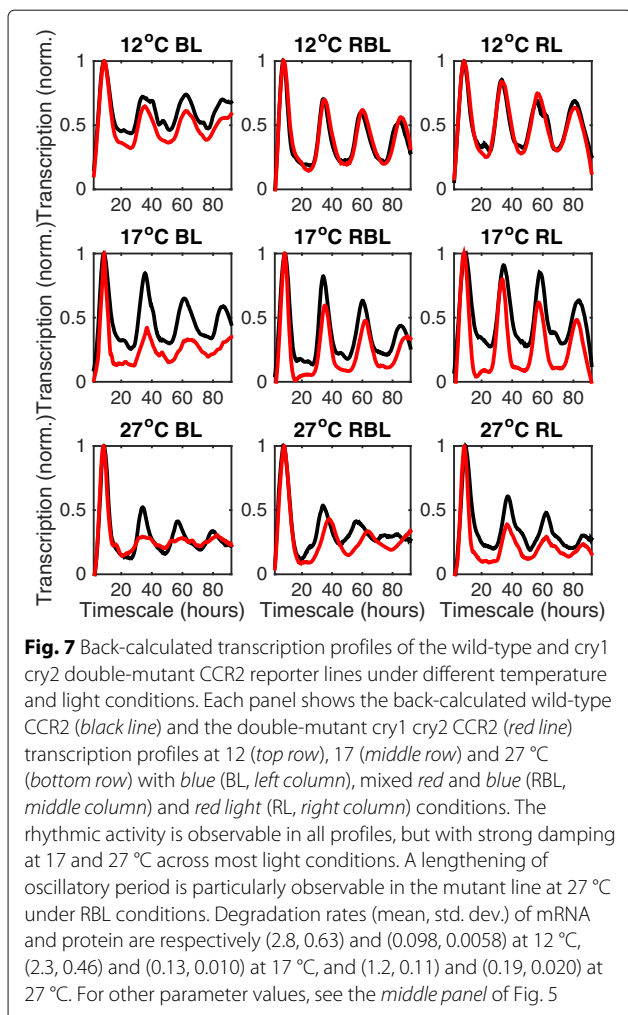




**Table 1** Selected *Arabidopsis thaliana* clock genes

Gene	Group	Locus
LHY	LHY/CCA1	AT1G01060
CCA1	LHY/CCA1	AT2G46830
PRR9	PRRs	AT2G46790
PRR7	PRRs	AT5G02810
PRR5 (NI)	PRRs	AT5G24470
TOC1	-	AT5G61380
GI	-	AT1G22770
LUX	EC	AT3G46640
ELF4	EC	AT2G40080
ELF3	EC	AT2G25930

CAB2 promoters in a range of wild-type and mutant lines, temperature conditions and light regimes [13]. Applying the ReTrOS-Smooth algorithm to the detrended data and combining replicate observations from the same lines yields clear rhythmic back-calculated transcriptional profiles under most experimental conditions. Figure 7 shows the back-calculated transcription profiles for the wild-type CCR2(3) line (shown by the black line) and the *cry1 cry2* CCR2(3) double-mutant line (shown by the red line at temperatures of 12, 17 and 27 °C and blue (BL), red (RL) and mixed red-blue (RBL) light conditions). We observe similar behaviours to those identified within the original study, such as an increased period in the *cry1 cry2* CCR2 mutant line at 27 °C under RBL conditions, however,



the back-calculated transcription profile of the *cry1 cry2* CCR2 still shows damped rhythmic dynamics. As the back-calculation model takes into account the degradation processes of the luciferase reporter, finer-scale structures in the time series are able to be extracted which may allow, for instance, increased accuracy in periodicity inference when using methods robust to asymmetric oscillations such as spectrum resampling [14].

### Conclusions

Analysis of large-scale and high-throughput data is becoming an increasingly common task for many researchers. We provide an easy-to-use toolbox for the analysis of mRNA or protein-reporter time series data, that generates a fine-scale profile of transcriptional activity by removing the effects of degradation processes from the observed data (Additional file 1). The ReTrOS toolbox has been applied to a variety of datasets from a range of different technologies and platforms. ReTrOS can be easily incorporated into a computational or analysis

pipeline as either, a data preprocessing step for rapidly obtaining high-resolution back-calculated transcriptional profiles which are then used in other analysis steps, or directly as an analysis tool extracting distributions of transcriptional ‘switching’ activity and degradation rate estimates. There are a number of possible extensions to the analysis including parallelisation of the ReTrOS-Smooth bootstrap procedure and parallel MCMC sampling in the ReTrOS-Switch algorithm, the application of further analysis methods such as clustering on the model outputs and the use of Bayesian hierarchical modelling for multiple time series [3]. The methodology for a stochastic transcriptional switch model is considerably more challenging and has recently been developed in [5] and [6].

### Availability and requirements

The ReTrOS toolbox, user manual and example data are freely available from <http://www2.warwick.ac.uk/fac/sci/systemsbiology/research/software/> and <https://github.com/giorgosminas/ReTrOS> under the GNU General Public License <https://www.gnu.org/licenses/gpl-3.0.en.html>. The toolbox runs in MATLAB® that can operate in Windows, Mac and Linux (see <https://uk.mathworks.com/support/sysreq.html>).

### Additional file

**Additional file 1:** Toolbox and data. The MATLAB code for running ReTrOS toolbox. This also includes the data used in the illustrative examples and the user manual that provides the instructions on how to run the toolbox. (ZIP 52838 kb)

### Abbreviations

BL: Blue light; CAB2: Chlorophyll A/B-binding protein 2; CCA1: Circadian clock associated 1; CCR2: Cold, circadian rhythm, and RNA binding 2; EC: Evening complex; EGFP: Destabilized enhanced green fluorescent protein; GUI: Graphical user interface; LHY: Late elongated Hypocotyl; luc: firefly luciferase; MCMC: Markov Chain Monte Carlo; mRNA: Messenger Ribonucleic acid; ODE: Ordinary differential equation; PRR: Pseudo-response regulator; RBL: Mixed red-blue light; ReTrOS: Reconstructing transcription open software; RL: Red light

### Acknowledgments

This research is part of the PRESTA (Plant Responses to Environmental STress in Arabidopsis) and the ROBUST (Regulation Of Biological Signalling by Temperature) projects and we thank the groups for high-resolution *Arabidopsis thaliana* microarray and luciferase time series data and for helpful discussions and advice. We thank S. Calderazzo, K. Hey, P. Brown and D. Woodcock for discussions on the methodology and software.

### Funding

This work was supported through providing funds by the Biotechnology and Biological Sciences Research Council [BB/F005806/1, BB/F005237/1]; and the Engineering and Physical Sciences Research Council [EP/C544587/1 to DAR]. None of the funding bodies have played any part in the design of the study, in the collection, analysis, and interpretation of the data, or in the writing the manuscript.

### Authors’ contributions

GM developed ReTrOS-Switch for protein data; DJJ developed ReTrOS-Switch for mRNA data; HM and MJC developed ReTrOS-Smooth; DJJ, HM, MJC and GM all programmed and implemented relevant parts of the software; DJJ

implemented final version of software with user manual; GM tested the final version of software; DAR provided guidance on mathematical modeling; BF suggested and provided statistical guidance on statistical methodology; GM, BF and DJJ wrote the manuscript with the help from all authors; All authors read and approved manuscript submitted.

#### Competing interests

The authors declare that they have no competing interests.

#### Consent for publication

Not applicable.

#### Ethics approval and consent to participate

Not applicable.

#### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

#### Author details

<sup>1</sup>Systems Biology Centre, University of Warwick, Coventry CV34 4HD, UK.

<sup>2</sup>Department of Statistics, University of Warwick, Coventry CV34 4HD, UK.

<sup>3</sup>Mathematics Institute, University of Warwick, Coventry CV34 4HD, UK.

<sup>4</sup>School of Life Sciences, University of Warwick, Coventry CV34 4HD, UK.

Received: 5 December 2016 Accepted: 18 May 2017

Published online: 26 June 2017

#### References

- Finkenstädt B, Heron EA, Komorowski M, Edwards K, Tang S, Harper CV, Davis JRE, White MRH, Millar AJ, Rand DA. Reconstruction of transcriptional dynamics from gene reporter data using differential equations. *Bioinformatics*. 2008;24(24):2901–7.
- Harper CV, Finkenstädt B, Woodcock DJ, Friedrichsen S, Semprini S, Ashall L, Spiller DG, Mullins JJ, Rand DA, Davis JRE, White MRH. Dynamic analysis of stochastic transcription cycles. *PLoS Biol*. 2011;9(4):1000607.
- Jenkins DJ, Finkenstädt B, Rand DA. A temporal switch model for estimating transcriptional activity in gene expression. *Bioinformatics*. 2013;29(9):1158–65.
- Woodcock DJ, Vance KW, Komorowski M, Koentges G, Finkenstädt B, Rand DA. A hierarchical model of transcription dynamics allows robust estimation of transcription rates in populations of single cells with variable gene copy number. *Bioinformatics*. 2013;29(12):1519–25.
- Hey K, Momiji H, Featherstone K, Davis JRE, White MRA, Rand DA, Finkenstädt B. A stochastic transcriptional switch model for single cell imaging data. *Biostatistics*. 2015;16(4):655–69.
- Featherstone K, Hey K, Momiji H, McNamara A, Patist AL, Woodburn J, Spiller D, Christian H, McNeilly A, Mullins J, Finkenstädt B, Rand DA, White MRA, Davis JRE. Spatially coordinated dynamic gene transcription in living pituitary tissue. *eLife*. 2015;16:5–08494.
- Wand MP, Jones MC. Kernel Smoothing. *Monographs on Statistics and Applied Probability*, vol 60. Boca Raton: Chapman and Hall/CRC; 1994.
- Efron B, Tibshirani RJ. An introduction to the bootstrap. *Chapman & Hall/CRC monographs on statistics & applied probability*. Boca Raton: Taylor & Francis; 1994.
- Green PJ. Reversible jump Markov chain Monte Carlo computations and Bayesian model determination. *Biometrika*. 1995;82:711–32.
- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. *Bayesian data analysis*. Boca Raton: Chapman & Hall/CRC; 2004.
- Pokhilko A, Piñas Fernández A, Edwards KD, Southern MM, Halliday KJ, Millar AJ. The clock gene circuit in *Arabidopsis* includes a repressor with additional feedback loops. *Mol Syst Biol*. 2012;8:574.
- Windram O, Madhou P, McHattie S, Hill C, Hickman R, Cooke E, Jenkins DJ, Penfold CA, Baxter L, Breeze E, Kiddle SJ, Rhodes J, Atwell S, Kliebenstein DJ, Kim YS, Stegle O, Borgwardt K, Zhang C, Tabrett A, Legaie R, Moore J, Finkenstädt B, Wild DL, Mead A, Rand D, Beynon J, Ott S, Buchanan-Wollaston V, Denby KJ. *Arabidopsis* defense against *Botrytis cinerea*: Chronology and regulation deciphered by high-resolution temporal transcriptomic analysis. *Plant Cell*. 2012;24(9):3530–57.
- Gould PD, Ugarte N, Domijan M, Costa M, Foreman J, Macgregor D, Rose K, Griffiths J, Millar AJ, Finkenstädt B, Penfield S, Rand DA, Halliday

KJ, Hall AJ. Network balance via CRY signalling controls the *Arabidopsis* circadian clock over ambient temperatures. *Mol Syst Biol*. 2013;9(1):650.

- Costa MJ, Finkenstädt B, Roche V, Lévi F, Gould PD, Foreman J, Halliday K, Hall A, Rand DA. Inference on periodicity of circadian time series. *Biostatistics*. 2013;14(4):792–806.
- Gould PD, Ugarte N, Domijan M, Costa M, Foreman J, MacGregor D, Rose K, Griffiths J, Millar AJ, Finkenstädt B, Penfield S, Rand DA, Halliday KJ, Hall AJW. Network balance via CRY signalling controls the *Arabidopsis* circadian clock over ambient temperatures. *Mol Syst Biol*. 2013;9(1). doi:10.1038/msb.2013.7.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
www.biomedcentral.com/submit

