The effects of cell asynchrony on gene expression levels: analysis and application to *Plasmodium falciparum*

Wei ZHAO, *Student Member, IEEE,* Justin DAUWELS, *Member, IEEE,* and Jianshu CAO

***Abstract*—To investigate the intraerythrocytic developmental cycle of *Plasmodium falciparum*, time-series gene expression data is commonly measured of infected red blood cells. However, the observed data is usually blurred due to cell asynchrony during experiments. In this paper, the effect of cell asynchrony is investigated by conducting numerical experiments. The simulation results suggest that cell asynchrony has varying effects on different intrinsic expression patterns. Specifically, the intrinsic patterns with high expression around the late life stage are more likely to be affected by cell asynchrony. It is also investigated how the effect of cell asynchrony depends on the experimental conditions. From this analysis, the burst rate $r\%$ in infection period and the standard deviation σ of growth rate are identified to have a strong impact on the blurring due to cell asynchrony. Consequently, it is important to measure these two parameters during biological experiments in order to deblur time-series gene expression data.**

***Index Terms*—Malaria, *Plasmodium falciparum*, Gene expression level, Cell synchrony, Computational synchronization.**

I. Introduction

Approximately 207 million people are infected by malaria, and in 2012, about 627,000 people died from this disease [2]. Malaria is transmitted when female Anopheles mosquitoes bite humans, and inject sporozites that invade the liver, resulting in the production of merozoites. These merozoites are released into the blood, invading red blood cells (RBCs). The infected red blood cells (iRBCs) undergo an intraerythrocytic developmental cycle (IDC) that is characterized by three stages, namely: rings, trophozoites, and schizonts [1]. *Plasmodium falciparum* (*P. falciparum*) is the most fatal *Plasmodium* species which causes human malaria. In many efforts to understand the blood stages of the *P. falciparum* infection, time-series gene expression data are measured over the 48-hour IDC [1], [3]–[5]. To better investigate the underlying mechanisms, the parasites are required to be at equivalent stages while the gene expression data are being measured. Therefore, sorbitol treatments are usually conducted to obtain iRBCs synchronized at the schizont stage [6], [7]. Once these schizonts burst, merozoites are released, which further invade fresh RBCs. Therefore, a complete IDC can be observed from the newly infected RBCs over the next 48 hours.

All authors are with Singapore-MIT Alliance for Research and Technology, 1 CREATE Way, Singapore 138602.
W. ZHAO and J. DAUWELS are with School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798.
J. CAO is with Department of Chemistry, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge MA 02139, USA.
Thank Prof. Zbynek Bozdech for granting permission to include Fig. 1(a) from [1] in our manuscript.
The authors thank Ms. Smitha Velayil for proofreading this manuscript.
Manuscript received December 4, 2014; revised March 20, 2015.

However, a perfect synchronization is difficult to achieve. Although the experiment is initialized with synchronized schizonts, the parasite cultures gradually lose synchrony. Consequently, the intrinsic gene expression patterns remain blurred in the observed gene expression data. In our earlier work, we developed a linear system to model the superposition across cells over the IDC [8]. In particular, the decay of cell synchrony is described as a cell age distribution which changes over the course of the experiment. The cell age distributions at different time points constitute the observation matrix of the linear system. The cell asynchrony in other cells has been studied earlier, such as yeast [9] and *Caulobacter crescentus* [10]. The decay of cell synchrony strongly depends on the type of cell and the treatments conducted on them to enhance the cell synchrony. Therefore, dedicated effort is required to model the asynchrony of different cell types. We model the cell asynchrony of *P. falciparum* over a 48-hour IDC. To the best of our knowledge, no such models have been proposed before for *P. falciparum*.

This manuscript is an extended version of an earlier conference paper [11]. Additional figures, results, and discussions are included in this journal version. In this paper, we analyze the linear model proposed in [8] to better understand the effects of cell asynchrony. There are two questions that we are specifically interested in:

1) Are there specific shapes of intrinsic expression patterns that are more likely to be affected by cell asynchrony?
2) How do the effects of cell asynchrony on gene expression levels depend on the experimental conditions?

We conduct simulations on synthetic intrinsic gene expression patterns to answer these two questions. We investigate the effect of cell asynchrony by assessing the difference between the intrinsic patterns and the observed expression data. From this analysis, we observe that cell asynchrony depends on the shape of intrinsic expression patterns. Specifically, intrinsic patterns with high expression around the late life stage are more likely to be affected. We also model different experimental conditions by changing the values of the parameters in the linear system. We observe that the burst rate $r\%$ during infection period, and the standard deviation σ of the growth rate, have a strong effect on cell asynchrony. This observation also highlights the importance of estimating the parameters $r\%$ and σ in order to deblur the observed time-series data. With better estimates of the parameters that have a strong effect on cell asynchrony, more accurate intrinsic gene expression patterns may be reconstructed [8].

In Section II, we review our linear model of cell asynchrony. In Section III, we analyze the effect of different parameters, within that model, on cell asynchrony, and in Section IV, we

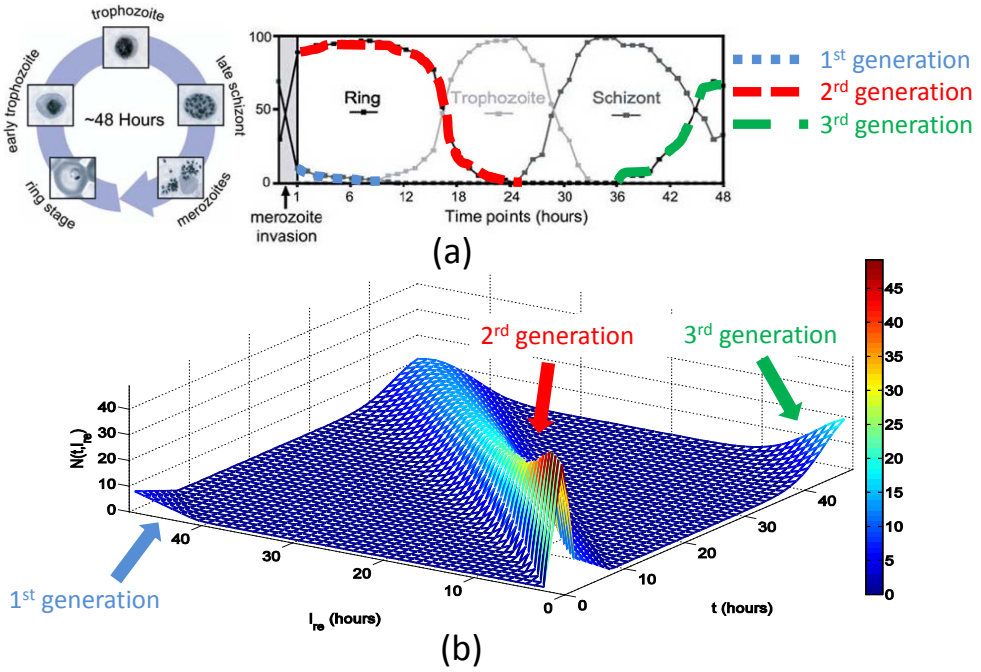

Fig. 1. (a) Percentage representation of the three generations of iRBCs observed in the experiment [1]. (Permission to reproduce this figure was obtained from the authors of [1]). (b) The decay of cell asynchrony is described as the cell age distribution $N(t, \ell_{\text{re}})$ of the three generations of iRBCs which changes over the course of the experiment.

discuss our results. In section V, we discuss how the parameters can be manipulated in experimental settings. Conclusions are drawn at the end of the paper.

II. Model

Here we briefly review the model for cell asynchrony proposed in our earlier work [8]. This forms a guide to the rest of this paper.

Although the IDC of *P. falciparum* lasts for about 48 hours, it slightly varies between different strains. For example, three microarray experiments were respectively conducted over 48, 50 and 53 hours respectively, to cover the IDC of three different *P. falciparum* strains [4]. Therefore, the length of IDC is assumed to be a variable L in our model. For simplification, the value of L will be set to 48 in our numerical experiments.

Gene expression levels are measured at discrete time points over L-hour IDC. The resulting observed expression data $e_i(t)$ at time point t can be modeled as an integral over one life span of infected red blood cells (iRBCs):

$$e_i(t) = \int_0^L N(t, \ell_{\text{re}}) f_i(\ell_{\text{re}}) \mathrm{d}\ell_{\text{re}}, \tag{1}$$

which can be approximated as the following sum [8]:

$$e_i(t) \approx \sum_{\ell_{\text{re}}=1}^{L} N(t, \ell_{\text{re}}) f_i(\ell_{\text{re}}) \triangle \ell_{\text{re}}, \tag{2}$$

where $\{N(t,1), N(t,2), ..., N(t,L)\})$ denotes the cell age distribution of iRBCs at the time point t, and $\{f_i(1), f_i(2), ...f_i(L)\}$ denotes the intrinsic gene expression pattern of a specific gene i; the latter is a virtual expression pattern measured from a perfectly-synchronized cell population over the exact L-hour life span. From (2), a linear system can be derived to model the relationship between intrinsic pattern $f_i(\ell_{\text{re}})$ and observed expression data $e_i(t)$:

$$\underbrace{\begin{pmatrix} N(1,1) & \dots & N(1,L) \\ N(2,1) & \dots & N(2,L) \\ \vdots & \ddots & \vdots \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} f_i(1) \\ \vdots \\ f_i(L) \end{pmatrix}}_{\mathbf{x}} = \underbrace{\begin{pmatrix} e_i(1) \\ \vdots \end{pmatrix}}_{\mathbf{b}}. \tag{3}$$

The constant vector b denotes the observed gene expression data $e_i(t)$. The unknown variable vector x stands for the intrinsic expression pattern $f_i(\ell_{\text{re}})$. The element of the observation matrix $N(t, \ell_{\text{re}})$ denotes the relative number of iRBC that stays at the rescaled cell age l_{re} at time point t, which is calculated as [8]:

$$\begin{aligned} N(t, \ell_{\text{re}}) = &\int_t^{+\infty} S(t') p_{\widetilde{L}} \left(\frac{t'-t}{L-\ell_{\text{re}}} \right) \frac{t'-t}{(L-\ell_{\text{re}})^2} \mathrm{d}t' \\ &+ \int_{-\infty}^{t} R(t') p_{\widetilde{L}} \left(\frac{t-t'}{\ell_{\text{re}}} \right) \frac{t-t'}{{\ell_{\text{re}}}^2} \mathrm{d}t' \\ &+ \int_{-\infty}^{t} R_f(t') p_{\widetilde{L}} \left(\frac{t-t'}{\ell_{\text{re}}} \right) \frac{t-t'}{{\ell_{\text{re}}}^2} \mathrm{d}t', \end{aligned} \tag{4}$$

where the details of $S(t)$, $R(t)$, and $R_f(t)$ will be discussed in the following paragraphs.

As shown in Figure 1 (a), three generations of iRBCs appear in the experiment over the L-hour IDC. The first generation stands for the late-stage iRBCs which are used to infect fresh RBCs to initialize the experiment. The fresh RBCs infected at the beginning of the experiment constitute the second

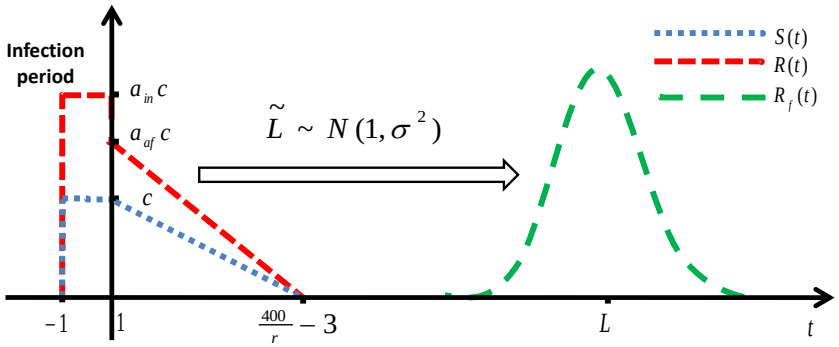

Fig. 2. Each of the functions $S(t)$, $R(t)$, and $R_f(t)$ corresponds to one of the three generations of iRBCs appearing in the experiment. These functions are respectively described by the following key parameters: the burst rate in the infection period $r\%$, the infection factors $\{a_{\text{in}}, a_{\text{af}}\}$, and the standard deviation of the growth rate σ. The first data point is assumed to be at $t = 1$, and therefore, the infection period is the time range $\{t| -1 < t < 1\}$.

generation. Due to the diversity of growth rates, a few fast-growing iRBCs of the second generation will burst and infect additional RBCs. As a result, the third generation of iRBCs appears at the end of the experiment. The three integrals that appear in (4) correspond to the three generations of iRBCs. Specifically, $S(t)$ denotes the number of first generation iRBCs that have burst at time t; $R(t)$ stands for the number of second generation iRBCs infected at time t; $R_f(t)$ denotes the number of third generation iRBCs infected at time t. These three functions are essential to calculate the observation matrix $N(t, \ell_{\text{re}})$, which describes the cell age distribution of the three generations iRBCs over the L-hour IDC as shown in Figure 1 (b). In the remainder of this section, we review the key parameters involved in describing $S(t)$, $R(t)$, and $R_f(t)$.

A. Burst rate in infection period

The number of first generation iRBCs which burst at time t is denoted as $S(t)$. We derive the expression of $S(t)$ based on the percentage of first generation iRBCs which burst in the two-hour infection period. According to the experimental specifications [8], $r\%$ of the first generation iRBCs burst in the two-hour infection period prior to the experiment. The remaining $1 - r\%$ iRBCs remain alive and continually infect fresh RBCs till around h hours after the two-hour infection period. Therefore, $S(t)$ is approximated as a piecewise function:

$$S(t) = \begin{cases} c, & \text{if } -1 \le t < 1, \\ at + b, & \text{if } 1 \le t \le h, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

which satisfies the conditions:

$$\begin{cases} S(1) = c, \\ S(h) = 0, \\ \frac{\int_{-1}^{1} S(t)\mathrm{d}t}{\int_{1}^{h} S(t)\mathrm{d}t} = \frac{r}{100-r}. \end{cases} \tag{6}$$

Hence, (5) can be written as a function of $r\%$:

$$S(t) = \begin{cases} c, & \text{if } -1 \le t < 1, \\ \frac{rc}{4(r-100)}t + \frac{3r-400}{4(r-100)}c, & \text{if } 1 \le t \le \frac{400}{r} - 3, \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

where c takes a positive values, as illustrated in Figure 2.

B. Infection factors

The number of second generation iRBCs infected at time t is denoted as $R(t)$. Since the second generation iRBCs are infected by first generation iRBCs, $R(t)$ is proportional to the number of first generation iRBCs that burst at time t:

$$R(t) = \begin{cases} a_{\text{in}} S(t), & \text{if } t \in [\text{infection period}], \\ a_{\text{af}} S(t), & \text{if } t \in [\text{after infection period}], \end{cases} \tag{8}$$

where the average number of RBCs infected by one iRBC during and after the infection period is respectively denoted by a_{in} and a_{af} respectively. Since the cell concentration is reduced by diluting the culture after the infection period [1], [4], the value of a_{af} is expected to be smaller than a_{in}.

C. Distribution of normalized life span

Individual iRBCs grow at different growth rates. The normalized life span of iRBCs, $\widetilde{L}$, is assumed to be a Gaussian random variable $\widetilde{L} \sim N(1, \sigma^2)$. Given the probability density function $p_{\widetilde{L}}(l)$ of the normalized life span $\widetilde{L}$, the number of third-generation iRBCs infected at time t can be derived from $R(t)$, as [8]:

$$R_f(t) = \frac{a_{\text{af}}}{L} \int_{-\infty}^{+\infty} R(t') p_{\widetilde{L}} \left(\frac{t - t'}{L} \right) \mathrm{d}t'. \tag{9}$$

Overall, the three functions $S(t)$, $R(t)$, and $R_f(t)$ are described by four key parameters $\{r\%, a_{\text{in}}, a_{\text{af}}, \sigma\}$, as shown in Figure 2. The elements of the observation matrix $N(t, \ell_{\text{re}})$ can be calculated by substituting the expressions of $S(t)$, $R(t)$, and $R_f(t)$ into (4).

III. Analysis

In this section, we conduct simulations on synthetic intrinsic gene expression patterns. The observed expression pattern b is obtained by substituting the intrinsic pattern $f_i(\ell_{\text{re}})$ into the linear system (3). The difference between the observed pattern b and intrinsic pattern $f_i(\ell_{\text{re}})$ is calculated to investigate the effects of cell asynchrony. As discussed in the previous section, the linear system is dominated by three groups of parameters: the burst rate $r\%$ in the infection period, the infection factors $\{a_{\text{in}}, a_{\text{af}}\}$, and the standard deviation σ of the growth rate. We also consider different values within the

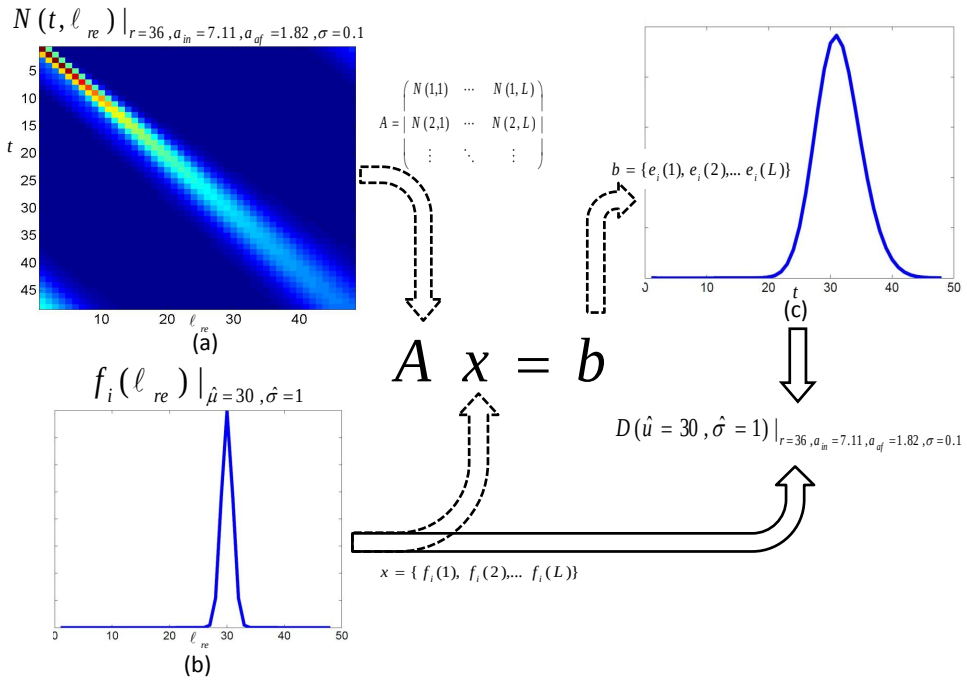

Fig. 3. The flow diagram of the numerical experiments: (a) The elements of the observation matrix $N(t, \ell_{\text{re}})$ are calculated for values of the burst rate in the infection period $r\%$, the infection factors $\{a_{\text{in}}, a_{\text{af}}\}$, and the standard deviation of the growth rate σ. (b) The synthetic intrinsic gene expression pattern $f_i(\ell_{\text{re}})$ is generated with specific positions of the peak $(e.g.: \hat{\mu} = 30)$ and the bell shape $(e.g.: \hat{\sigma} = 1)$. (c) The observed expression data $e_i(t)$ is simulated by substituting the intrinsic pattern $f_i(\ell_{\text{re}})$ into the linear system denoted by $N(t, \ell_{\text{re}})$. The difference between intrinsic pattern and observed data $D(\hat{\mu}, \hat{\sigma})$ is calculated to measure the effects of cell asynchrony.

parameters of the linear system $\{r\%, a_{\text{in}}, a_{\text{af}}, \sigma\}$ to model different experimental conditions.

A. Synthetic gene expression patterns

A 'just-in-time' principle has been proposed to describe the gene expression of *P. falciparum* within the 48-hour IDC [1], [12]. The transcript level of a gene involved in the key biological processes of IDC is expected to peak, just before the encoded protein is needed. Therefore, synthetic gene expression patterns $f_i(\ell_{\text{re}})$ are generated by utilizing the bell curve of a normal distribution. Each of them simulates a synthetic gene with high expression level at different stages in its life span. The mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$ of a normal distribution correspond to the position of the peak and width of the bell curve.

The gene expression patterns $f_i(\ell_{\text{re}})$ present the change of gene expression levels over one life span. Once the iRBCs reach the end of their life span, they will burst and start the next life cycle. Hence, the expression level at the first data point $f_i(1)$ is highly correlated with the expression level at the last data point $f_i(L)$. Therefore, we simply assume that $f_i(\ell_{\text{re}})$ has the same value at $\ell_{\text{re}} = 1$ and $\ell_{\text{re}} = L$. The synthetic gene expression patterns are generated as follows:

$$f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}} = \sum_{i=\{-1,0,1\}} \frac{1}{\hat{\sigma}\sqrt{2\pi}} \left[e^{\frac{-(\ell_{\text{re}}-\hat{\mu}+iL)^2}{2\hat{\sigma}^2}} \right]. \qquad (10)$$

B. Effects of cell asynchrony

We assess the difference between the synthetic intrinsic pattern $f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}}$ and observed expression data $e_i(t)$ by a measure $D(\hat{\mu}, \hat{\sigma})$ defined as follows:

$$D(\hat{\mu}, \hat{\sigma}) = \int_0^L \left| \frac{e_i(t)}{\int_0^L e_i(t)\mathrm{dt}} - \frac{f_i(t)|_{\hat{\mu},\hat{\sigma}}}{\int_0^L f_i(t)|_{\hat{\mu},\hat{\sigma}}\mathrm{dt}} \right| \mathrm{dt}. \qquad (11)$$

Since only the trend of the expression data is of interest here, the values of $f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}}$ and $e_i(t)$ are normalized across the cell life span. The observed expression data $e_i(t)$ are measured at discrete data points. By substituting (3) into (11), the expression of $D(\hat{\mu}, \hat{\sigma})$ can be written in discrete form as:

$$D(\hat{\mu}, \hat{\sigma}) = \text{Sum}\left(\left| \frac{Ax}{\text{Sum}\,(Ax)} - \frac{x}{\text{Sum}\,(x)} \right| \right), \qquad (12)$$

where x denotes the intrinsic expression pattern $\{f_i(1), f_i(2), \ldots f_i(L)\}|_{\hat{\mu},\hat{\sigma}}$, and A stands for the observation matrix consisting of the cell age distribution $N(t, \ell_{\text{re}})$ (3).

As shown in Figure 3, there are three steps in our numerical experiments. Firstly, the value of the parameters $\{r\%, a_{\text{in}}, a_{\text{af}}, \sigma\}$ are chosen to model the experimental conditions, following which the linear system is generated from these parameters. Specifically, the elements of the observation matrix A of the linear system are calculated according to (4). Secondly, the bell-curved synthetic intrinsic patterns $f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}}$ are generated with values $\{\hat{\mu}, \hat{\sigma}\}$, which respectively denote the position and the shape of the bell curve respectively. Thirdly, the observed patterns are obtained by substituting the intrinsic pattern $f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}}$ into the linear system. The effect of cell asynchrony is measured as the difference between the observed pattern and intrinsic pattern according to (12). The details of the experiments are described

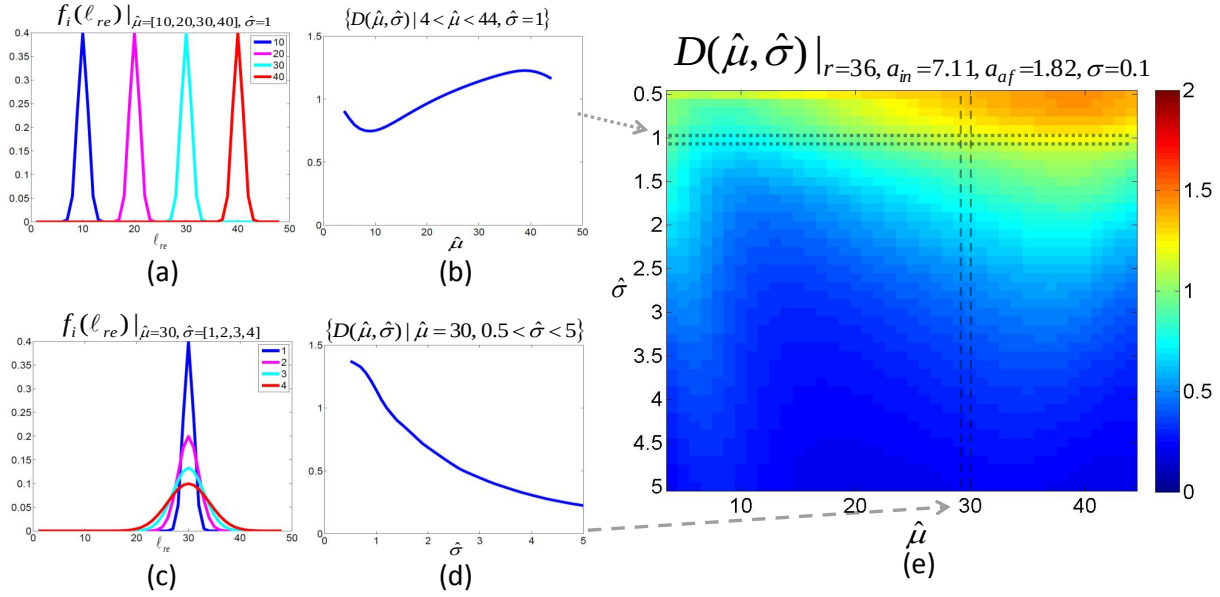

Fig. 4. (a) Synthetic intrinsic patterns $f_i(\ell_{\text{re}})$ with fixed shape ($\hat{\sigma}=1$) and varying position of the bell curve ($\hat{\mu}=[10,20,30,40]$). (b) The row vector of the matrix $D(\hat{\mu},\hat{\sigma})$ with $\hat{\sigma}=1$. (c) Synthetic intrinsic patterns $f_i(\ell_{\text{re}})$ with varying shape ($\hat{\sigma}=[1,2,3,4]$) and fixed position of the bell curve ($\hat{\mu}=30$). (d) The column vector of the matrix $D(\hat{\mu},\hat{\sigma})$ with $\hat{\mu}=30$. (e) The 2-D plot of $D(\hat{\mu},\hat{\sigma})$ with parameters $\{r\%=36\%, a_{\text{in}}=7.11, a_{\text{af}}=1.82, \sigma=0.1\}$.

in Algorithm 1. The experimental results will be discussed in next section.

Algorithm 1 Calculating the matrix $D(\hat{\mu},\hat{\sigma})$ with given parameters $\{r\%, a_{\text{in}}, a_{\text{af}}, \sigma\}$.

1: Initialize the value of $\{r\%, a_{\text{in}}, a_{\text{af}}, \sigma\}$ and substitute them into the expressions of $S(t)$, $R(t)$, and $R_f(t)$, given by (7), (8), and (9), respectively.
2: Calculate $N(t,\ell_{\text{re}})$ by substituting the expressions of $S(t)$, $R(t)$, and $R_f(t)$ into (4), and next compute the observation matrix A:

$$\mathbf{A}=\begin{pmatrix} N(1,1) & \dots & N(1,L) \\ N(2,1) & \dots & N(2,L) \\ \vdots & \ddots & \vdots \end{pmatrix}.$$

3: **for** all reasonable values of $\{\hat{\mu},\hat{\sigma}\}$ **do**
4: Generate the synthetic gene expression pattern with $\{\hat{\mu},\hat{\sigma}\}$, according to equation (10):

$$x=\{f_i(1),f_i(2),...f_i(L)\}|_{\hat{\mu},\hat{\sigma}}.$$

5: Calculate the observed expression data b by substituting x in the linear system (3):

$$b=Ax.$$

6: Calculate the difference between intrinsic pattern x and observed data Ax according to equation (12)
7: **end for**
8: Return $D(\hat{\mu},\hat{\sigma})|_{\{r\%,a_{\text{in}},a_{\text{af}},\sigma\}}$.

IV. Results

In this section, we investigate the effects of cell asynchrony on different expression profiles, and also how these effects depend on model parameters (and hence experimental conditions).

A. Effects on different expression patterns

Algorithm 1 is executed to calculate the measure $D(\hat{\mu},\hat{\sigma})$ for different synthetic intrinsic patterns $f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}}$. The values of the parameters $\{r\%, a_{\text{in}}, a_{\text{af}}, \sigma\}$ are fixed as $\{36\%, 7.11, 1.82, 0.1\}$, estimated from experimental specifications in our earlier study [8]. The synthetic intrinsic patterns $f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}}$ are generated with values of $\{\hat{\mu},\hat{\sigma}\}$ in the range of $\{\hat{\mu},\hat{\sigma}|4<\hat{\mu}<44, 0.5<\hat{\sigma}<5\}$. The difference $D(\hat{\mu},\hat{\sigma})$ is calculated between each synthetic intrinsic pattern x and its corresponding observed pattern b, according to (12).

To have a better understanding, we interpret a row and column vector of the 2-D plot of $D(\hat{\mu},\hat{\sigma})$ separately, as depicted in Figure 4. The parameters $\{\hat{\mu},\hat{\sigma}\}$ denote the position and the shape of the bell curve used, respectively, to generate the synthetic intrinsic pattern $f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}}$, as demonstrated in Figure 4(a)(c).

The row vector $\{D(\hat{\mu},\hat{\sigma})|4<\hat{\mu}<44,\hat{\sigma}=1\}$ indicates how the difference $D(\hat{\mu},\hat{\sigma})$ depends on the intrinsic patterns $f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}}$ with different positions of the peak ($4<\hat{\mu}<44$) but with a fixed shape ($\hat{\sigma}=1$). As shown in Figure 4(b), the row vector $\{D(\hat{\mu},\hat{\sigma})|4<\hat{\mu}<44,\hat{\sigma}=1\}$ decreases when the value of $\hat{\mu}$ changes from 4 to 10, after which the trend reverses after $\hat{\mu}$ further increases towards 44. The highest value of the row vector $\{D(\hat{\mu},\hat{\sigma})|4<\hat{\mu}<44,\hat{\sigma}=1\}$ is obtained around $\hat{\mu}=40$. The pattern of this row vector suggests that cell asynchrony has the strongest effect on the intrinsic patterns $f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}}$ when the expression level peaks at a late stage in the life cycle, and is substantially weaker at the early stages of the life cycle. This is a common observation for all row vectors in the 2-D plot of $D(\hat{\mu},\hat{\sigma})$ (see Figure 4(e)).

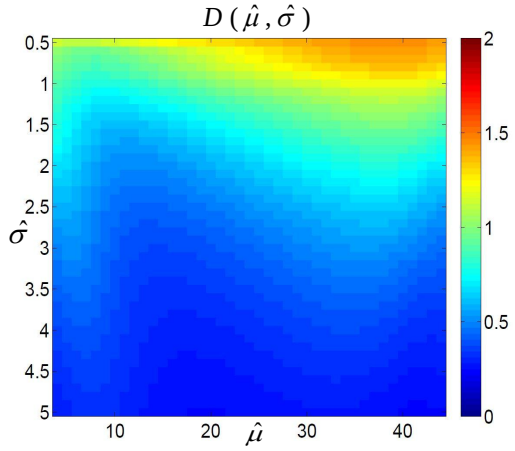

Fig. 5. The average value of $D(\hat{\mu}, \hat{\sigma})$ which is obtained by executing Algorithm 1 one thousand times with the randomly selected values of $\{r\%, a_{\text{in}}, a_{\text{af}}, \sigma\}$.

Similarly, the column vector $\{D(\hat{\mu}, \hat{\sigma}) | \hat{\mu} = 30, 0.5 < \hat{\sigma} < 5\}$ shows how $D(\hat{\mu}, \hat{\sigma})$ varies with intrinsic patterns $f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}}$ with a fixed position of the peak $\hat{\mu} = 30$ but with varying shape ($0.5 < \hat{\sigma} < 5$). As shown in Figure 4(d), values in the column vector $\{D(\hat{\mu}, \hat{\sigma}) | \hat{\mu} = 30, 0.5 < \hat{\sigma} < 5\}$ continuously decrease when $\hat{\sigma}$ changes from 0.5 to 5. This suggests that the cell asynchrony has a continuously decreasing effects on intrinsic patterns $f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}}$, if its bell-shaped expression curve becomes more dispersed. This is also the common conclusion that can be drawn from all column vectors of the $D(\hat{\mu}, \hat{\sigma})$.

As discussed earlier on in this paper, the experimental conditions are described by four parameters: the burst rate in infection period $r\%$, the infection factors $\{a_{\text{in}}, a_{\text{af}}\}$, and the standard deviation of growth rate σ. Although their values were estimated as $\{36\%, 7.11, 1.82, 0.1\}$ from the experimental specifications [8], these values can vary across different runs of experiments due to unavoidable uncertainty and variability in experimental conditions. Therefore, to observe the effects of cell asynchrony in a more realistic scenario, more simulations are conducted where the values of $\{r\%, a_{\text{in}}, a_{\text{af}}, \sigma\}$ are generated by multiplying $\{36\%, 7.11, 1.82, 0.1\}$ with a random sample from the normal distribution $N(1, 0.3^2)$. Specifically, Algorithm 1 is executed one thousand times with randomly selected values of $\{r\%, a_{\text{in}}, a_{\text{af}}, \sigma\}$. Similar patterns can be observed in the average value of $D(\hat{\mu}, \hat{\sigma})$ (see Figure 5) as in the deterministic case discussed earlier in Figure 4.

To Summarize, the 2-D plots of $D(\hat{\mu}, \hat{\sigma})$ presented in Figure 4 and Figure 5 suggest that intrinsic patterns with high expression (smaller value of $\hat{\sigma}$) around the late life stages (larger value of $\hat{\mu}$) are more likely to be affected by cell asynchrony [8]. In the next section, we will further investigate how the effects of cell asynchrony depend on the parameters $\{r\%, a_{\text{in}}, a_{\text{af}}, \sigma\}$, and hence the experimental conditions.

B. Effects under various experimental conditions

In this section, we calculate $D(\hat{\mu}, \hat{\sigma})$ with different parameters $\{r\%, a_{\text{in}}, a_{\text{af}}, \sigma\}$ to simulate how the effects of cell asynchrony depends on experimental conditions.

The parameters are first initialized as $\{36, 7.11, 1.82, 0.1\}$, the estimated values from our earlier study [8]. Then, one of the four parameters is selected and changed to either half or twice of its initial value, to simulate different experimental conditions. For example, Figure 6(a) and 6(b) show $D(\hat{\mu}, \hat{\sigma})$ with parameter $r\%$ changed to 18 and 72 respectively. By comparing these two figures, we can observe how the difference $D(\hat{\mu}, \hat{\sigma})$ is influenced by the value of $r\%$. When the parameter $r\%$ decreases from 72 to 18, $D(\hat{\mu}, \hat{\sigma})$ increases considerably on all intrinsic patterns $f_i(\ell_{\text{re}})|_{\hat{\mu},\hat{\sigma}}$. The same phenomenon is also observed when the parameter σ increases from 0.05 to 0.2, as shown in Figure 6(g) and 6(h) respectively. By contrast, as illustrated in Figure 6(c) and 6(d), $D(\hat{\mu}, \hat{\sigma})$ only increases slightly when a_{in} decreases from 14.22 to 3.56, respectively. Similarly, from Figure 6(e) and 6(h), it can be seen that $D(\hat{\mu}, \hat{\sigma})$ does not vary much when a_{af} increases from 0.91 to 3.64 except for small values of $\hat{\mu}$.

From this analysis, it becomes clear that the parameters $r\%$ and σ have a strong effect on cell asynchrony, whereas the parameters a_{in} and a_{af} have far lesser influence on cell asynchrony. Therefore, when measuring expression levels experimentally, it is crucial to properly measure parameters $r\%$ and σ, either directly or indirectly.

V. Discussions

Here, we discuss how the parameters $\{r\%, a_{\text{in}}, a_{\text{af}}, \sigma\}$ can be manipulated in experimental settings.

Sorbitol treatments are typically conducted in experiments to synchronize the iRBCs at the schizont stage [1]. These schizonts infect fresh RBCs and initialize the experiment. As mentioned earlier on in this paper, $r\%$ of schizonts burst in the infection period, followed by the remaining $1-r\%$ of schizonts which burst after the infection period. Therefore, the burst rate $r\%$ in the infection period is highly correlated with the degree of synchrony of schizonts. In addition to the standard sorbitol treatments, new protocols which combine Percoll and sorbitol treatments have been proposed to achieve tight synchronization of *P. falciparum* [13]. Therefore, higher values of $r\%$ can be achieved if a more homogeneous population of schizonts is used to initialize the experiment.

The infection factors a_{in} and a_{af} respectively denote the average number of RBCs that can be infected by one schizont during and after the infection period, respectively. One schizont usually contains 12 to 16 merozoites, which can further infect RBCs [14]. As mentioned earlier in the model section, cell culture is diluted to reduce the concentration of RBCs in order to avoid further infection after the infection period. Hence, a_{af} is smaller than a_{in}. This suggests that the infection factors a_{in} and a_{af} are proportional to the concentration of RBCs. Therefore, they can be manipulated by changing the concentration of RBCs accordingly.

The standard deviation σ of the normalized life span indicates the diversity of growth rate between individual iRBCs. So far, no experimental technique has been shown to manipulate the value of σ.

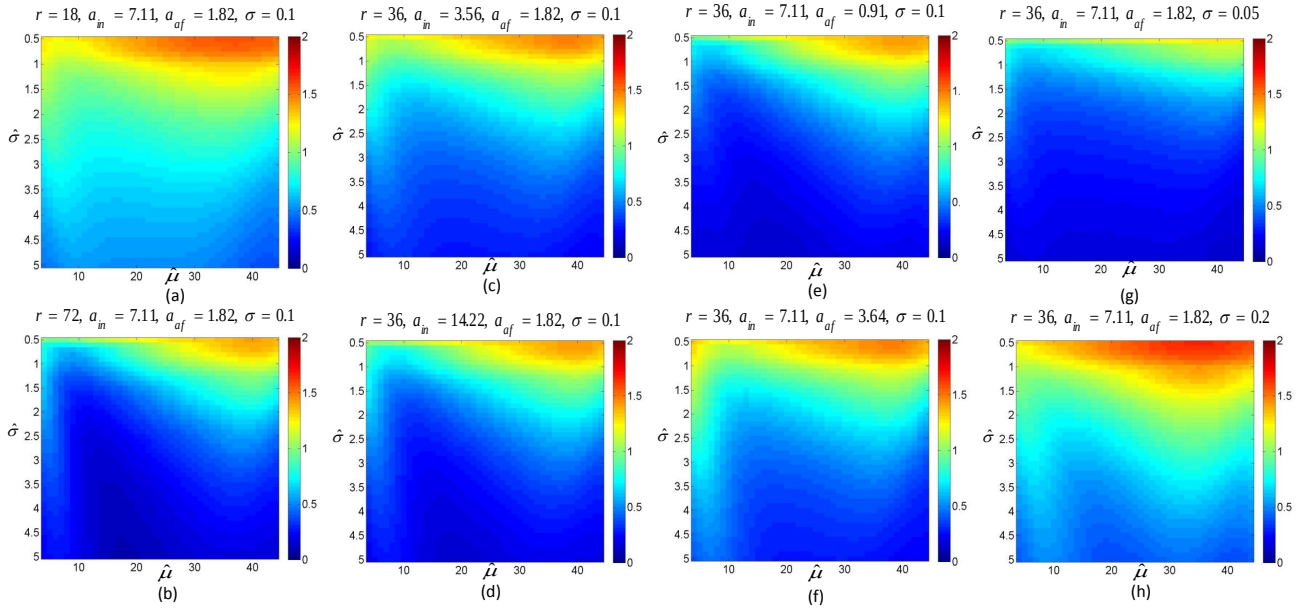

Fig. 6. The 2-D plots of $D(\hat{\mu}, \hat{\sigma})$ calculated with different parameters $\{r\%, a_{\text{in}}, a_{\text{af}}, \sigma\}$. The value of $r\%$ is 18 in (a) and 72 in (b), and 36 in the other figures. Similarly, a_{in} is equal to 3.56 in (c) and 14.22 in (d), and 7.11 in the other figures; a_{af} is equal to 0.91 in (e) and 3.64 in (f), and 1.82 in the other figures; σ is equal to 0.05 (g) and 0.2 (h), and 0.1 in the other figures.

VI. Conclusions

In this paper, we investigated the effects of cell asynchrony on time-series gene expression data of *P. falciparum* by conducting numerical experiments. By analyzing a linear model of cell asynchrony, we demonstrated how cell asynchrony has varying effects on different intrinsic expression patterns, and how these effects are influenced by the experimental conditions (model parameters). The presented analysis may help to gain a better understanding of the effects of cell asynchrony on expression data, and underlines the importance of measuring certain variables, i.e., the burst rate $r\%$ in the infection period, and the standard deviation σ of the growth rate, during experimental measurements of expression levels.

References

[1] Z. Bozdech, M. Llinás, B. Lee, E. D. Wong, J. Zhu, and J. L. DeRisi, "The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium falciparum," *PLoS Biol*, vol. 1, no. 1, pp. e5+, Aug. 2003. [Online]. Available: http://dx.doi.org/10.1371/journal.pbio.0000005

[2] *World Malaria Report*. Geneva, Switzerland: World Health Organization, 2013. [Online]. Available: http://www.who.int

[3] Z. Bozdech, J. Zhu, M. P. Joachimiak, F. E. Cohen, B. Pulliam, and J. L. DeRisi, "Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray." *Genome biology*, vol. 4, no. 2, 2003. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/12620119

[4] M. Llinás, Z. Bozdech, E. D. Wong, A. T. Adai, and J. L. DeRisi, "Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains." *Nucleic acids research*, vol. 34, no. 4, pp. 1166–1173, 2006. [Online]. Available: http://dx.doi.org/10.1093/nar/gkj517

[5] B. Javier, N. Zhang, B. Kaur, S. Kwan, P. Rainer, and Z. Bozdech, "Quantitative time-course profiling of parasite and host cell proteins in the human malaria parasite Plasmodium falciparum." *Molecular & cellular proteomics : MCP*, vol. 10, no. 8, Aug. 2011. [Online]. Available: http://dx.doi.org/10.1074/mcp.m110.006411

[6] W. Trager and J. B. Jensen, "Human malaria parasites in continuous culture." *Science (New York, N.Y.)*, vol. 193, no. 4254, pp. 673–675, Aug. 1976. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/781840

[7] C. Lambros and J. P. Vanderberg, "Synchronization of Plasmodium falciparum erythrocytic stages in culture." *The Journal of parasitology*, vol. 65, no. 3, pp. 418–420, Jun. 1979. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/383936

[8] W. Zhao, J. Dauwels, J. Niles, and J. Cao, "Computational synchronization of microarray data with application to Plasmodium falciparum," *Proteome Science*, vol. 10, no. Suppl 1, pp. S10+, 2012. [Online]. Available: http://dx.doi.org/10.1186/1477-5956-10-s1-s10

[9] Z. Bar-Joseph, S. Farkash, D. K. Gifford, I. Simon, and R. Rosenfeld, "Deconvolving cell cycle expression data with complementary information." *Bioinformatics (Oxford, England)*, vol. 20 Suppl 1, Aug. 2004. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/bth915

[10] D. Siegal-Gaskins, J. N. Ash, and S. Crosson, "Model-Based Deconvolution of Cell Cycle Time-Series Data Reveals Gene Expression Details at High Resolution," *PLoS Comput Biol*, vol. 5, no. 8, pp. e1 000 460+, Aug. 2009. [Online]. Available: http://dx.doi.org/10.1371/journal.pcbi.1000460

[11] W. Zhao, J. Dauwels, and J. Cao, "The Effects of Cell Asynchrony on Time-Series Data: An Analysis on Gene Expression Level of Plasmodium Falciparum," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, vol. 2014. IEEE, Aug. 2014, pp. 5–8. [Online]. Available: http://dx.doi.org/10.1109/embc.2013.6609423

[12] K. G. Le Roch, J. R. Johnson, L. Florens, Y. Zhou, A. Santrosyan, M. Grainger, S. F. Yan, K. C. Williamson, A. A. Holder, D. J. Carucci, J. R. Yates, and E. A. Winzeler, "Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle," *Genome Research*, vol. 14, no. 11, pp. 2308–2318, Nov. 2004. [Online]. Available: http://dx.doi.org/10.1101/gr.2523904

[13] R. A. Childs, J. Miao, C. Gowda, and L. Cui, "An alternative protocol for Plasmodium falciparum culture synchronization and a new method for synchrony confirmation," *Malaria Journal*, vol. 12, no. 1, pp. 386+, Nov. 2013. [Online]. Available: http://dx.doi.org/10.1186/1475-2875-12-386

[14] W. Crewe and D. R. W. Haddock, *Parasites and human disease*. Wiley, 1985. [Online]. Available: http://www.worldcat.org/isbn/9780471010630