Supplementary materials to accompany: "Synchronised oscillations in growing cell populations are explained by demographic noise"

Enrico Gavagnin*[1], Sean T. Vittadello[2], Gency Guanasingh[3], Nikolas K. Haass[3], Matthew J. Simpson[2], Tim Rogers[1], and Christian A. Yates[1]

[1] *Department of Mathematical Sciences, University of Bath, Claverton Down, Bath, UK*

[2] *School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia*

[3] *The University of Queensland, The University of Queensland Diamantina Institute, Brisbane, Queensland, Australia*

This document contains the supplementary materials which accompany the paper Gavagnin et al. [2020]. Sections S.1, S.2 and S.3 contain some details of the mathematical derivation of the analytical formula for the envelope of two standard deviation Q. In Section S.4 we explain the method adopted to parametrise our multi-stage model from the experimental images. Section S.5 contains the computation of the relative entropy between an Erlang and a Gaussian distribution.

S.1 The OU approximation

We can simplify the Langevin model given by equation (8) of the main document by replacing the dependence on $\boldsymbol{x}$ in the correlator of $\boldsymbol{\eta}(t)$ with $\mathbb{E}[\boldsymbol{x}] = K\boldsymbol{u}e^{K\lambda t}$. The resulting equation consists of the high-dimensional non-autonomous Ornstein-Uhlenbeck (OU) process

$$\frac{\mathrm{d}\hat{\boldsymbol{x}}}{\mathrm{d}t} = K\mathcal{S}\hat{\boldsymbol{x}} + K\sqrt{\frac{e^{K\lambda t}}{N_0}}\mathcal{S}\boldsymbol{\psi}(t)\,, \tag{S.1}$$

*Corresponding author: e.gavagnin@bath.ac.uk

where $\boldsymbol{\psi}(t)$ is a K-dimensional white noise vector with correlator $\mathbb{E}[\eta_i(t)\eta_j(t')] = u_i\delta_{ij}\delta(t-t')$. We test the behaviour of the two models, the OU process given by equation (S.1), and the Langevin equation (8) of the main document in Figure S.1. The results suggest that the OU process is an accurate approximation of the Langevin equation, in particular the presence of the oscillations is evident in both the modelling regimes (Figure S.1(a)). In Figure S.1(b), we compare the distributions of $Q(t)$ at times $t = 1, 3$ and 5 obtained by averaging over 1000 independent simulations, which show good agreement between the two models.

S.2 The correlation matrix

For a stochastic initial condition, $\boldsymbol{x}_0$, as described in Section 2.3 of the main document, we can compute the correlation matrix at time $t = 0$, as

$$C_0 = \mathbb{E}\left[x_i(0)x_j(0)\right] = \begin{cases} u_i u_j & \text{for } i \neq j \\ u_i^2 + \frac{u_i}{N_0} & \text{for } i = j \end{cases}. \tag{S.2}$$

We can rewrite this as $C_0 = \boldsymbol{u}\boldsymbol{u}^T + \frac{1}{N_0}M$, where $M = Diag(\boldsymbol{u})$.

We then focus on computing the correlation matrix $C(t,t') = \mathbb{E}\left[\hat{\boldsymbol{x}}(t)\,\hat{\boldsymbol{x}}^T(t')\right]$ for the OU process (S.1), as an approximation for the correlation matrix of the Langevin model. By applying general results for OU processes (See Section 4.5 of [Gardiner, 2009]) we have:

$$C(t,t') = e^{Kt\mathcal{S}}C_0 e^{Kt'\mathcal{S}^T} + \frac{K}{N_0}\int_0^{\min(t,t')} e^{K(t-\tau)\mathcal{S}}\mathcal{S}\left(Me^{K\lambda\tau}\right)\mathcal{S}^T e^{K(t'-\tau)\mathcal{S}^T}\,d\tau\,. \tag{S.3}$$

We can use expression (5) of the main document and the fact that $e^{Kt\mathcal{S}}\boldsymbol{u}\boldsymbol{u}^T e^{Kt'\mathcal{S}^T} = \boldsymbol{u}\boldsymbol{u}^T e^{K(t+t')\lambda}$,

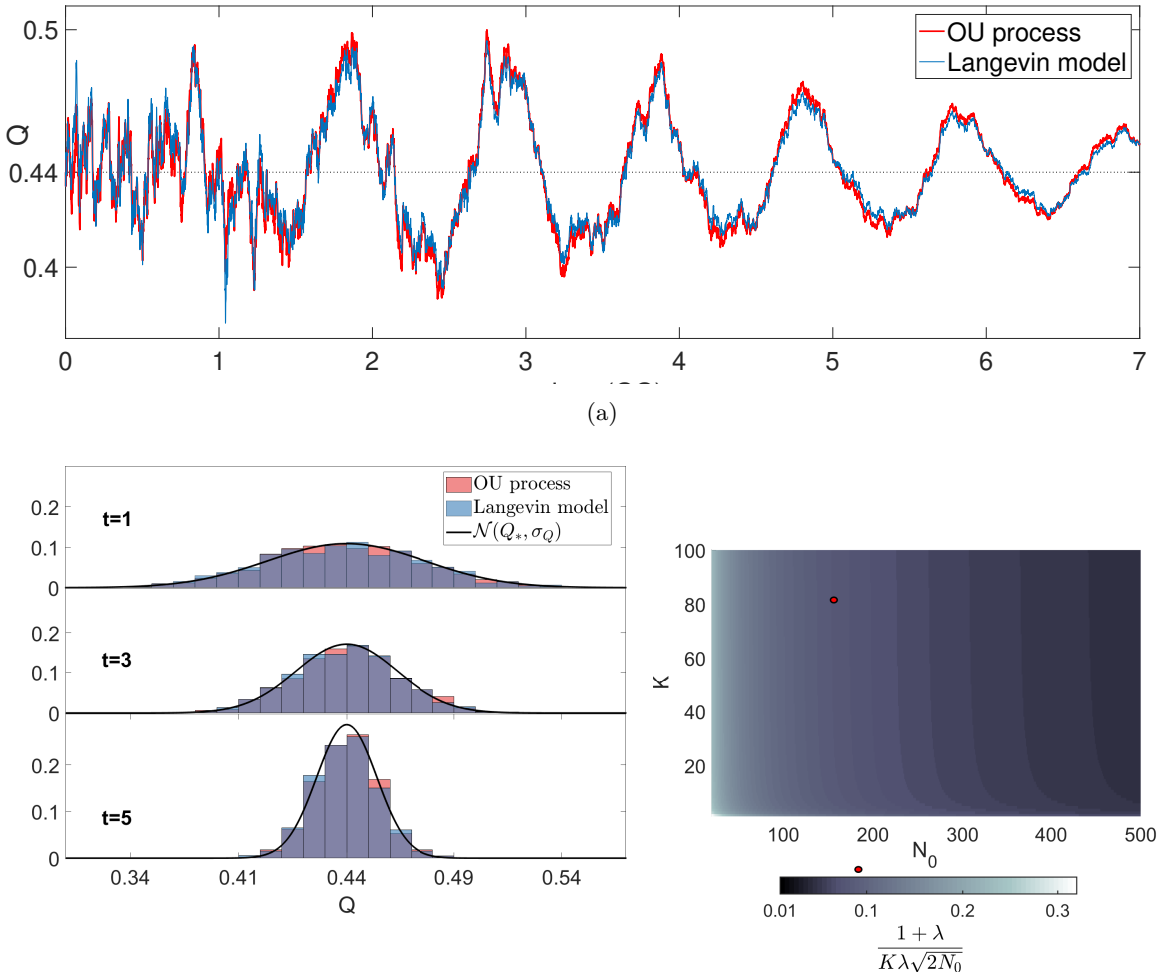

Figure S.1: Comparison of OU process and the Langevin equation model. Panel (a) shows two evolutions of Q for the OU process (red) and the Langevin equation (blue). The two trajectories are realised using Euler-Maruyama method with time step $\Delta t = 10^{-3}$ and the same randomly generated numbers. The parameters are the same as Figure 2 of the paper and time is normalised with reference to the average cell-cycle time. In panel (b) we plot the distribution of Q at three time points ($t = 1$, $t = 3$ and $t = 5$). The two overlaid histograms represent the distributions of 1000 independent simulations of the OU process (red) and the Langevin Equation (blue). The black line represent the distribution $N(Q_*, \sigma_Q(t))$. All the parameters are the same as Figure 2 of the paper. Panel (c) shows the asymptotic value of CV_G and CV_N for a range of parameters choices, N_0 and K. The red dot corresponds to the parameters inferred from the data in Section S.4, indicating that Q will be Gaussian distributed for biologically realistic parameter values.

to write down the (i,j) element of (S.3) for $t < t'$ as

$$\begin{aligned} C_{i,j}(t,t') =& u_i u_j e^{K(t+t')\lambda} + \frac{1}{N_0 K^2} \sum_{k,l,m=1}^{K} \frac{1}{(1+\lambda_k)^{i-m}} \frac{1}{(1+\lambda_l)^{j-m}} \frac{2\lambda}{(1+\lambda)^m} e^{K(t\lambda_k + t'\lambda_l)} \\ &+ \frac{1}{N_0 K} \sum_{k,l,m=1}^{K} \frac{\lambda_k}{(1+\lambda_k)^{i-m}} \frac{\lambda_l}{(1+\lambda_l)^{j-m}} \frac{2\lambda}{(1+\lambda)^m} \int_0^t e^{K[(t-\tau)\lambda_k + (t'-\tau)\lambda_l + \lambda\tau]} . \end{aligned} \tag{S.4}$$

Substituting the expressions (3) of the main text for $\boldsymbol{u}^k$ and $\boldsymbol{v}^k$, using the formula $\sum_{m=1}^{K}(1+\lambda_k)^m(1+\lambda_l)^m/(1+\lambda)^m = (1+\lambda_k)(1+\lambda_l)/[(1+\lambda_k)(1+\lambda_l)-(1+\lambda)]$ and upon rearranging terms, we obtain

$$\begin{aligned} C_{i,j}(t,t') =& \frac{4\lambda^2}{(1+\lambda)^{i+j}} e^{K(t+t')\lambda} + \frac{2\lambda}{N_0 K^2} \sum_{k,l=1}^{K} \frac{(1+\lambda_l)^{1-j}}{(1+\lambda_k)^{i-1}} \frac{1}{(\lambda_k+\lambda_l-\lambda)} \Bigg[e^{K(t\lambda_k + t'\lambda_l)} \\ &- \frac{\lambda_k \lambda_l}{(1+\lambda_k)(1+\lambda_l)-(1+\lambda)} e^{K((t'-t)\lambda_l + t\lambda)} \Bigg] . \end{aligned} \tag{S.5}$$

S.3 The envelope of two standard deviations of Q

We recall the definition of the envelope of two standard deviations of $Q(t)$ as $\Omega(t) = [Q_* - 2\sigma_Q(t), Q_* + 2\sigma_Q(t)]$, where $\sigma_Q(t)$ denotes the standard deviation of $Q(t)$. To compute σ_Q we employ the OU approximation (see Section S.1). From a fixed initial condition, the solutions of (S.1) evolve as a Gaussian process with mean $\bar{\boldsymbol{x}}(t)/N_0$. We can write $G(t) \sim \mathcal{N}(\mu_G(t), \sigma_G(t))$ and $N(t) \sim \mathcal{N}(\mu_N(t), \sigma_N(t))$ where

$$\mu_G(t) = Q_* e^{K\lambda t}, \qquad \sigma_G^2(t) = \sum_{i,j=1}^{\alpha K} C_{i,j}(t,t) - Q_*^2 e^{2K\lambda t}, \tag{S.6a}$$

$$\mu_N(t) = e^{K\lambda t}, \qquad \sigma_N^2(t) = \sum_{i,j=1}^{K} C_{i,j}(t,t) - e^{2K\lambda t} . \tag{S.6b}$$

Notice that, Q is defined as a ratio between two Gaussian distribution and, in general, this does not imply that $Q(t)$ is Gaussian. However, Hayya et al. [1975] showed that the ratio of two

Gaussian can be well approximated as a Gaussian, under certain conditions on the *coefficient of variation* (CV) of the numerator and denominator. Precisely, provided that

$$CV_N = \frac{\sigma_N}{\mu_N} < 0.39 \qquad \text{and} \qquad CV_{G1} = \frac{\sigma_{G1}}{\mu_{G1}} > 0.005 \tag{S.7}$$

Hayya et al. [1975] demonstrate that Q is close to a Gaussian distribution. Moreover, we can approximate the variance of Q by Taylor expanding to the second order which leads to

$$\begin{aligned} \sigma_Q^2 &\approx \sigma_N^2 \frac{\mu_G^2}{\mu_N^4} + \frac{\sigma_G^2}{\mu_N^2} - 2\rho\mu_G \frac{\sigma_N \sigma_G}{\mu_N^3} \\ &= \frac{1}{\mu_N^2} \left[\sigma_N^2 Q_*^2 + \sigma_G^2 - 2\sigma_N \sigma_G Q_* \, \rho\left[G, N\right] \right] , \end{aligned} \tag{S.8}$$

where ρ denotes the correlation coefficient, defined as

$$\rho\left[Y_1, Y_2\right] = \frac{\mathbb{E}\left[Y_1 Y_2\right] - \mathbb{E}\left[Y_1\right] \mathbb{E}\left[Y_2\right]}{\sqrt{\operatorname{Var}\left[Y_1\right] \ \operatorname{Var}\left[Y_2\right]}} . \tag{S.9}$$

Notice that we can compute $\mathbb{E}\left[G(t)N(t)\right]$ in equation (S.9) in terms of the correlation matrix C as

$$\mathbb{E}\left[G(t)N(t)\right] = \sum_{i=1}^{K} \sum_{j=1}^{\alpha K} C_{i,j}(t,t) .$$

We now need to check that the conditions (S.7) are satisfied for biologically relevant parameter choices. By studying the expressions (S.5) and (S.6), we obtain that $CV_N(0) = 1/\sqrt{N_0}$ and $CV_G(0) = 1/\sqrt{Q_* N_0}$ which satisfy the conditions (S.7) for $\alpha \in [0, 1]$ and $N_0 \in [10, 10^4]$. In order to check the validity of the conditions in the long-term, we look at the leading terms of the expression (S.5). We find that

$$\lim_{t \to +\infty} CV_N(t) = \lim_{t \to +\infty} CV_G(t) = \frac{1+\lambda}{K\lambda\sqrt{2N_0}} . \tag{S.10}$$

In Figure. S.1(c) we evaluated this expression for any $K \in [1, 100]$ and $N_0 \in [20, 500]$. Our findings show that the limit of CV_G and CV_N for $t \to +\infty$ lies in the interval $(0.01, 0.32)$ for the range of parameters considered which suggests that the conditions (S.7) are satisfied for biologically relevant choices of the parameters. Notice that the plots in Figure S.1(b) provide further confirmation of this by showing good agreement between the distribution of Q and the Gaussian distribution $\mathcal{N}(Q_*, \sigma_Q(t))$.

S.4 Parameter inference

To infer the parameters of the multi-stage model, we simultaneously fit the distribution of the total cell-cycle time and of the G1 duration of 200 randomly selected cells.

Let $\boldsymbol{H}_T$ and $\boldsymbol{H}_{G1}$ denote the histogram representations of the probability density function (pdf) of the total cell-cycle time and the G1 duration, respectively, with a bin width of one hour. For example, $(\boldsymbol{H}_T)_i$ denotes the proportion of cells with a cell-cycle time in the interval $[ih, (i+1)h)$. We denote with $\boldsymbol{H}_{E(K,\beta)}$ the histogram obtained by discretising an Erlang distribution with parameters (K, β) with the same bin width, *i.e.* $(\boldsymbol{H}_{E(K,\beta)})_i = \frac{\beta^K}{(K-1)!} \int_i^{i+1} x^{K-1} e^{-\beta x} dx$.

For a given combination of parameters, (K, β, α), one can consider the statistic

$$I(K, \beta, \alpha) = \|\boldsymbol{H}_T - \boldsymbol{H}_{E(K,\beta)}\|_1 + c\,\|\boldsymbol{H}_{G1} - \boldsymbol{H}_{E(\alpha K,\beta)}\|_1\,, \tag{S.11}$$

where $c > 0$ is a constant and $\| - \|_1$ denotes the 1-norm. Notice that the constant c can be interpreted as a weight to give more ($c > 1$) or less ($c < 1$) priority at the fitting of the G1 distribution compared to the one of the total cell-cycle time distribution. For simplicity we choose $c = 1$, which corresponds to equal levels of priority for the two distribution fits.

To determine the parameter combination which provides the best simultaneous fit of the two distribution, we evaluated the function I in the parameter range $K \in [10, 150]$ $\beta \in [1, 10]$ and $\alpha \in [0, 1]$. We find that the combination $K^* = 92$, $\beta^* = 4.96$ and $\alpha^* = 33/92$ minimises the

statistic I in the parameter region considered and, hence, we select these parameters for the multi-stage model.

To infer the average population size at the moment of the initial sampling, N_0, we first measure the average population size at the beginning of the recording, $N_{24} = 381.1$, averaged over the 30 experiments. Since the average population size grows exponentially at rate λ, we project back from time $t = 24h$ and we obtain the average sample size as $N_0 = N_{24} \exp(-24\lambda) \approx 155$.

S.5 The Kullback Leibler divergence between Erlang and Gaussian distribution

We compute the relative entropy (Kullback-Leibler divergence, D_{KL}) between an Erlang and a Gaussian distribution as a measure of the distance between the two distributions.

For two distributions, $p(x)$ and $q(x)$, the KL divergence is defined as:

$$D(p,q) = \int_{-\infty}^{\infty} p(x) \log \left[\frac{p(x)}{q(x)} \right] dx \,. \tag{S.12}$$

We set $p(x)$ to be the probability density function (pdf) of an $Erlang(K, \beta)$ and $q(x)$ to be the pdf of a Gaussian with same mean and variance, *i.e.* $\mathcal{N}\left(\frac{K}{\beta}, \frac{K}{\beta^2}\right)$. We then obtain

$$
\begin{aligned}
D(K,\beta) =& D\left(p\left(K,\beta\right), q\left(\frac{K}{\beta}, \frac{K}{\beta^2}\right)\right) \\
=& \frac{\beta^K}{(K-1)!}\int_0^\infty x^{K-1}e^{-\beta x}\log\left[\frac{\beta^{K-1}\sqrt{2\pi K}}{(K-1)!}x^{K-1}e^{\frac{\beta^2}{2K}\left(x-\frac{K}{\beta}\right)^2-\beta x}\right]dx \\
=& \log\left[\frac{\beta^{K-1}\sqrt{2\pi K}}{(K-1)!}\right]\frac{\beta^K}{(K-1)!}\int_0^\infty x^{K-1}e^{-\beta x}dx \\
&+ (K-1)\frac{\beta^K}{(K-1)!}\int_0^\infty x^{K-1}\log(x)e^{-\beta x}dx \\
&+ \frac{\beta^K}{(K-1)!}\int_0^\infty\left[-\beta x + \frac{\beta^2}{2K}\left(x-\frac{K}{\beta}\right)^2\right]x^{K-1}e^{-\beta x}dx\,.
\end{aligned}
\tag{S.13}
$$

Notice that the first integral of Equation (S.13) is exactly the pdf of an Erlang, which simplifies to unity. The second and third integral in (S.13) require more work. By using integration by parts and upon simplification, we get to the final expression

$$
D(K,\beta) = \log\left[\frac{\beta^{K-1}\sqrt{2\pi K}}{(K-1)!}\right] + (K-1)\left(\mathcal{H}_{K-1} - \log(\beta) - \gamma\right) - K + \frac{1}{2}\,, \tag{S.14}
$$

where $\mathcal{H}_{K-1} = \sum_{i=1}^{K-1}\frac{1}{i}$ is the $(K-1)$-th harmonic number and $\gamma = \lim_{n\to+\infty}\left(\sum_{i=1}^{n}\frac{1}{i} - \log(n)\right)$ denotes the Euler-Mascheroni constant.

Using the expression (S.14) it is possible to show that $D(K,\beta)$ is a decreasing function of K and $D(K,\beta) \sim \mathcal{O}(K^{-1})$ for $K \to +\infty$. This is not a surprise, since by central limit theorem we know that the Erlang distribution converges to a Gaussian with same mean and variance. Since the CV of an Erlang(K,β) is given by $K^{-\frac{1}{2}}$, we can rephrase by saying that the KL divergence scales proportionally to the square of the CV of the Erlang distribution.

Figure S.5 shows the plot of $D(K,\beta)$ with $\beta = K$ for different values of K. In the overlaid panels the two distributions are compared for $K = 5, 10, 20, 30$ and 60. The results highlight the good level of similarity between the Erlang and Gaussian distributions for large K - small values

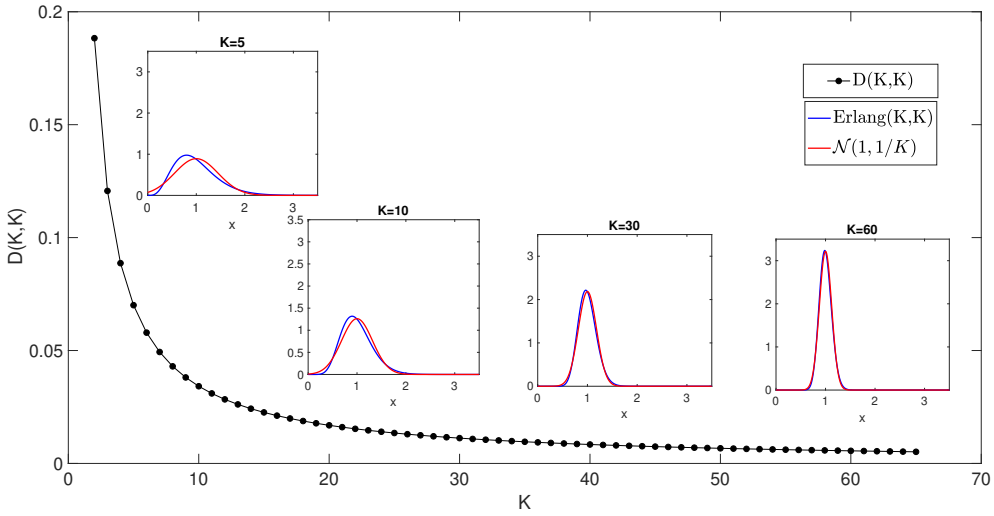

Figure S.5: The Kullback Leibler (KL) divergence between an Erlang distribution and Gaussian distribution. The black dotted line in the main panel shows the KL divergence between an Erlang distribution of parameters (K, K) and a Gaussian distribution of parameters $(1, 1/K)$ as function of K. The four overlaid panels show the comparison of the two distributions, Erlang (blue) and Gaussian (Red), for $K = 5, 10, 30$ and 60 (from left to right).

of the CV. For example, for $K > 25$, *i.e.* CV< 0.2, we have $D(K, K) < 0.02$ which corresponds to good agreement between the two distributions.

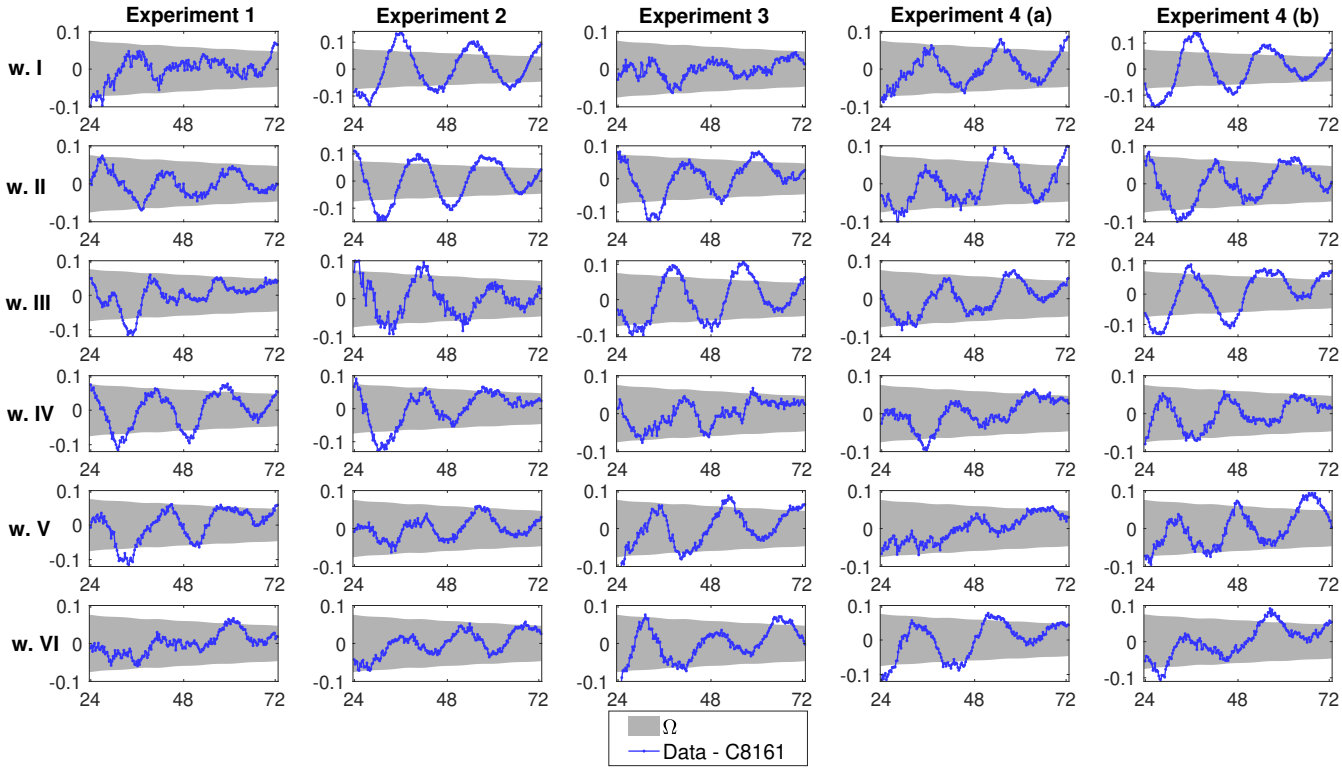

Figure S.5: Comparison 30 time series obtained from the data (blue lines), together with the envelope of two standard deviations, Ω (light grey regions) predicted using the multi-stage model. The parameters of the multi-stage models are obtain by fitting the distribution of the total cell-cycle time and G1 duration (see Section S.4): $K = 92, \alpha K = 33, \beta = 4.96h^{-1}$ and $N_0 = 155$.

References

C. Gardiner. *Stochastic methods*, volume 4. Springer Berlin, 2009.

E. Gavagnin, S.T Vittadello, G. Gunasingh, N.K Haass, M.J. Simpson, T. Rogers, and C.A. Yates. Synchronised oscillations in growing cell populations are explained by demographic noise. *t.b.a.*, 2020.

J. Hayya, D. Armstrong, and N. Gressis. A note on the ratio of two normally distributed variables. *Manag. Sci.*, 21(11):1338–1341, 1975.