Dynamics of double-strand break and repair in growing populations

Yeast DNA repair group

Contents

1 Introduction

2 Analysis of pooled data reveals time-scales of break and repair

2.1 Model

We begin by ignoring the single-cell resolution afforded to us by the individual wells, and treat each condition as a bulk experiment by studying the number of bright field and GFP cells averaged over all the wells. The trajectories of the average number of bright field and GFP cells are shown in Figure 1 (A). We compare this data to a simple ODE mode in which an initial population of cells with modified DNA grows at a rate α and switches into a non-growing, broken state, at a rate β. The broken cells can then become repaired at a rate ρ and once again begin to grow at a rate α. Letting m, b and g be the number of modified, broken and repaired (or green) cells, we have

$$\frac{d}{dt}m = (\alpha - \beta)m \tag{1}$$

$$\frac{d}{dt}b = \beta m - \rho b \tag{2}$$

$$\frac{d}{dt}g = \alpha g + \rho b. \tag{3}$$

It is easy to see that the total population size $n = m + b + g$ will eventually grow exponentially at a rate α. We can therefore obtain α by fitting the bright field data to an exponential. We can also see from these equations that m and b will eventually grow (assuming $\alpha > \beta$) exponentially at a rate $\alpha - \beta$, while g, and hence the total number of cells, will eventually grows at a rate α. Therefore, a prediction of this model is that the fraction of non-GFP cells, $\phi_{m+b} = (b + m)/(m + b + g)$ decays exponentially at a rate β:

$$\phi_{m+b} \sim e^{-\beta t}. \tag{4}$$

Note that this result is independent of whether we impose a carrying capacity on the population, the ratios will remain unchanged even if the population is not growing exponentially. Fitting $\ln \phi_{m+b}$ to a line thus gives us a way to infer the break rate which (at least in the long-term) is independent of ρ. This exponential

decay is seen in some, but not all experiments; see Figure 1 (B). In CTG-Cpf1 and CGG-Cpf1 the green cells appear very slowly, making is difficult to separate the long term decay from the the transient dynamics. In GAA-SpCas9 and CTG-SpCas9 we don't see the number of GFP cells converge to the number of bright field cells, resulting an a biphasic trajectory of ϕ_{m+b}.

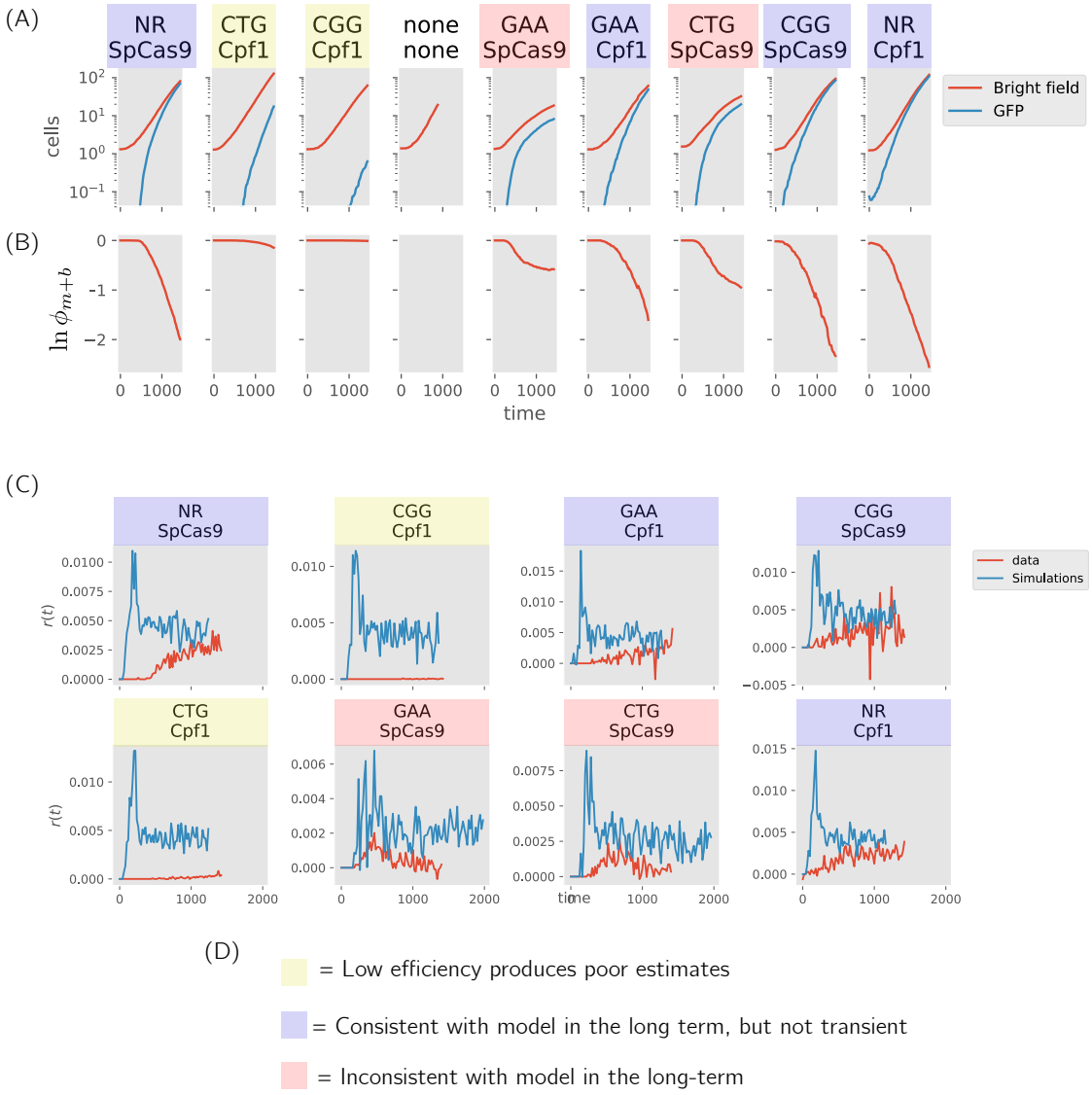

Figure 1: (A) The average number of bright field and GFP cells as a function of time in each condition. (B) The fraction of non-GFP cells as function of time for each condition. (C) The rate to turn green obtained from Equation 6 compared to simulations of the model without a broken state using values of inferred from the trajectories in (B). Classification of experiments in terms of comparison between predictions from (B) and (C).

2.2 Trajectories of rate to turn green

An alternative approach to inferring β is to look at the instantaneous production of green cells. Solving for β in Equation 1, we get

$$\begin{aligned}\beta &= \frac{1}{m}\left[\frac{d}{dt}b + \rho b\right] = \frac{1}{m}\left[\frac{d}{dt}b + \frac{d}{dt}g - \alpha g\right] \\ &= \frac{1}{n-g-b}\left[\frac{d}{dt}b + \frac{d}{dt}g - \frac{g}{n-b}\frac{d}{dt}n\right]\end{aligned} \tag{5}$$

Since we don't know the number of broken and not yet repaired cells, we can't compute all the terms in this equation directly. Intuitively, if $\rho \gg \beta$, we can neglect the broken cells, leading to the approximation

$$\beta \approx r(t) \equiv \frac{1}{n-g}\left[\frac{d}{dt}g - \frac{g}{n}\frac{d}{dt}n\right]. \tag{6}$$

Within the model, we can obtain an exact formula for $r(t)$ for finite ρ by plugging the solution of Equation 1 into Equation 6 (see Appendix A). We find that if $\beta < \alpha$, $r(t) \to \beta$ in the long-time limit, regardless of ρ.

In Figure 1 (C) we compare $r(t)$ in the data and simulations of the model with no broken cells (meaning $\rho \to \infty$). We first focus on the experiments for which ϕ_{m+b} exhibits clear exponential decay. For these experiments, the simulations and data appear to be converging to similar values of $r(t)$, although the transient dynamics different significantly. Since the convergence of $r(t)$ to β depends on ρ, we check whether the trajectories can be reproduced by simulations of the model with finite ρ. Selecting values of ρ on the order of β, we find that the qualitative features of the trajectories are reproduced; see Figure 3 (B). However, this does not explain the "bump" in the GAA-SpCas9 and CTG-SpCas9. To reproduce this, we assume that some fraction of the repaired cells fail to turn green, but still grow. A simulation of this model is shown in Figure 3 (B), where we can see that this will indeed produce the behavior observed in some experiments.

We now consider that there may be a lag between the repair of broken cells and the expression of GFP. The only difference between the lag period and the broken state is that cells may still grow when during the lag period. Therefore, it could be that the lag period rather than the (more biologically relevant) broken state explains the slow increase in $r(t)$. Letting r denote the cells that are repaired and growing but non-green and $1/\gamma$ be the time-scale between repair and GFP production, we have the expanded system

$$\frac{d}{dt}m = (\alpha - \beta)m \tag{7}$$

$$\frac{d}{dt}b = \beta m - \rho b \tag{8}$$

$$\frac{d}{dt}r = \alpha r + \rho b - \gamma r \tag{9}$$

$$\frac{d}{dt}g = \alpha g + \gamma r. \tag{10}$$

As with the original model, these ODEs can be solved exactly to obtain an expression for $r(t)$. In the large ρ limit, we find

$$r(t) = \frac{\beta\gamma\left(e^{\beta t} - e^{\gamma t}\right)}{\beta e^{\beta t} - \gamma e^{\gamma t}} \tag{11}$$

which converges to $\min(\gamma, \beta)$. Thus the long-term behavior of $r(t)$ shown in Figures 1 and 1 is not inconsistent with values of γ is close to or greater than β. By simulating the stochastic model with no broken state, but values of γ which are close to β, we find that a lag between repair and GFP production provides an alternative explanation of our previous observations.

3 Inference from ODE model predicts well-to-well variability

We now consider the variability between wells within a condition. In order to remove the effects of lag time and asynchronous growth, we consider the variance of G conditioned on the total number of cells, $\mathrm{var}(G|N)$.

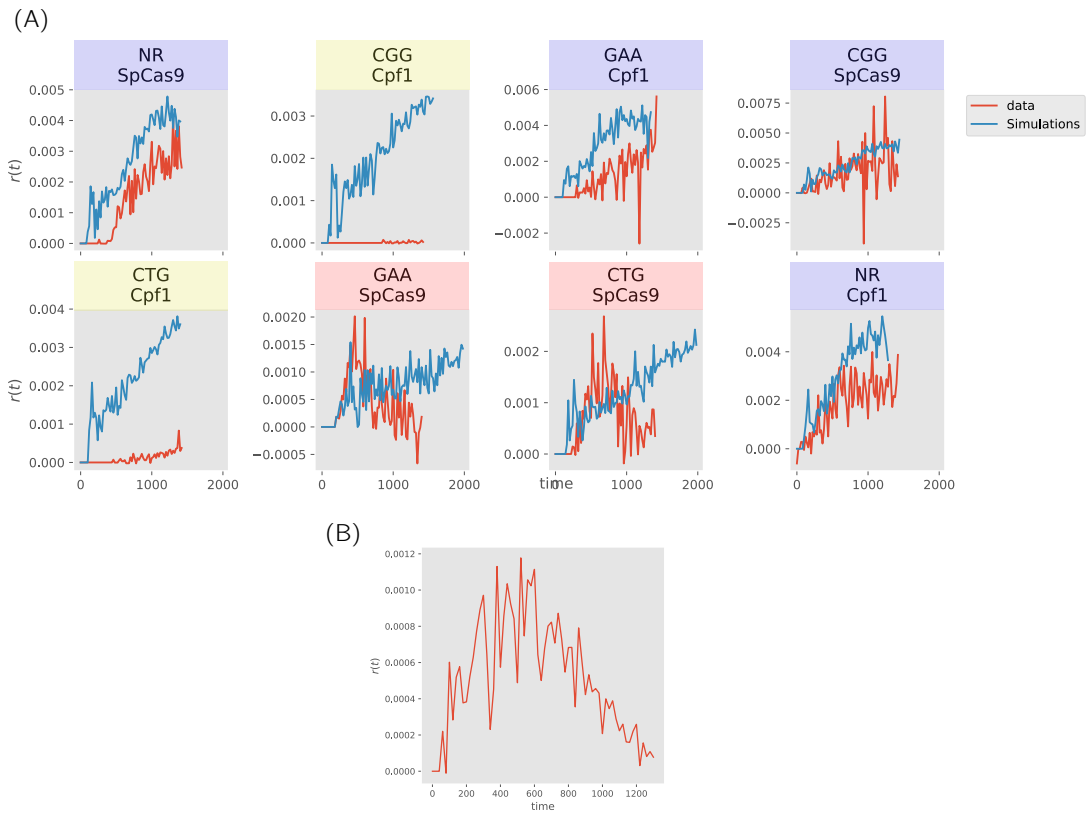

Figure 2: (A) The rate to turn green obtained from Equation 6 compared to simulations of the model with a broken with the values of α and β inferred from the trajectories in Figure 1 (B) and values of ρ comparable to β. The values of β found to give agreement between the simulations are data are $\rho = \beta/5$. (B) A simulation of the model with 1/2 of the repaired cells failing to turn green.

In a simple model where we neglect the broken, non-green state, $\mathrm{var}(G|N)$ undergoes a phase transition at $p_c = 1/2$ (see Section C for derivation). When $p < 1/2$, $\mathrm{var}(G|N)$ grows exponentially, while for $p < 1/2$ it decays. In all the experiments, we find that $\mathrm{var}(G|N)$ grows exponentially. In general, the growth of $\mathrm{var}(G|N)$ is given by the formula

$$\mathrm{var}(G_t|N) = \begin{cases} \frac{p}{1-2p} N^{2+2\ln(1-p)/\ln(2)} & \text{if } p < 1/2 \\ \frac{p}{2p-1} N^{1+1\ln(1-p)/\ln(2)} & \text{if } p > 1/2 \end{cases} \tag{12}$$

This formula seems to be fairly accurate for a more realistic asynchronous model where cells have generation times drawn from a distribution. Interestingly, the inclusion of a broken state does not affect the long-term behavior of $\mathrm{var}(G_t|N_t)$; see Section C. In Figure 4, we compare the fluctuations in the data to those in simulations of the the random generation time model [NOTE: Need to define this somewhere] using parameter values computed from Figure 1 as described in the previous section. We simulate the model both with and without the broken state and see that the broken state model does a slightly better job of predicting the growth of the fluctuations, suggesting the broken state is indeed important. However, there are still some inconsistencies between the simulations and the data; namely, the decrease in $\mathrm{var}(G|N)$ at the end of the experiment.

To-do:

- Test if lag between repair and GFP production (but no broken state) can also reproduce qualitative features of data
- Find more quantitive way to infer ρ (currently just tried ρ close to β). Fitting the curves doesn't work so well due lag time.

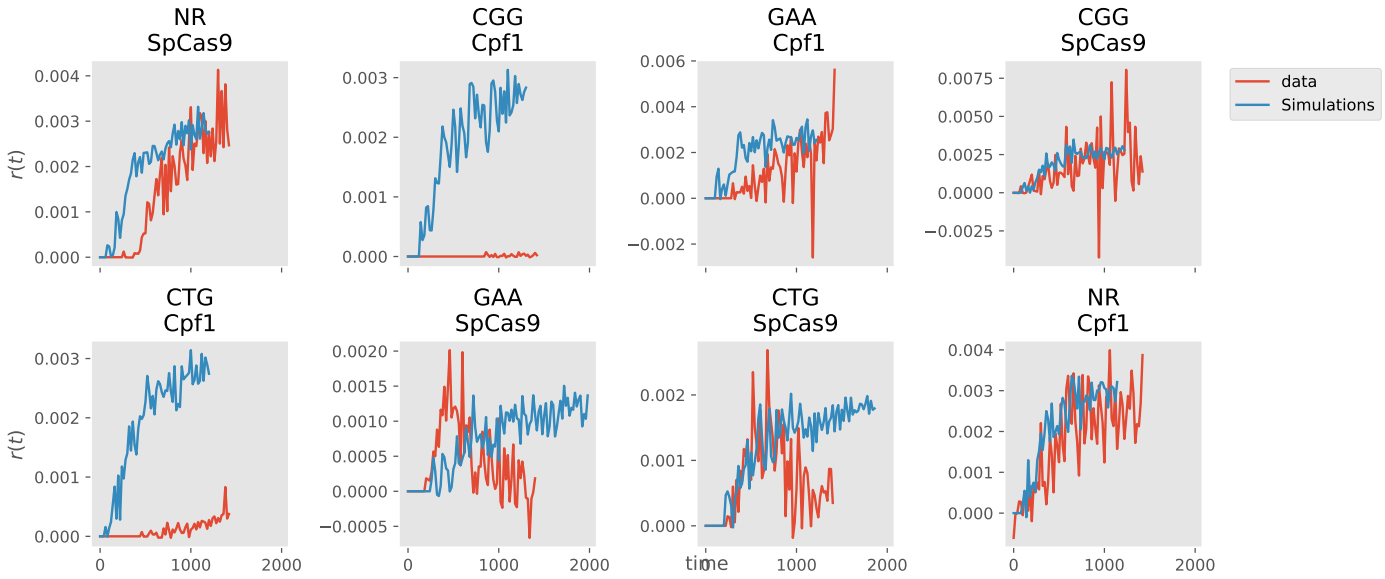

Figure 3: The rate to turn green obtained from Equation 6 compared to simulations of the model with a lag between repair and GFP production. The values of the rate for GFP to be expressed after repair (γ) are $\gamma = 0.8\beta$.

- Plot variance in green cells as a function of time in data vs. simulations
- Show distribution of simulations

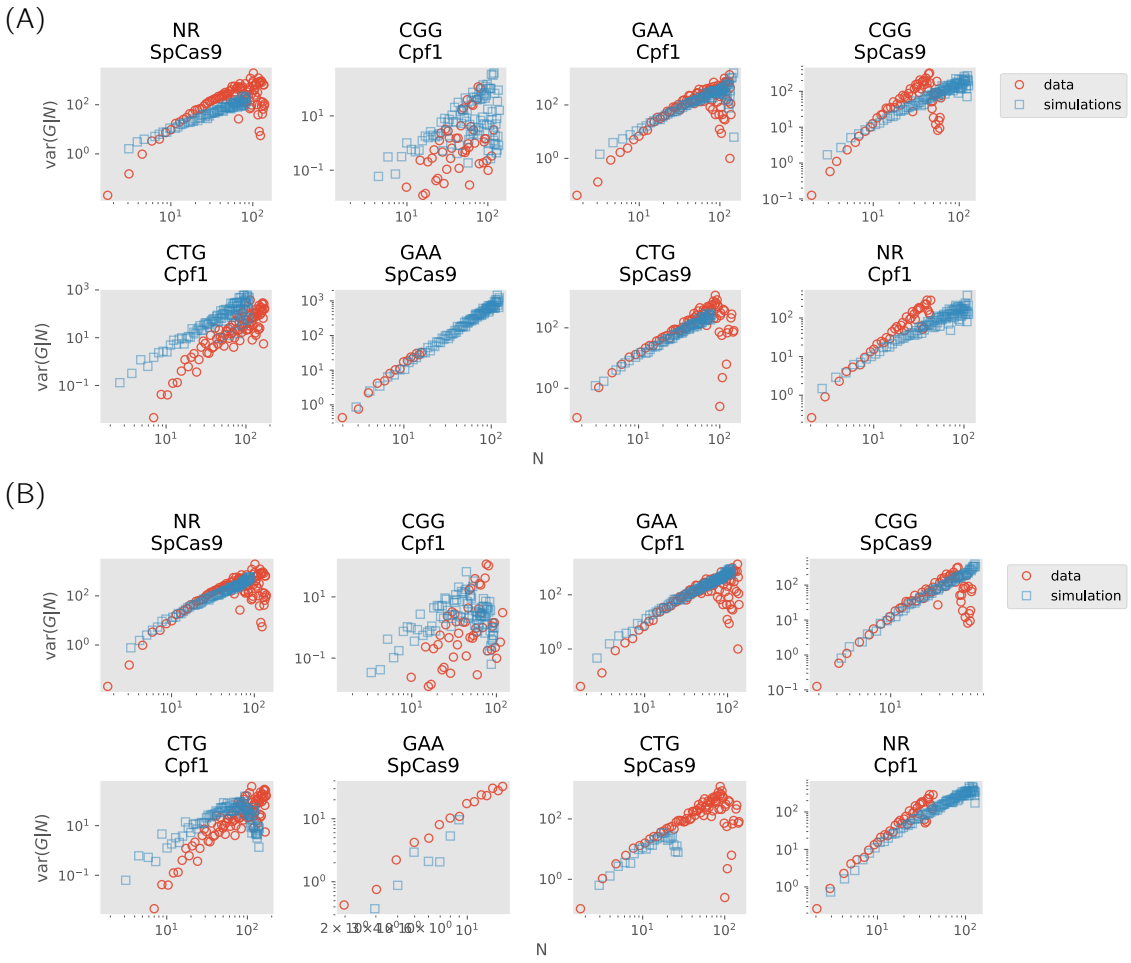

Figure 4: (A) var$(G|N)$ compared to simulations of the constant probability model with parameters computed from the previous section. (B) The same as (A) but including a broken state, with the same parameter ρ used in the simulations in Figure 3 (A).

A Solution of ODE model

The ODEs given in Equation 1 can be solved explicitly as follows. First note that the equation for m is decoupled from the other three, and has the solution

$$m(t) = m(0)e^{t(\alpha-\beta)}. \tag{S.13}$$

The equation for $b(t)$ can then be solved using an integrating factor. With $b(0) = 0$, this leads to

$$b(t) = \frac{\beta m(0)\left(e^{t(\alpha-\beta)} - e^{-\rho t}\right)}{\alpha - \beta + \rho}. \tag{S.14}$$

Doing the same for $g(t)$ yields

$$g(t) = \frac{m(0)\rho\left(\beta e^{-\rho t} - (\alpha+\rho)e^{t(\alpha-\beta)} + (\alpha-\beta+\rho)e^{t\alpha}\right)}{(\alpha+\rho)(\alpha-\beta+\rho)}. \tag{S.15}$$

Plugging the solutions for $b(t)$, $m(t)$ and $g(t)$ into $r(t)$ yields (after some simplification)

$$r(t) = \frac{\beta\rho^2 e^{t(\beta+\rho)}\left(\alpha\beta e^{t(\alpha+\beta+\rho)} + \beta\rho e^{t(2\alpha+\rho)} - \beta(\alpha+\rho)e^{t(\alpha+\beta)} + (\alpha-\beta)(\alpha+\rho)e^{t(\alpha+\rho)} - \alpha e^{\alpha t}(\alpha-\beta+\rho) + \beta\rho e^{\beta t}\right)}{\left(\rho e^{t(\alpha+\rho)} + \beta e^{t(\beta+\rho)} + \beta\left(-e^{\beta t}\right)\right)\left(\alpha^2\beta\left(-e^{\beta t}\right) + \beta\rho^2 e^{t(\alpha+\beta+\rho)} - \beta(\rho-\alpha)(\alpha+\rho)e^{t(\beta+\rho)} + \alpha\rho(\alpha+\rho)e^{t(\alpha+\rho)}\right)}. \tag{S.16}$$

We can obtain the large time limit by looking at the fastest growing terms in the numerator and denominator. If $\alpha > \beta$ (which is true for all experiments) the fastest growing terms grow exponentially with rates $2\alpha+\beta+2\rho$ and

$$r_\infty \equiv \lim_{t\to\infty} r(t) = \frac{\beta^2\rho^3}{\beta\rho^3} = \beta \tag{S.17}$$

regardless of ρ.

B Analysis of asynchronous model

Here consider a stochastic model where divisions are asynchronous; that is, cells have different generation times. For simplicity, we will consider the case where generation times are independent and exponentially distributed. Although not biologically realistic, it captures the essential qualitative features of the dynamics, namely the growth of variance with time. We consider the probability $P(n, g, t)$ of observing n cells, g of which are green at time t and suppose that cells divide at a constant rate α, while non-green cells turn green at a rate β. Under these assumptions $P(n, g, t)$ obeys the master equation

$$\begin{aligned}\frac{d}{dt}P(n,g,t) = {}& \alpha(n-g-1)P(n-1,g,t) + \alpha(m-1)P(n-1,g-1,t) \\ & + \beta(n-g+1)p(n,g-1,t) - [\alpha n + \beta(n-g)]\,P(n,g,t).\end{aligned} \tag{S.18}$$

For the averages, we have

$$\frac{d}{dt}\langle N\rangle = \alpha\langle N\rangle \tag{S.19}$$

$$\frac{d}{dt}\langle G\rangle = \alpha\langle G\rangle + \beta(\langle N\rangle - \langle G\rangle) \tag{S.20}$$

which implies $\langle N\rangle \sim e^{\alpha t}$ and

$$\langle G\rangle = e^{(\alpha-\beta)t}\left[e^{\beta t} - 1\right]. \tag{S.21}$$

We now consider the second moment:

$$\frac{d}{dt}\langle G^2\rangle = \sum_{g,n} g^2 \frac{d}{dt}P(n,g,t) \tag{S.22}$$

$$= \sum_n \sum_g \alpha(n-g)g^2 P(n,g,t) + \sum_n \sum_g \alpha g(g+1)^2 P(n,g,t) \tag{S.23}$$

$$+ \sum_g \sum_g \beta(n-g)(g+1)^2 P(n,g,t) - \sum_n \sum_g [\alpha n + \beta(n-g)]\, g^2 P(n,g,t) \tag{S.24}$$

$$= \alpha\langle NG^2\rangle - \alpha\langle G^3\rangle + \alpha\langle G^3\rangle + 2\alpha\langle G^2\rangle + \alpha\langle G\rangle + \beta\langle NG^2\rangle - \beta\langle G^3\rangle \tag{S.25}$$

$$+ 2\beta\langle NG\rangle - 2\beta\langle G^2\rangle + \beta\langle N\rangle - \beta\langle G\rangle - \langle NG^2\rangle - \beta\langle NG^2\rangle + \beta\langle G^3\rangle. \tag{S.26}$$

After cancelling terms, we find

$$\frac{d}{dt}\langle G^2\rangle = 2\alpha\langle G^2\rangle + \alpha\langle G\rangle + 2\beta\langle NG\rangle \tag{S.27}$$

$$- 2\beta\langle G^2\rangle + \beta\langle N\rangle - \beta\langle G\rangle. \tag{S.28}$$

For the variance

$$\frac{d}{dt}\mathrm{var}(G) = \frac{d}{dt}\left[\langle G^2\rangle - \langle G\rangle^2\right] \tag{S.29}$$

$$= \frac{d}{dt}\langle G^2\rangle - 2\langle G\rangle\frac{d}{dt}\langle G\rangle \tag{S.30}$$

$$= 2(\alpha-\beta)\mathrm{var}(G) + 2\beta\mathrm{cov}(G,N) + (\alpha-\beta)\langle G\rangle - \langle N\rangle. \tag{S.31}$$

We need to calculate the dynamics of the covariance to obtain a closed set of equations. Proceeding in a similar manner as we did for the variance, we start with the product:

$$\frac{d}{dt}\langle GN\rangle = \alpha\langle G\rangle + 2\alpha\langle GN\rangle - \beta\langle GN\rangle + \beta\langle N^2\rangle. \tag{S.32}$$

Subtracting the equations for the means gives

$$\frac{d}{dt}\mathrm{cov}(G,N) = \frac{d}{dt}\langle GN\rangle - \langle G\rangle\frac{d}{dt}\langle N\rangle - \langle N\rangle\frac{d}{dt}\langle G\rangle \tag{S.33}$$

$$= \beta\mathrm{var}(N) + (2\alpha-\beta)\mathrm{cov}(G,N) + \alpha\langle G\rangle. \tag{S.34}$$

Since $\mathrm{var}(N)\sim e^{2\alpha t}$, we conclude that $\mathrm{cov}(N,G)\sim e^{2\alpha t}$ and therefore,

$$\mathrm{var}(G)\sim e^{2\alpha t} \tag{S.35}$$

for any value of β.

C Analysis of synchronous model

Here we consider a class of synchronous model in which all growing cells have a fixed generation time.

C.1 Two state model

We start by ignoring any intermediate state the breaking of DNA and the appearance of growing GFP cells. In this case, the total number of cells at the tth generation is simply 2^t. Letting G_t be the number of green cells at time t, we have

$$G_t = 2G_t + 2\sum_{j=1}^{2^{t-1}-G_{t-1}}\xi_{t,j} \tag{S.36}$$

where $\xi_{t,j}$ are independent Bernoulli random variables with probability of success p (the probability to turn green in a given generation). The first term represents the division of existing green cells, while the second represents the fact that each of the $2^{t-1}-G_{t-1}$ non-green cells in the previous generation has a probability p to turn green before dividing. The second term has a binomial distribution with mean $p(2^t-2G_{t-1})$ and variance $p(1-p)(2^t-2G_{t-1})$ and can therefore be approximated by a normal distribution, allowing us to rewrite Equation S.36 as

$$G_t \approx 2(1-p)G_{t-1} + p2^t + \sqrt{p(1-p)(2^t-2G_{t-1})}\eta_t \tag{S.37}$$

where η_t are independent standard normal random variates. For the mean (this can actually be obtained directly from Equation S.36 and is exact), we have

$$\langle G_t\rangle = 2(1-p)\langle G_{t-1}\rangle + p2^t = 2^t\sum_{j\le}(1-p)^{t-j}p \tag{S.38}$$

$$= 2^t\left[1-(1-p)^t\right] \tag{S.39}$$

From this equation, we can deduce that the fraction of green cells approaches one for all values of p. For the variance, we make the approximation that G_{t-1} appearing in the square root an be replaced by $\langle G_{t-1}\rangle$ mean, which gives

$$G_t \approx 2(1-p)G_{t-1} + p2^t + \sqrt{2^t p(1-p)^t}\eta_t. \tag{S.40}$$

Taking the variance of both sides, we have

$$\mathrm{var}(G_t) \approx 4(1-p)^2\mathrm{var}(G_{t-1}) + 2^t p(1-p)^t = \sum_{j\le 2}2^{2t-j}(1-p)^{2(t-j)}p(1-p)^j \tag{S.41}$$

$$= \frac{p\left((2-2p)^t-1\right)(2-2p)^t}{1-2p} = \frac{p}{1-2p}\left[\left(2^t(1-p)\right)^2 - 2^t(1-p)^t\right]. \tag{S.42}$$

Considering the cases $p < 1/2$ and $p > 1/2$ separately, we obtain the simple expressions:

$$\mathrm{var}(G_t) = \begin{cases} \frac{p}{1-2p}\left[2(1-p)\right]^{2t} & \text{if } p < 1/2 \\ \frac{p}{2p-1}\left[2(1-p)\right]^{t} & \text{if } p > 1/2 \end{cases} \tag{S.43}$$

Rewriting this expression in terms of the total number of cells $N = 2^t$, we have

$$\mathrm{var}(G_t|N) = \begin{cases} \frac{p}{1-2p} N^{2+2\ln(1-p)/\ln(2)} & \text{if } p < 1/2 \\ \frac{p}{2p-1} N^{1+1\ln(1-p)/\ln(2)} & \text{if } p > 1/2 \end{cases} \tag{S.44}$$

This formula seems to do a very good job of predicting the growth of the conditional variance for an asynchronous model as well; see Figure 5.

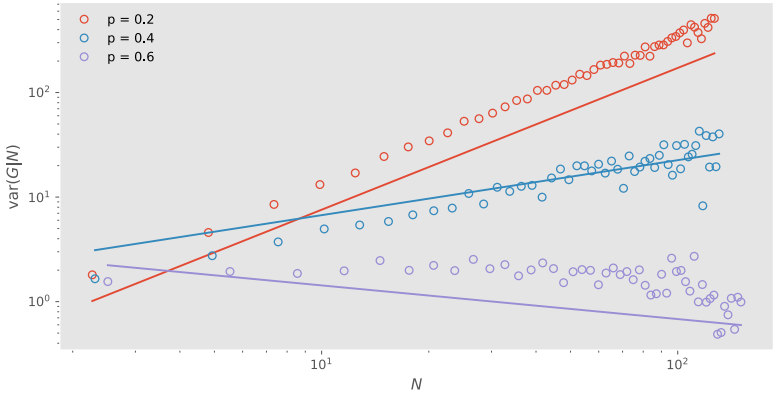

Figure 5: An illustration of the phase transition in Equation S.44 with theoretical predictions of the synchronous model (lines) and simulations of the asynchronous model (circles).

C.2 Broken state model

Let the probability for a modified cell to break be p and the probability for a broken cell to repair in a given generation be q. Letting B_t and G_t be the number of broken and green cells, we have

$$M_t = 2M_{t-1} - \sum_{i}^{2M_{t-1}} \xi_{t,\mathrm{break},i} \tag{S.45}$$

$$B_t = \sum_{i=1}^{B_{t-1}} (1 - \xi_{t,\mathrm{repair},i}) + \sum_{i}^{2M_{t-1}} \xi_{t,\mathrm{break},i} \tag{S.46}$$

$$G_t = 2G_{t-1} + \sum_{i=1}^{B_{t-1}} \xi_{t,\mathrm{repair},i} \tag{S.47}$$

where $\xi_{\mathrm{repair},i}$ and $\xi_{\mathrm{break},i}$ are 1 if the ith cells in the sums are broken or repaired respectively, and zero otherwise. Making a Gaussian approximation, we have

$$\begin{aligned} M_t &= 2M_{t-1}(1-p) + \sqrt{2p(1-p)M_{t-1}}\eta_{2,t} \\ B_t &= (1-q)B_{t-1} + 2pM_{t-1} - \sqrt{q(1-q)B_{t-1}}\eta_{1,t} - \sqrt{2p(1-p)M_{t-1}}\eta_{2,t} \\ G_t &= 2G_{t-1} + qB_{t-1} + \sqrt{q(1-q)B_{t-1}}\eta_{1,t}. \end{aligned} \tag{S.48}$$

Here, $\eta_{i,t}$ are independent standard normal random variates, but note that we have been careful to flip the signs in-front of the noise terms above so that, e.g., a positive value of $\eta_{2,t}$ in the first equation corresponds to a negative value in the second.

Conditioned on the previous generation, $\mathbf{z}_t = (M_t, B_t, G_t)^T$ has a multivariate Gaussian distribution:

$$f_t(\mathbf{z}|\mathbf{z}') = \frac{1}{\sqrt{2\pi}\det\Sigma_t(\mathbf{z}')^{1/2}} e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_t(\mathbf{z}'))^T\Sigma_t(\mathbf{z}')^{-1}(\mathbf{z}-\boldsymbol{\mu}_t(\mathbf{z}'))} \tag{S.49}$$

The mean and covariance matrix can be determined from Equation (S.48). For the mean, we have

$$\boldsymbol{\mu}_t = \begin{bmatrix} \langle M_t|\mathbf{z}_{t-1}\rangle \\ \langle B_t|\mathbf{z}_{t-1}\rangle \\ \langle G_t|\mathbf{z}_{t-1}\rangle \end{bmatrix} = A\mathbf{z}_{t-1} \tag{S.50}$$

where

$$A = \begin{bmatrix} 2-2p & 0 & 0 \\ 2p & 1-q & 0 \\ 0 & q & 2 \end{bmatrix}. \tag{S.51}$$

This leads to the recursive equation for the average:

$$\langle \mathbf{z}_t\rangle = A\langle \mathbf{z}_{t-1}\rangle \tag{S.52}$$

Taking the initial conditions to be $\mathbf{z}_0 = (1, 0, 0)$ (that is, there is initially one modified cell), it is a straightforward exercise in linear algebra (or a few lines in Mathematica) to obtain

$$\langle \mathbf{z}_t\rangle = \begin{bmatrix} (2-2p)^t \\ \frac{2p\left((1-q)^t-(2-2p)^t\right)}{2p-q-1} \\ -\frac{q\left(2p\left(2^t-(1-q)^t\right)+(q+1)2^t\left((1-p)^t-1\right)\right)}{(q+1)(-2p+q+1)} \end{bmatrix}. \tag{S.53}$$

We find (as we could have deduced without this calculation) that the fraction of green cells approaches one as $t \to \infty$.

Now we turn to the covariance matrix. The elements of Σ_t, are determined by variance and covariance of B_t and G_t:

$$\begin{aligned} \Sigma_{t,1,1} &= \mathrm{var}(M_t|\mathbf{z}_{t-1}) = p(1-p)2M_{t-1} \\ \Sigma_{t,2,2} &= \mathrm{var}(B_t|\mathbf{z}_{t-1}) = q(1-q)B_{t-1} + p(1-p)2M_{t-1} \\ \Sigma_{t,3,3} &= \mathrm{var}(G_t|\mathbf{z}_{t-1}) = q(1-q)B_{t-1} \end{aligned} \tag{S.54}$$

and

$$\begin{aligned} \Sigma_{t,1,2} &= \mathrm{cov}(M_t, B_t|\mathbf{z}_{t-1}) = 2p(1-p)M_{t-1} \\ \Sigma_{t,2,3} &= \mathrm{cov}(B_t, G_t|\mathbf{z}_{t-1}) = q(1-q)B_{t-1} \\ \Sigma_{t,3,1} &= \mathrm{cov}(M_t, G_t|\mathbf{z}_{t-1}) = 0. \end{aligned} \tag{S.55}$$

The covariance matrix is therefore,

$$\Sigma_t = \begin{bmatrix} 2p(1-p)M_{t-1} & 2p(1-p)M_{t-1} & 0 \\ 2p(1-p)M_{t-1} & q(1-q)B_{t-1} + 2p(1-p)M_{t-1} & q(1-q)B_{t-1} \\ 0 & q(1-q)B_{t-1} & q(1-q)B_{t-1} \end{bmatrix}. \tag{S.56}$$

We are interested in the unconditional distribution $P_t(\mathbf{z})$, so we write the recursive equation

$$P_t(\mathbf{z}) = \int f_t(\mathbf{z}|\mathbf{z}')P_{t-1}(\mathbf{z}')d\mathbf{z}'. \tag{S.57}$$

We make ansatz that $P_t(\mathbf{z})$ is a Gaussian

$$P_t(\mathbf{z}) = \frac{1}{\sqrt{2\pi}\ [\det\Omega_t]^{1/2}} e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\nu}_t)^T\Omega_t^{-1}(\mathbf{z}-\boldsymbol{\nu}_t)}. \tag{S.58}$$

This is equivalent to assuming that B_{t-1} and G_{t-1} appearing in the elements of the covariance matrix can be replaced by their averages, thus making the joint distribution of $\mathbf{z}$ and $\mathbf{z}'$ also a multivariate Gaussian. This allows us to write

$$\begin{aligned} P_t(\mathbf{z},\mathbf{z}') = f_t(\mathbf{z}|\mathbf{z}')P_{t-1}(\mathbf{z}') &\propto e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_t(\mathbf{z}'))^T\Sigma_t^{-1}(\mathbf{z}-\boldsymbol{\mu}_t(\mathbf{z}'))-\frac{1}{2}(\mathbf{z}'-\boldsymbol{\nu}_{t-1})^T\Omega_{t-1}^{-1}(\mathbf{z}'-\boldsymbol{\nu}_{t-1})} \\ &\propto e^{-\frac{1}{2}(\mathbf{z}-\hat{\boldsymbol{\mu}}_t)^T\hat{\Sigma}_t^{-1}(\mathbf{z}-\hat{\boldsymbol{\mu}}_t)} \end{aligned} \tag{S.59}$$

where $\hat{\boldsymbol{\mu}}_t$ and $\hat{\Sigma}_t$ are the mean and covariance matrices of the joint distribution of the current and previous generations. Since $\mathbf{z}$ and $\mathbf{z}'$ have marginal distributions given by $P_t(\mathbf{z})$ and $P_{t-1}(\mathbf{z})$ respectively, $\hat{\Sigma}_t$ has the block form:

$$\hat{\Sigma}_t = \begin{bmatrix} \Omega_t & U_t \\ U_t^T & \Omega_{t-1} \end{bmatrix}. \tag{S.60}$$

where

$$U_t = \begin{bmatrix} \mathrm{cov}(M_t, M_{t-1}) & \mathrm{cov}(M_t, B_{t-1}) & \mathrm{cov}(M_t, G_{t-1}) \\ \mathrm{cov}(B_t, M_{t-1}) & \mathrm{cov}(B_t, B_{t-1}) & \mathrm{cov}(B_t, G_{t-1}) \\ \mathrm{cov}(G_t, M_{t-1}) & \mathrm{cov}(G_t, B_{t-1}) & \mathrm{cov}(G_t, G_{t-1}) \end{bmatrix}. \tag{S.61}$$

After some algebra, we find that we can write the matrices U_t in terms of the entries of Ω_{t-1} (dropping the time subscript for clarity):

$$U_t = \begin{bmatrix} 2(1-p)\Omega_{1,1} & 2(1-p)\Omega_{1,2} & 2(1-p)\Omega_{t-1,1,3} \\ (1-q)\Omega_{2,1} + 2p\Omega_{1,1} & (1-q)\Omega_{2,2} + 2p\Omega_{1,2} & (1-q)\Omega_{2,3} + 2p\Omega_{1,3} \\ 2\Omega_{3,1} + q\Omega_{2,1} & 2\Omega_{3,2} + q\Omega_{2,2} & 2\Omega_{3,3} + q\Omega_{3,2} \end{bmatrix} \tag{S.62}$$

$$= D\Omega_{t-1} \tag{S.63}$$

where

$$D = \begin{bmatrix} 2(1-p) & 0 & 0 \\ 2p & (1-q) & 0 \\ 0 & q & 2 \end{bmatrix}. \tag{S.64}$$

In order to derive an equation for Ω_t, we notice that standard properties of multivariate normal distributions allow us to write Σ_t (which is known) as a Shur complement of the blocks in $\hat{\Sigma}$:

$$\Sigma_t = \Omega_t - U_t\Omega_{t-1}^{-1}U_t^T \tag{S.65}$$

which gives us a recursive relation for Ω_t:

$$\Omega_t = \Sigma_t + D(D\Omega_{t-1})^T = \Sigma_t + D\Omega_{t-1}D^T \tag{S.66}$$

Taking Ω_0 to be a matrix of all zeros,

$$\Omega_t = \sum_{j\le t} D^{t-j}\Sigma_j \left(D^T\right)^{t-j} \tag{S.67}$$

We are actually interested in the variation in green cells conditioned on the total number of cells, thus, we introduce the transformation

$$L = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}. \tag{S.68}$$

This gives us the covariance matrix for (N_t, G_t):

$$\bar{\Omega}_t = L\Omega_t L^T = \sum_{j\le t} LD^{t-j}\Sigma_j \left(D^T\right)^{t-j} L^T \tag{S.69}$$

In terms of the entries of $\tilde{\Omega}_t$,

$$\mathrm{var}(G_t|N_t) = \bar{\Omega}_{t,2,2} - \frac{\bar{\Omega}_{t,1,2}^2}{\bar{\Omega}_{t,1,1}}. \tag{S.70}$$

Although we don't have a closed form for $\mathrm{var}(G_t|N_t)$, it can easily be computed numerically. We find that the broken state seems to have no effect on the long term growth of $\mathrm{var}(G_t|N_t)$; see Figure 6.

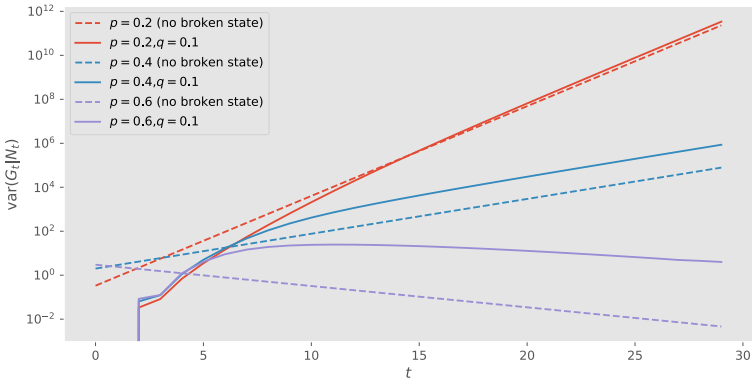

Figure 6: Theoretical predictions for the growth of $\mathrm{var}(G|N)$ for the model with a broken state using Equation S.70 compared to the predictions for a model with no broken state (Equation S.44).