Dynamics of double-strand break and repair in growing populations

Yeast team[1]

[1] *Harvard, Pasteur, Peking*

Contents

1 Introduction

2 Analysis of pooled data reveals inconsistencies with simple 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 2 (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 ρ. The broken cells can then become repaired 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 equation that m and b will eventually grow (assuming $\alpha > \beta$) exponentially at a rate $\alpha - \beta$, while g will eventually grows at a rate α. In particular, 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:

$$\phi_{m+b} \sim 1 - e^{-\beta t}. \tag{4}$$

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 exponentially decay is seen in some, but no all experiments; see Figure 2 (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}.

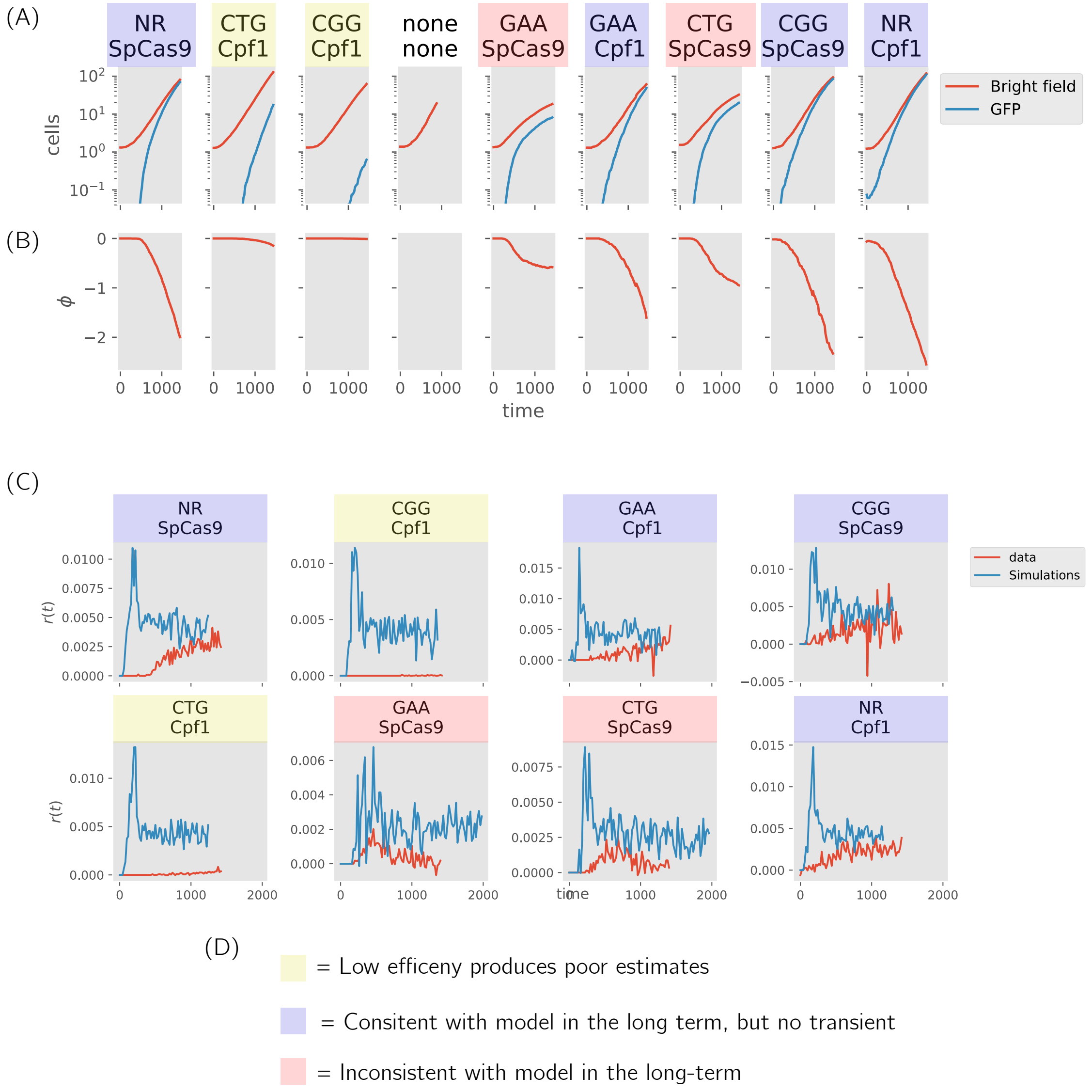

Figure 1: **Analysis of pooled data** (A) The average number of bright field and GFP cells as a function of time in each condition. (B) The fraction of non-GFP cells as function of time for each condition. (C) The rate to turn obtained from Equation 6 compared to simulations with the values of inferred from the trajectories in (B). Classification of experiments in terms of comparison between predictions from (B) and (C).

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

$$\beta = \frac{1}{n-g-b}\left[\frac{d}{dt}b + \frac{d}{dt}g - \frac{g}{n-b}\frac{d}{dt}(n-b)\right] \tag{5}$$

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. However, if $\rho \gg \beta$, we can make the approximation that the number of broken cells is small to obtain

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

In Figure 2 (C) we compare $r(t)$ compute for the data and simulations of the model with $\rho \to \infty$ so that there are no broken cells. We first focus on the experiments for which ϕ_{m+b} exhibits clear exponential decay, as in the simulations and data appear to converge (even if the transient dynamics differ). However, notice that even in these experiments $r(t)$ appears to be larger in the simulations. This makes sense, since $r(t)$ is an approximation to β. Plugging the solution of the ODEs into Equation 6 we find

$$r(t) = \frac{\beta\rho\left(e^{t(\alpha-\beta+\rho)} - 1\right)}{(\alpha+\rho)e^{t(\alpha-\beta+\rho)} - \beta} \tag{7}$$

and therefore $r(t)$ converges to

$$r_\infty = \frac{\beta\rho}{\alpha+\rho} \tag{8}$$

as $t \to \infty$, which, as expected, converges to β when $\rho \to \infty$ (the repairing of broken cells as fast). Note that $\beta\rho/(\alpha+\rho) < \beta$, therefore it makes sense that computation of $r(t)$ seems to produce an underestimate of β in the long-time limit.

We can estimate ρ from the formula

$$\rho = \frac{\alpha r_\infty}{r_\infty - \beta}. \tag{9}$$

Using these estimates of ρ along with estimates of β obtained previously from fitting ϕ_{m+g}, we simulate the full model with the broken state and see if any of the discrepancies in the transient dynamics are explained by a slow repair rate. The results are shown in Figure **??** (A). Including a finite time between break and repair seems to explain the transient dynamics in $r(t)$ for the experiments where ϕ_{m+b} decays exponentially (highlighted in purple) but no the experiments where few green cells appear (yellow) or the ones with biphasic decay (red). In order to explain the "bump" in the CAA-SpCas9 and CTG-SpCas9, we assume that some fraction of the repaired cells fail to turn green, but still growth. Some simulations for carrying parameters of this model are shown in Figure **??** (B), where we can see that this will indeed produce the behavior observed in some experiments [can we estimate the probability not to turn green?].

3 Fluctuation analysis

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

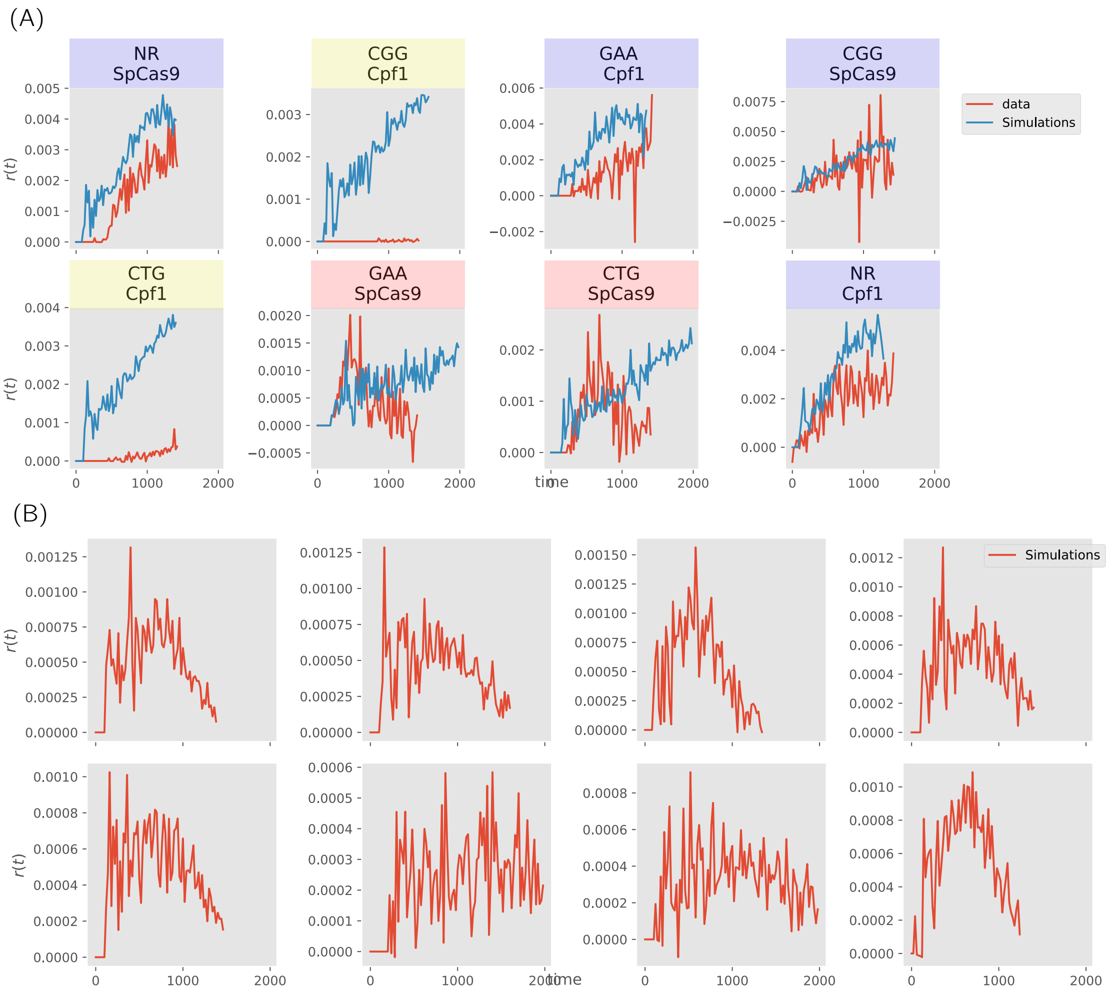

Figure 2: (A) The average number of bright field and GFP cells as a function of time in each condition. (B) The fraction of non-GFP cells as function of time for each condition. (C) The rate to turn obtained from Equation **??** compared to simulations with the values of inferred from the trajectories in (B).

A Analysis of ODE model

[UNFINISHED]

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, at 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

$$\begin{aligned}\frac{d}{dt}P(n,g,t) &= \alpha(n-g-1)P(n-1,g,t) + \alpha(m-1)P(n-1,g-1,t) \\ &+ \beta(n-g+1)p(n,g-1,t) - [\alpha n + \beta(n-g)]\, P(n,g,t)\end{aligned} \tag{10}$$

For the averages, we have

$$\frac{d}{dt}\langle N\rangle = \alpha\langle N\rangle \tag{11}$$

$$\frac{d}{dt}\langle G\rangle = \alpha\langle G\rangle + \beta(\langle N\rangle - \langle G\rangle) \tag{12}$$

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

$$\langle G\rangle = e^{(\alpha-\beta)t}\left[e^{\beta t} - 1\right]. \tag{13}$$

We now consider the second moment:

$$\frac{d}{dt}\langle G^2\rangle = \sum_{g,n} g^2 \frac{d}{dt}P(n,g,t) \tag{14}$$

$$= \sum_n \sum_g \alpha(n-g)g^2 P(n,g,t) + \sum_n \sum_g \alpha g(g+1)^2 P(n,g,t) \tag{15}$$

$$+ \sum_g \sum_g \beta(n-g)(g+1)^2 P(n,g,t) - \sum_n \sum_g \left[\alpha n + \beta(n-g)\right] g^2 P(n,g,t) \tag{16}$$

$$= \alpha\langle NG^2\rangle - \alpha\langle G^3\rangle + \alpha\langle G^3\rangle + 2\alpha\langle G^2\rangle + \alpha\langle G\rangle + \beta\langle NG^2\rangle - \beta\langle G^3\rangle \tag{17}$$

$$+ 2\beta\langle NG\rangle - 2\beta\langle G^2\rangle + \beta\langle N\rangle - \beta\langle G\rangle - \langle NG^2\rangle - \beta\langle NG^2\rangle + \beta\langle G^3\rangle \tag{18}$$

After cancelling terms, we find

$$\frac{d}{dt}\langle G^2\rangle = 2\alpha\langle G^2\rangle + \alpha\langle G\rangle + 2\beta\langle NG\rangle \tag{19}$$

$$- 2\beta\langle G^2\rangle + \beta\langle N\rangle - \beta\langle G\rangle. \tag{20}$$

For the variance

$$\frac{d}{dt}\mathrm{var}(G) = \frac{d}{dt}\left[\langle G^2\rangle - \langle G\rangle^2\right] \tag{21}$$

$$= \frac{d}{dt}\langle G^2\rangle - 2\langle G\rangle\frac{d}{dt}\langle G\rangle \tag{22}$$

$$= 2(\alpha-\beta)\mathrm{var}(G) + 2\beta\mathrm{cov}(G,N) + (\alpha-\beta)\langle G\rangle - \langle N\rangle. \tag{23}$$

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

$$\frac{d}{dt}\langle GN\rangle = \alpha\langle G\rangle + 2\alpha\langle GN\rangle - \beta\langle GN\rangle + \beta\langle N^2\rangle. \tag{24}$$

Subtracting the equations for the means gives

$$\frac{d}{dt}\mathrm{cov}(G,N) = \frac{d}{dt}\langle GN\rangle - \langle G\rangle\frac{d}{dt}\langle N\rangle - \langle N\rangle\frac{d}{dt}\langle G\rangle \tag{25}$$

$$= \beta\mathrm{var}(N) + (2\alpha-\beta)\mathrm{cov}(G,N) + \alpha\langle G\rangle. \tag{26}$$

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

$$\mathrm{var}(G) \sim e^{2\alpha t} \tag{27}$$

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

$$G_t = 2G_t + 2 \sum_{j=1}^{2^{t-1}-G_{t-1}} \xi_{t,j} \tag{28}$$

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 28 as

$$G_t \approx 2(1-p)G_{t-1} + p2^t + \sqrt{p(1-p)(2^t - 2G_{t-1})}\eta_t \tag{29}$$

where η_t are independent standard normal random variates. For the mean (this actually be obtained directly from Equation 28 and is exact), we have

$$\langle G_t \rangle = 2(1-p)\langle G_{t-1} \rangle + p2^t = 2^t \sum_{j\leq} (1-p)^{t-j} p \tag{30}$$

$$= 2^t \left[1 - (1-p)^t\right] \tag{31}$$

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

$$G_t \approx 2(1-p)G_{t-1} + p2^t + \sqrt{2^t p(1-p)^t}\eta_t. \tag{32}$$

Taking the variance of both sides, we have

$$\text{var}(G_t) \approx 4(1-p)^2 \text{var}(G_{t-1}) + 2^t p(1-p)^t = \sum_{j\leq 2} 2^{2t-j}(1-p)^{2(t-j)} p(1-p)^j \tag{33}$$

$$= \frac{p\left((2-2p)^t - 1\right)(2-2p)^t}{1-2p} = \frac{p}{1-2p}\left[\left(2^t(1-p)\right)^2 - 2^t(1-p)^t\right]. \tag{34}$$

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

$$\text{var}(G_t) = \begin{cases} \frac{p}{1-2p}\left[2(1-p)\right]^{2t} & \text{if } p < 1/2 \\ \frac{p}{2p-1}\left[2(1-p)\right]^{t} & \text{if } p > 1/2 \end{cases} \tag{35}$$

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

$$\text{var}(G_t|N) = \begin{cases} \frac{p}{1-2p} N^{2+2\ln(1-p)/\ln(2)} & \text{if } p < 1/2 \\ \frac{p}{2p-1} N^{1+1\ln(1-p)/\ln(2)} & \text{if } p > 1/2 \end{cases} \tag{36}$$

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

$$M_t = 2M_{t-1} - \sum_{i}^{2M_{t-1}} \xi_{t,\text{break},i} \tag{37}$$

$$B_t = \sum_{i=1}^{B_{t-1}} (1 - \xi_{t,\text{repair},i}) + \sum_{i}^{2M_{t-1}} \xi_{t,\text{break},i} \tag{38}$$

$$G_t = 2G_{t-1} + \sum_{i=1}^{B_{t-1}} \xi_{t,\text{repair},i} \tag{39}$$

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

$$\begin{aligned} M_t &= 2M_{t-1}(1-p) + \sqrt{2p(1-p)M_{t-1}}\eta_{2,t} \\ B_t &= (1-q)B_{t-1} + 2pM_{t-1} - \sqrt{q(1-q)B_{t-1}}\eta_{1,t} - \sqrt{2p(1-p)M_{t-1}}\eta_{2,t} \\ G_t &= 2G_{t-1} + qB_{t-1} + \sqrt{q(1-q)B_{t-1}}\eta_{1,t}. \end{aligned} \tag{40}$$

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 to that, e.g., a positive value of $\eta_{2,t}$ in the first equation corresponds to a negative value in the second. Since we ultimately interested in the distribution of G_t conditioned on the total number of cells, $N_t = G_t + B_t + M_t$, we consider the equation of N_t:

$$N_t = 2(N_t - B_t). \tag{41}$$

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

$$f_t(\mathbf{z}|\mathbf{z}') = \frac{1}{\sqrt{2\pi}\mathrm{det}\Sigma_t(\mathbf{z}')^{1/2}} e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_t(\mathbf{z}'))^T \Sigma_t(\mathbf{z}')^{-1}(\mathbf{z}-\boldsymbol{\mu}_t(\mathbf{z}'))} \tag{42}$$

The mean and covariance matrix can be determined from Equation (40). For the mean, we have

$$\boldsymbol{\mu}_t = \begin{bmatrix} \langle M_t|\mathbf{z}_{t-1}\rangle \\ \langle B_t|\mathbf{z}_{t-1}\rangle \\ \langle G_t|\mathbf{z}_{t-1}\rangle \end{bmatrix} = A\mathbf{z}_{t-1} \tag{43}$$

where

$$A = \begin{bmatrix} 2-2p & 0 & 0 \\ 2p & 1-q & 0 \\ 0 & q & 2 \end{bmatrix}. \tag{44}$$

Writing the initial $\mathbf{z}_0 = (1,0,0)$, we find

$$\langle \mathbf{z}_t \rangle = \begin{bmatrix} (2-2p)^t \\ \frac{2p\left((1-q)^t-(2-2p)^t\right)}{2p-q-1} \\ -\frac{q\left(2p\left(2^t-(1-q)^t\right)+(q+1)2^t\left((1-p)^t-1\right)\right)}{(q+1)(-2p+q+1)} \end{bmatrix}. \tag{45}$$

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:

$$\Sigma_{t,1,1} = \mathrm{var}(M_t|\mathbf{z}_{t-1}) = p(1-p)2M_{t-1} \tag{46}$$

$$\Sigma_{t,2,2} = \mathrm{var}(B_t|\mathbf{z}_{t-1}) = q(1-q)B_{t-1} + p(1-p)2M_{t-1} \tag{47}$$

$$\Sigma_{t,3,3} = \mathrm{var}(G_t|\mathbf{z}_{t-1}) = q(1-q)B_{t-1} \tag{48}$$

and

$$\Sigma_{t,1,2} = \mathrm{cov}(M_t, B_t|\mathbf{z}_{t-1}) = 2p(1-p)M_{t-1} \tag{49}$$

$$\Sigma_{t,2,3} = \mathrm{cov}(B_t, G_t|\mathbf{z}_{t-1}) = q(1-q)B_{t-1} \tag{50}$$

$$\Sigma_{t,3,1} = \mathrm{cov}(M_t, G_t|\mathbf{z}_{t-1}) = 0 \tag{51}$$

The covariance matrix is therefore,

$$\Sigma_t = \begin{bmatrix} 2p(1-p)M_{t-1} & 2p(1-p)M_{t-1} & 0 \\ 2p(1-p)M_{t-1} & q(1-q)B_{t-1} + 2p(1-p)M_{t-1} & q(1-q)B_{t-1} \\ 0 & q(1-q)B_{t-1} & q(1-q)B_{t-1} \end{bmatrix}. \tag{52}$$

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

$$P_t(\mathbf{z}) = \int f_t(\mathbf{z}|\mathbf{z}')P_{t-1}(\mathbf{z}')d\mathbf{z}'. \tag{53}$$

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

$$P_t(\mathbf{z}) = \frac{1}{\sqrt{2\pi}\,[\det\Omega_t]^{1/2}} e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\nu}_t)^T\Omega_t^{-1}(\mathbf{z}-\boldsymbol{\nu}_t)}. \tag{54}$$

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 Guassian:

$$P_t(\mathbf{z},\mathbf{z}') = f_t(\mathbf{z}|\mathbf{z}')P_{t-1}(\mathbf{z}') \propto e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_t(\mathbf{z}'))^T\Sigma_t^{-1}(\mathbf{z}-\boldsymbol{\mu}_t(\mathbf{z}'))-\frac{1}{2}(\mathbf{z}'-\boldsymbol{\nu}_{t-1})^T\Omega_{t-1}^{-1}(\mathbf{z}'-\boldsymbol{\nu}_{t-1})} \tag{55}$$

$$\propto e^{-\frac{1}{2}(\mathbf{z}-\hat{\boldsymbol{\mu}}_t)^T\hat{\Sigma}_t^{-1}(\mathbf{z}-\hat{\boldsymbol{\mu}}_t)} \tag{56}$$

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:

$$\hat{\Sigma}_t = \begin{bmatrix} \Omega_t & \hat{\Sigma}_{1,2} \\ \hat{\Sigma}_{1,2}^T & \Omega_{t-1} \end{bmatrix}. \tag{57}$$

where

$$\hat{\Sigma}_{1,2} = \begin{bmatrix} \mathrm{cov}(M_t, M_{t-1}) & \mathrm{cov}(M_t, B_{t-1}) & \mathrm{cov}(M_t, G_{t-1}) \\ \mathrm{cov}(B_t, M_{t-1}) & \mathrm{cov}(B_t, B_{t-1}) & \mathrm{cov}(B_t, G_{t-1}) \\ \mathrm{cov}(G_t, M_{t-1}) & \mathrm{cov}(G_t, B_{t-1}) & \mathrm{cov}(G_t, G_{t-1}) \end{bmatrix}. \tag{58}$$

After some algebra, we can now write $\hat{\Sigma}_{1,2}$ in terms of the entries of Ω_{t-1} (dropping the time subscript for clarity):

$$\hat{\Sigma}_{t,1,2} = \begin{bmatrix} 2(1-p)\Omega_{1,1} & 2(1-p)\Omega_{1,2} & 2(1-p)\Omega_{t-1,1,3} \\ (1-q)\Omega_{2,1} + 2p\Omega_{1,1} & (1-q)\Omega_{2,2} + 2p\Omega_{1,2} & (1-q)\Omega_{2,3} + 2p\Omega_{1,3} \\ 2\Omega_{3,1} + q\Omega_{2,1} & 2\Omega_{3,2} + q\Omega_{2,2} & 2\Omega_{3,3} + q\Omega_{3,2} \end{bmatrix} \tag{59}$$

$$= D\Omega_{t-1} \tag{60}$$

where

$$D = \begin{bmatrix} 2(1-p) & 0 & 0 \\ 2p & (1-q) & 0 \\ 0 & q & 2 \end{bmatrix}. \tag{61}$$

In order to derive an equation for Ω_t, we notice that Σ_t (which is known) can be expressed as the shur complement

$$\Sigma_t = \Omega_t - \hat{\Sigma}_{t,1,2}\Omega_{t-1}^{-1}\hat{\Sigma}_{t,1,2}^T \tag{62}$$

which gives us a recursive relation for Ω_t:

$$\Omega_t = \Sigma_t + D(D\Omega_{t-1})^T = \Sigma_t + D\Omega_{t-1}D^T \tag{63}$$

Taking $\Omega_0 = 0$,

$$\Omega_t = \sum_{j\le t} D^{t-j}\Sigma_j \left(D^T\right)^{t-j} \tag{64}$$

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

$$L = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}. \tag{65}$$

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

$$\bar{\Omega}_t = L\Omega_t L^T = \sum_{j\le t} LD^{t-j}\Sigma_j \left(D^T\right)^{t-j} L^T \tag{66}$$

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

$$\mathrm{var}(G_t|N_t) = \bar{\Omega}_{t,2,2} - \frac{\bar{\Omega}_{t,1,2}^2}{\bar{\Omega}_{t,1,1}}. \tag{67}$$