APPENDIX: MATHEMATICAL DETAILS

Standard applications of branching processes (e.g. (1)) assume that we observe only a few realizations (and often only one) of the process over many generations. We, on the other hand, have only a single observation (the final live and dead cell counts) for a large number of realizations (in our context, each embryo is a single realization of the process). This means that many available results on parameter estimation for branching processes are not applicable. Nevertheless, as indicated in the main text, it turns out that we can estimate all the required parameters of the model using standard maximum likelihood techniques, and in particular by maximizing the likelihood $L(\alpha,\delta,n_s,n_f)$ of the data. It is in fact more convenient to work with $\log L$ rather than L itself, because this is numerically more manageable, and clearly the choice of $L(\alpha,\delta,n_s,n_f)$ that maximizes $\log L$ is the same as that which maximizes L.

Computing the likelihood

To compute L, or $\log L$, we need the probability $Q_{jk}^{(n_f)}(\alpha,\delta,n_s)$ of an embryo having j live and k dead cells at generation n_f for a model with parameters α, δ and initial condition 2^{n_s} cells at generation n_s. Then if we have observed R embryos, with live and dead cell counts $j(r)$ and $k(r)$ for the rth embryo (with $1 \le r \le R$), the log likelihood is simply

$$\log L(\alpha,\delta,n_s,n_f) \quad = \quad \sum_{r=1}^{R} \log Q_{j(r)k(r)}^{(n_f)}(\alpha,\delta,n_s) \; . \qquad [\mathbf{1}]$$

To compute $\log L$, it thus remains only to show how to calculate $Q_{jk}^{(n_f)}(\alpha,\delta,n_s)$. For ease of exposition, let us from now on drop α, δ and n_s from the notation for $Q_{jk}^{(n_f)}$. Recall that we assume that the dead cells counted in Fig. 2*A* are precisely those that have died in the transition from generation n_f-1 to generation n_f. This depends purely on the number of live cells in generation n_f-1. Thus, we may compute $Q_{jk}^{(n_f)}$ from the probability distribution $P_i^{(n_f-1)}$ of the number of live cells in generation n_f-1 (where $P_i^{(n)}$ is the probability that at the nth generation an embryo has i live cells) using the formula

$$Q_{jk}^{(n+1)} \quad = \quad \sum_{i=0}^{2^n} B_{jki}^{(n)} P_i^{(n)} \; , \qquad [\mathbf{2}]$$

where $B_{jki}^{(n)}$ is the probability that an embryo with i cells at generation n gives rise to precisely j live cells and k dead cells at generation n+1. Note that the summation needs to be taken only to $i = 2^n$, because $P_i^{(n)} = 0$ for $i > 2^n$. This is because the number of cells at most doubles in each generation, and hence 2^n is the largest possible number of cells at generation n. We can derive an explicit expression for $B_{jki}^{(n)}$ in terms of the probabilities α, δ, and γ by considering an embryo with i cells at generation n that gives rise to one with j live and k dead cells in generation n+1. Suppose that exactly q cells divide, creating $2q$ daughter cells. If we are to end up with exactly j cells in the next generation, this means that $j - 2q$ cells must neither divide nor die, with probability $\delta^{(j-2q)}$ for any particular choice of such cells. The remaining $i - q - (j-2q) = i - j + q$ must therefore die, with probability $\alpha^{(i-j+q)}$. But we are assuming that we have k dead cells in the next generation, so $i - j + q = k$, or $q = k - i + j$. Taking into account the number of possible permutations of choosing q cells to divide, $j - 2q$ cells do nothing, and $i - j + q$ to die, we obtain the multinomial probability:

$$B_{jki}^{(n)} = \frac{i!}{(k-i+j)!(2i-2k-j)!k!}\gamma^{(k-i+j)}\delta^{(2i-2k-j)}\alpha^{k} \qquad [3]$$

for all k such that $0 \le k \le i$, $0 \le k\text{-}i\text{+}j \le i$ and $0 \le 2i\text{-}2k\text{-}j \le i$. For any other combination of i,j,k, and embryo with i cells cannot give rise to one with j live and k dead cells, and hence $B_{jki}^{(n)} = 0$. Together, Eqs. **2** and **3** allow us to compute $Q_{jk}^{(n+1)}$ from $P_i^{(n)}$. But additionally note that

$$P_i^{(n)} = \sum_{k=0}^{2^n} Q_{ik}^{(n)}, \qquad [4]$$

and hence we have an explicit method of computing $P_i^{(n_f-1)}$ from $P_i^{(n_f-2)}$ and so on, all the way to the initial distribution $P_i^{(n_s)}$, which assuming the initial condition of 2^{n_s} cells in generation n_s is given by $P_i^{(n_s)} = 1$ for $i = 2^{n_s}$ and $P_i^{(n_s)} = 0$ for $i \neq 2^{n_s}$. Eqs. **1** to **4** thus give a complete algorithm for computing $\log L(\alpha,\delta,n_s,n_f)$ for any given (α,δ,n_s,n_f). They also allow us to calculate the probability $P_0^{(n)}$ of having no live cells at generation n, which is plotted in Fig. 5.

Maximizing the likelihood

A variety of numerical methods then exist for finding the maximum of $\log L$, e.g. (2). For ease of implementation we chose an extremely simplistic approach, namely of successive searches in each parameter in turn. We thus start with an initial guess α_0,δ_0 and in turn hold α and δ fixed while maximizing the other, using a standard gold section line search, (2). Thus, in more detail, we first compute an α_1 that maximizes $\log L(\alpha,\delta_0)$, then a δ_1 that maximizes $\log L(\alpha_1,\delta)$, then an α_2 that maximizes $\log L(\alpha,\delta_1)$ and so on, until the desired accuracy is obtained. To maximize over the discrete parameters, we simply maximize $\log L(\alpha,\delta,n_s,n_f)$ over α and δ for reasonable choices of n_s,n_f and choose those that give the highest value. Note that since there are data points with more than 128 cells in Fig. 2*A*, we must have $n_f \ge 8$.

This approach is numerically very inefficient, and far faster methods are available such as the Nelder-Mead simplex algorithm (2). However, it is extremely easy to set up and to modify as the model changes and evolves. In particular, adding further parameters involves little effort. It is thus especially appropriate for initial data exploration, where the model, and the parameters to be maximized are in a frequent state of flux. Furthermore, despite its inefficiency, this method was able to determine the maximum in a few minutes on an average personal computer for all but the most complicated models that we investigated.

Simulating the model

Having chosen the parameters it is straightforward to implement the model on a computer to simulate a population of embryos for any given number of cell cycles. For each embryo, we pass from one generation to the next simply by generating one or more random numbers for each cell in the embryo to determine the fate of that cell. Thus for our preliminary model, we generate a single random number ξ, which is uniformly distributed between 0 and 1. If $0 \le \xi < \alpha$ then the cell dies, if $\alpha \le \xi \le \alpha + \delta$ then the cell does nothing, and if $\alpha + \delta \le \xi \le 1$ (which occurs with probability $1 - \alpha - \delta = \gamma$, then the cell divides to give rise to two daughter cells. This procedure is repeated for every cell for each generation, for however many generations we wish to simulate.

The refined model

If α is selected randomly from $\alpha_1, \ldots, \alpha_m$, with probabilities $p_1, \ldots, p_m$, then the distribution of live and dead cells at the nth generation is simply the corresponding combination of the distributions for the simple model with each of the $\alpha_1, \ldots, \alpha_m$ in turn. Thus

$$Q_{jk}^{(n)} = \sum_{i=1}^{m} p_i \, Q_{jk}^{(n)}(\alpha_i) \,,$$

where $Q_{jk}^{(n)}(\alpha)$ denotes the distribution of the simple model with death rate α.

Other extensions: Dead cell clearance

In fitting the model we have hitherto assumed that each dead cell in Fig. 2*A* corresponds to one cell dying during the last cell cycle simulated by the model, i.e., in the transition from generation n_f-1 to generation n_f. To test the validity of this, we extended the model in two different ways to examine the effect of the removal of dead cells from the system. The first corresponds to the situation that dead cells are cleared relatively quickly compared to the cell cycle, so that at a given instant we observe only a fraction of those cells dying in one cycle. We model this by assuming that each of the dead cells created in the transition to the final generation had a probability β of being counted (and hence 1-β of being cleared before it could be observed). This amounts to replacing $Q_{jk}^{(n_f)}$ by $\tilde{Q}_{jk}^{(n_f)}$, where $\tilde{Q}^{(n)}$ is derived from $Q^{(n)}$ in a similar fashion to Eqs. **2** to **4** but now using binomial probabilities:

$$\tilde{Q}_{ij}^{(n)} = \sum_{k=0}^{2^n - i} \frac{k!}{j!(k-j)} (1-\beta)^{(k-j)} \beta^j \, Q_{ik}^{(n)} \,. \qquad [\mathbf{5}]$$

The second possibility, corresponding to dead cells persisting for times in excess of a cell cycle is more complex to model and a number of different implementations is possible. We chose to suppose that in the transition from generation n to generation n+1, each dead cell has a probability of β' of being cleared. This gives a distribution $\hat{Q}^{(n)}$ of live and dead cells, which is derived from $Q^{(n)}$ exactly as in Eq. **5**, except that β is replaced by 1-β'. We then assume that the live cells die, remain, or divide with probabilities α, δ, and γ as in the standard model. This allows us to compute $Q^{(n+1)}$ from $\hat{Q}^{(n)}$ using a calculation similar to that for Eq. **3**. Suppose that an embryo with i live and q dead cells gives rise to one with j live and k dead cells. Then if d cells divide, we must have j - 2d cells neither dividing nor dying. Hence i - j + d cells die, and because we start with q dead cells and end up with k, we have $q + i - j + d = k$. This implies that $d = k - q - i + j$, and substituting into j - 2d and i - j + d respectively shows that $2q + 2i - 2k - j$ cells neither divide nor die, and k - q cells die. Using multinomial probabilities similar to Eq. **3** and summing over all possibilities gives:

$$Q_{jk}^{(n+1)} = \sum_{i \geq j/2}^{j+k} \sum_{q \geq k-i+j/2}^{\min\{k,k-i+j\}} \frac{i!}{(k-q-i+j)!(2i+2q-2k-j)!(k-q)!} \gamma^{(k-q-i+j)} \delta^{(2i+2q-2k-j)} \alpha^{(k-q)} \, \hat{Q}_{iq}^{(n)}$$

for $0 \leq j \leq 2^{n+1}$, $0 \leq k \leq 2^{n+1}$-j. The summation ranges ensure that $0 \leq k$-$q \leq i$, $0 \leq 2q$+$2i$-$2k$-$j \leq i$ and $0 \leq k$-q-i+$j \leq i$.

Fitting the extended models yielded $\beta = 1$ and $\beta' = 0$, indicating that our initial assumption was reasonable. We intend to use such extended models to obtain estimates of mean clearance times and cell cycle lengths in a future paper.

Other extensions: Size dependence

The model can easily be extended to test the hypothesis that the cell death rate may depend on the size of the embryo and in particular, as suggested in the introduction and by Fig. 2*A*, that larger embryos have lower cell death rates. The simplest possible such dependence is a linear one, which can be introduced into the model by assuming that the death rate for an embryo with i live cells is given by

$$\alpha = \alpha_c + \alpha_s i\,, \qquad [\mathbf{6}]$$

where α_c and α_s are parameters that are estimated by maximizing the likelihood as before. The best fit yielded a value of $\alpha_s \approx -1.1\times10^{-4}$ corresponding to a decrease in the death rate of approximately 1% for every additional 100 cells in the embryo. Unfortunately, this value turns out not to be statistically significantly different from 0. In particular, if L_0 is the restricted maximum likelihood when $\alpha_s = 0$ in Eq. **6** and L^* the unrestricted maximum, then $2\log(L^*/L_0)$ is distributed approximately as χ^2 with 1 degree of freedom, e.g. (3). In our case we have $\log L^* \approx -2031.0$ and $\log L_0 \approx -2031.4$, yielding a χ^2 value of ≈ 0.8. We thus cannot reject the hypothesis that $\alpha_s = 0$ even at the 10% significance level. We can similarly compute the 95% confidence limits for α_s which turn out to be $[-3.9 \times 10^{-4}, 1.5 \times 10^{-4}]$.

One possible criticism of Eq. **6** is that the size dependence becomes effective only in the last few generations, when the embryo is large. A possible alternative is to scale α_s by the maximum number of possible cells in a particular generation. The effective death rate for an embryo with i cells in generation n then becomes

$$\alpha = \alpha_c + \frac{\alpha_s}{2^n} i\,. \qquad [\mathbf{7}]$$

The likelihood is now maximized at $\alpha_s \approx -4 \times 10^{-3}$ with $\log L^* \approx -2030.6$ indicating a marginally better fit to the data than with Eq. **6**. The restricted maximum likelihood when $\alpha_s = 0$ is still of course $\log L_0 \approx -2031.4$ so that we now have a χ^2 value of ≈ 1.6, which is still too small to allow us to reject the hypothesis $\alpha_s = 0$. The corresponding 95% confidence limits for α_s are now $[-1.3 \times 10^{-2}, 1.5 \times 10^{-3}]$.

We intend to investigate whether more sophisticated models can reveal a statistically significant relationship in the future.

References

1. Jagers, P. (1975) *Branching Processes with Biological Applications* (Wiley, London).
2. Press, W., Flannery, B., Teukolsky, S., Vetterling & W. (1998) *Numerical Recipes in C, The Art of Scientific Computing* (Cambridge University Press, Cambridge, U.K.).
3. Silvey, S. (1975) *Statistical Inference* (Chapman and Hall, London).