Supplemental Information

Appendix A: Some additional plots

In this section we present some additional figures that are referenced in the main text. The first figure contains four plots that show how the probability to cross N_{max} and the average extinction time vary with N_0 and with N_{max}. The second one shows how the probability to cross N_{max} and the average time to reach N_{max} vary with N_{max} plotted on the x-axis at the critical point $a = 1/2$. The last one has our three quantities plotted for the multiple cell type model of figure 3a of the main text with $n = 1$. For more details on the theoretical curves used on each plot, see appendix B.

Appendix B: Theoretical results about Galton-Watson processes

0.1. The Galton-Watson Processes

In this section we will look at how we obtained the theoretical results used in the main text by studying branching process theory. We consider a population obeying the following rules, known as a Galton-Watson process: at each time step, each individual gives birth to a random number of descendants. The probability distribution of the number of descendants is fixed throughout the problem. We will be looking at the number of individuals in the t-th generation. Considering that the individuals do not interact, it is only necessary to derive the results for a population that starts off with one individual.

Let $Z(t)$ stand for the number of individuals after t generations. Let X_i be a sequence of independent random variables whose laws correspond to the number of individuals that an individual gives birth to at each generation. The evolution of $Z(t)$ then obeys the following equation:

$$Z(t+1) = \sum_{i=1}^{Z(t)} X_i$$

We also have the condition $Z(0) = 1$. We will use the notations $m = E(X_i)$ for the average and $\sigma^2 = Var(X_i)$ for the variance.

A first result that we can obtain is the value of the mean and the variance of the random variables $Z(t)$. For this we will need to use Wald's formula. Let N be a random variable independent of all the X_i with $N \in N$. Let $Z = \sum_{i=1}^{N} X_i$. Then we have:

$$E(Z) = E(X)E(N)$$

$$Var(Z) = E(N)Var(X) + E(X)^2 Var(N)$$

Using this formula we obtain:

$$E(Z(t+1)) = E(Z(t))E(X)$$

$$Var(Z(t+1)) = E(Z(t))Var(X) + E(X)^2 Var(Z(t))$$

Solving for $E(Z(t))$ and $Var(Z(t))$, we obtain:

$$E[Z(t)] = m^t$$

$$Var[Z(t)] = \frac{\sigma^2 m^t (m^t - 1)}{m^2 - m} \quad \text{if } m \neq 1$$

$$Var[Z(t)] = t\sigma^2 \quad \text{if } m = 1$$

0.2. Generating functions and extinction probability

In order to obtain additional results, we will have to introduce certain generating functions. We define $s \to f(s)$ in the following way:

$$f(s) = \sum_{k=0}^{\infty} P(X = k) s^k$$

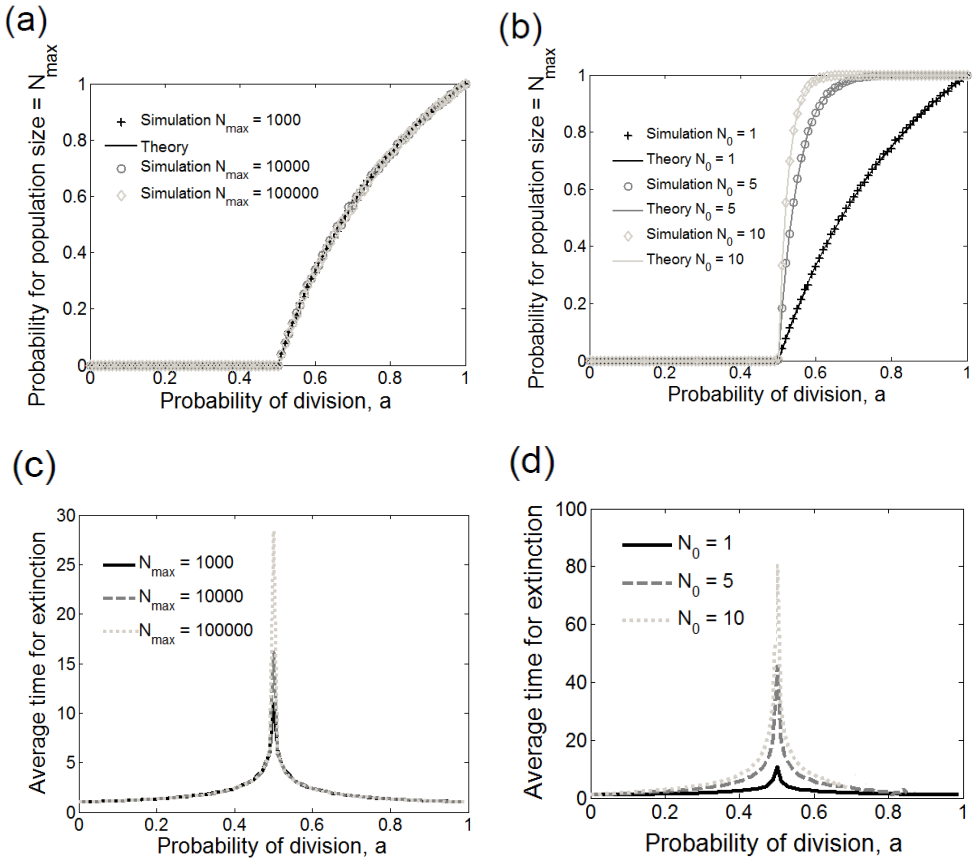


Figure 1. The plots show how the probability to reach the threshold and the average time for extinction vary when N_0 and N_{max} vary (plots done with 10000 runs of the simulation). (a) shows the probability to reach N_{max} for $N_{max} = 1000, 10000, 100000$ and $N_0 = 1$. (b) is the same probability but for $N_{max} = 1000$ and $N_0 = 1, 5, 10$. (c) shows the time to extinction for $N_{max} = 1000, 10000, 100000$ and $N_0 = 1$ and (d) is the same thing with $N_{max} = 1000$ and $N_0 = 1, 5, 10$. We see that when these parameters change the overall shapes of the curves stay the same.

We also define a generating function for each of the $Z(t)$:

$$f_t(s) = \sum_{k=0}^{\infty} P(Z(t) = k)s^k$$

By definition we have $f_0(s) = s$ and $f_1(s) = f(s)$. A first mathematical result is the following:

$$f_{t+1} = f_t(f(s))$$

This relationship can in theory be used to compute all the generating functions and find all the probability distributions of each of the $Z(t)$. Unfortunately, there are very few functions f for which it is possible to compute explicitly its iterates. To solve this problem we will have to derive asymptotic results for $t \to \infty$.

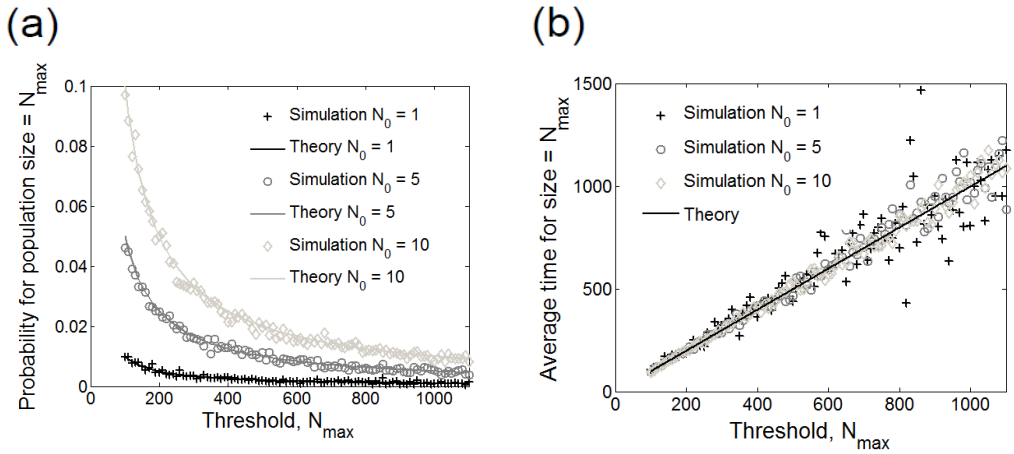


Figure 2. These plots show for $a = 1/2$ how the probability to reach N_{max} and the time it takes to reach it vary as a function of the threshold N_{max} for three values of N_0: 1, 5, 10. The plots were made with 10000 runs of the simulation. (a) shows the probability to reach N_{max} decreases as $\frac{1}{N_{max}}$ while (b) shows that the time to reach N_{max} increases linearly with N_{max}. There is also a very good fit with our theory (see appendix B).

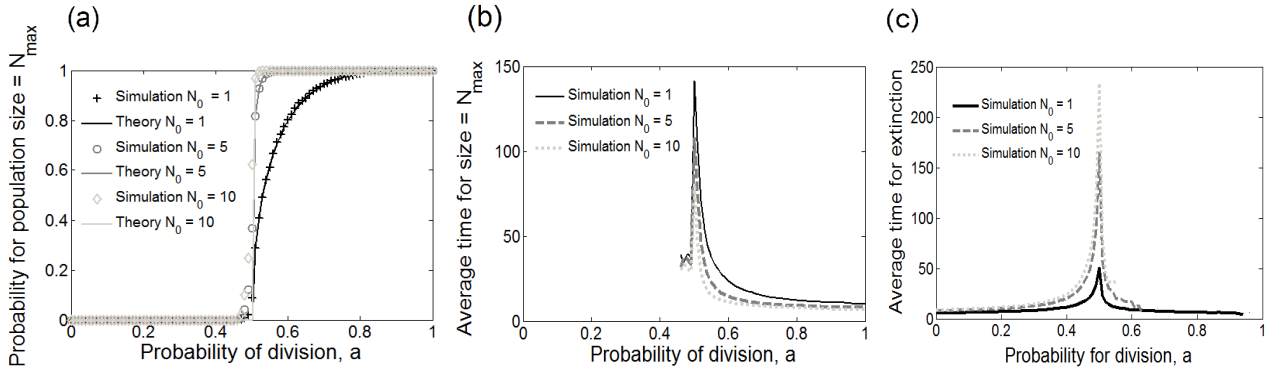


Figure 3. This figure shows our three main quantities plotted for a multiple cell type model. We have taken the model of figure 3a with $n = 1$: only one level of progenitors. We have chosen $d = 1/2$ and defined a by $a = a_0 = a_1$. The initial conditions of the simulations are one stem cell and zero of the other types of cells. We set $N_{max} = 10000$ and use again 10000 simulation runs per plot. We see that overall we obtain the same types of plots as with the one cell type two branch model. The only differences are due to the fact that here the threshold N_{max} has a bigger effect on the system because it can be reached more easily with the three different types of cells. On plot (a) we see that the transition between the two regimes at $a = 1/2$ is not as sharp as with the one cell type model: the probability is still non-zero for a slightly smaller than $1/2$. This means that the tumor sometimes reaches the threshold N_{max} for a smaller than $1/2$: it is possible to compute the average time to N_{max} in this case (b). (c) shows that for N_0 larger than one, every run reaches the threshold for a sufficiently large which explains why no extinction time can be computed after a certain value of a.

The first of these results concerns the extinction probability defined as the probability q for the population to die out at some point. Mathematically, the probability is defined by:

$$q = P(Z_t \to 0) = lim_{t\to\infty} f_t(0)$$

We have the following result:

If $m \leq 1$, then $q = 1$. If $m > 1$, then q is the smallest solution of the equation $f(s) = s$ in the interval $[0, 1[$.

We can note that $s = 1$ is always a solution of $f(s) = s$ because the sum of all probabilities must be equal to one. But the theorem states that if $m > 1$, a smaller solution will exist and this solution is the extinction

probability. It is also possible to show the following result valid for any value of m:

$$\forall s \in [0,1[\ ,\ lim_{t\to\infty} f_t(s) = q$$

In other words, $lim_{t\to\infty} P(Z_t = k) = 0$ for all $k \neq 0$. At infinity, there are only two possible outcomes for the population: it either grows infinitely large or it dies out. Convergence towards a stationary value is not possible and this can indeed pose a problem for the description of biological populations.

0.3. Asymptotic results based on the value of m

If $m \neq 1$, the global evolution of the population is pretty straightforward. If $m < 1$, the population will die out exponentially fast. If $m > 1$, the population will grow exponentially large on average. But it is still possible for the population to die out if $P(X = 0) \neq 0$; the probability of survival is worth $1 - q$.

In the case where $m = 1$, the results are less trivial than in the previous case. We know that the extinction probability is equal to 1, but because $m = 1$ we expect there to be more fluctuations before the population dies out. We have the following theorem (see [1]) :

$$\text{If } f^{''}(1) < \infty\ ,\ P(Z_t > 0) \sim_{t\to\infty} \frac{2}{tf^{''}(1)}$$

From this result, we can find the conditional expectation of Z_t knowing that $Z_t \neq 0$:

$$E(Z_t | Z_t \neq 0) = \frac{E(Z_t)}{P(Z_t \neq 0)} \sim \frac{tf^{''}(1)}{2}$$

There is also a theorem to estimate the probability for Z_t to be above a certain threshold knowing that $Z_t \neq 0$:

$$\text{If } f^{''}(1) < \infty\ ,\ lim_{t\to\infty} P(\frac{2Z_t}{tf^{''}(1)} > u | Z_t \neq 0) = e^{-u}\ ,\ \forall u \geq 0$$

We will use these results in the next section.

0.4. Application to the early stage of cancer growth

With these formulas established we are now going to derive the theoretical results used in the main text one by one. We are going to start with the one cell type model that can be written out as:

$$C \longrightarrow C + C \quad ra$$
$$C \longrightarrow C \quad 1 - r$$
$$C \longrightarrow Death \quad r(1 - a)$$

Let us first look at the average and variance of the system. We have $m = 2ar + 1 - r$ and $\sigma^2 = r - r^2(2a - 1)^2$. We see that the variance is the biggest when $a = 1/2$, when the system fluctuates the most. If $a > 1/2$, then we are in the case were the population grows on average exponentially large. The hypothesis we make to compute the probability to cross N_{max} is that if N_{max} is large enough compared to N_0, then crossing N_{max} is basically equivalent to not dying out. So all we have to do is compute the extinction probability q and the probability to cross N_{max} should roughly be equal to $1 - q$. The generating function is in this case:

$$f(s) = ras^2 + (1 - r)s + r(1 - a)$$

$$f(s) = s \Longleftrightarrow s = 1 \text{ or } \frac{1}{a} - 1$$

In the case where $a > 1/2$ there is indeed a solution of $f(s) = s$ in $[0, 1[$ that is the extinction probability:

$$q = \frac{1}{a} - 1$$

So we can conclude that the probability of crossing N_{max} is equal to $2 - \frac{1}{a}$. Figure 1c shows that the theory fits well with simulation so the approximation of a large N_{max} works very well here.

Let us now study the case where $a = 1/2 \Leftrightarrow m = 1$. In this case, the finite value of N_{max} is going to have an effect : the extinction probability is equal to 1 but we know that there is still a chance for the population to cross the threshold N_{max} at some point. Using the previous results and considering that $f''(1) = 2ra$, we have:

$$P(Z_t > 0) = \frac{1}{rat} \text{ , for } t \to \infty$$

$$E(Z_t | Z_t \neq 0) = \frac{E(Z_t)}{P(Z_t \neq 0)} \sim \frac{tf''(1)}{2}$$

These results allow to understand the shapes of the two plots of figure 2 of appendix A. We know that the probability for the population to be alive at time t is proportional to $\frac{1}{t}$. According to the conditional formula, we also know that if the population is alive at time t, then its size will be proportional to t. Combining these two results, we deduce that the probability to cross N_{max} is proportional to surviving for a time N_{max} which is proportional to $\frac{1}{N_{max}}$. This is coherent with the plot 2a which is indeed in $\frac{1}{N_{max}}$. Similarly, figure 2b shows that the time to reach N_{max} grows linearly with N_{max} which is again coherent with the result that the population grows linearly with time if it does not die out. Let us also note that the probability that the population is alive at time t decreases as $\frac{1}{t}$ when $a = 1/2$ is linked direclty to what plot 2d of the main text shows: the distribution of the extinction time decreases as one over time.

Finally, in the case where $a < 1/2$, the population size decreases on average exponentially fast, so the probability for it to cross N_{max} is going to be exponentially small; we will take it equal to 0.

0.5. Computation of the number of individual deaths and divisions (figure 2c)

In this section, we present the analytical computation that was done to obtain the plot of figure 2c of the main text. We want to understand the "phase transition" that is observed at $a = 1/2$ by using a formalism inspired by thermodynamics. We consider the simple 2-branch model ($r = 1$) with no threshold and with $a < 1/2$. Under these hypothesis, we know that the population is guaranteed to die out after some time. We will start from one cell and we want to write out a partition function for all the possible outcomes of the growth of the population before it dies out.

Looking at one process of growth (starting from one cell and going until no cells remain), we will call N the total number of individual duplications of cells throughtout the process. N is not linked in any simple way to the number of steps t that it takes for the population to die out. Starting from one cell, if N is the total number of cells that duplicated, then there must be a total of $N + 1$ cell deaths. In order to write a partition function we need to know how many possible ways they are for the population that starts with one cell to go extinct after N divisions and $N + 1$ deaths. It turns out this number can be computed and is eqaul to:

$$c_N = \frac{1}{N+1}\binom{2N}{N}$$

We can now write out a partition function for the system:

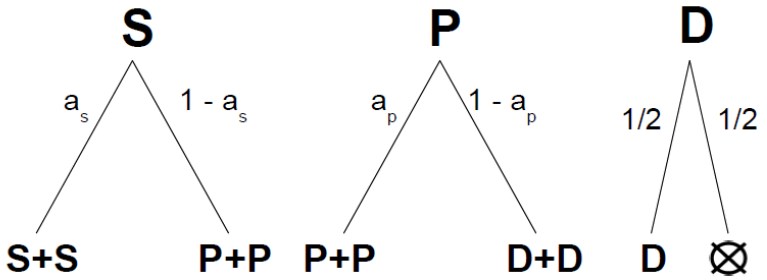
$$Z = \sum_{N=0}^{\infty} c_N a^N (1-a)^{N+1}$$

$a^N(1-a)^{N+1}$ corresponds to the probability of an outcome with N divisions and $N + 1$ deaths. We can rewrite this partition function in the following way:

$$Z = \sum_{N=0}^{\infty} c_N a^N (1-a)^{N+1} = (1-a) \sum_{N=0}^{\infty} c_N e^{-\beta N}$$

where we have

$$e^{-\beta} = a(1-a)$$



Average growth of the population	$a_s < 1/2$	$a_s = 1/2$	$a_s > 1/2$
$a_p < 1/2$	S_t and P_t both decrease exponentially fast	$S_t = S_0$, $P_t = \frac{S_0}{1-2a_p}$	$S_t = S_0(2a_s)^t$, $P_t = S_0(1-a_s)\frac{(2a_s)^t}{a_s - a_p}$
$a_p = 1/2$	S_t decreases exponentially fast, P_t is equal to the number of progenitor cells that were produced by stem cells before they died off	$S_t = S_0$, $P_t = S_0 t$	$S_t = S_0(2a_s)^t$, $P_t = S_0(1-a_s)\frac{(2a_s)^t}{a_s - 1/2}$
$a_p > 1/2$	S_t decreases exponentially fast, $P_t = P_i(2a_p)^t$ were P_i is the number of progenitors produced by stem cells	$S_t = S_0$, $P_t = S_0\frac{(2a_p)^t - 1}{2a_p - 1}$	$S_t = S_0(2a_s)^t$, $P_t = S_0(1-a_s)\frac{(2a_p)^t - (2a_s)^t}{a_p - a_s}$ $P_t = S_0(1-a_s)t(2a_s)^{t-1}$ if $a_p = a_s$.

Figure 4. This table is a detailed study of all the possible regimes of growth of the branching model shown right above the table. This model is a particular case of the general model of figure 3a with one progenitor level and $d = 1/2$. We describe the average growth of a population starting with S_0 stem cells (and none of the other types) based on the different possible values of a_s and a_p. S_t and P_t represent the average number of stem and progenitor cells at time t.

It turns out this partition function can be calculated exactly (we will drop the multiplicative constant $1-a$ in the rest of the computation as it has no impact on the result). Indeed, for $x \leq 1/4$, $\sum_{N=0}^{\infty} c_N x^N = \frac{1-\sqrt{1-4x}}{2x}$. Therefore we have:

$$Z = \frac{1-\sqrt{1-4e^{-\beta}}}{2e^{-\beta}}$$

In this formalism, we see that the "energy" of the system is equal to the average value of N:

$$\langle N \rangle = -\partial_\beta \ln(Z) = \frac{2e^{-\beta}}{\sqrt{1-4e^{-\beta}}(1-\sqrt{1-4e^{-\beta}})} - 1$$

We see this quantity is infinite for $a = 1/2$ but it is defined for $a > 1/2$. The total number of individual events (divisions plus deaths) is equal to : $N + (N+1) = 2N + 1$. This is what we have plotted on figure 2c.

0.6. Summary of different regimes

We generalize the previous theoretical results to the models with multiple cell types. Computing an extinction probability is still possible in this case: the only difference is that there will be n n-variable generating functions,

n being the number of types of cells. For example, in the simple one progenitor model that was used for figure 3 of appendix A, we will have:

$$f_0(s_0, s_1, s_2) = as_0^2 + (1-a)s_1^2$$
$$f_1(s_0, s_1, s_2) = as_1^2 + (1-a)s_2^2$$
$$f_2(s_0, s_1, s_2) = (1-d)s_2 + d$$

The fixed point equation to find the extinction probability is now a system of equations and it will have three solutions that we will call q_0, q_1, and q_2. The system is written as:

$$q_0 = aq_0^2 + (1-a)q_1^2$$
$$q_1 = aq_1^2 + (1-a)q_2^2$$
$$q_2 = (1-d)q_2 + d$$

We solve the system (with the condition $d > 0$) and we get:

$$q_2 = 1$$
$$q_1 = 1 \text{ or } \frac{1}{a} - 1$$
$$q_0 = 1 \text{ or } \frac{1}{a} - 1 \text{ or } \frac{1}{2a} - \frac{\sqrt{1 - \frac{4(1-a)^3}{a}}}{2a}$$

The correct value of each q depends on the value of a. We have to chose the smallest one under or equal to 1 (1 is always possible). If we call N_0, N_1, and N_2 the starting number of cells 0, 1 and 2, then the extinction probability is given by:

$$q = (q_0)^{N_0}(q_1)^{N_1}(q_2)^{N_2}$$

The probability to reach N_{max} is then given by $1 - q$. In the case of the figure 6a, we have $N_1 = N_2 = 0$ so the formula becomes:

$$p = 1 - (\frac{1}{2a} - \frac{\sqrt{1 - \frac{4(1-a)^3}{a}}}{2a})^{N_0} \text{ , if } a > 1/2 \text{ , else } 0$$

This is the equation of the theoretical curve that is plotted on figure 6a. We see a good fit with the data of the simulation. There is a little bit more error around $a = 1/2$ then in the one cell type model because of the large number of cells of different types that build up: they can still reach N_{max} for a a little bit below $1/2$.

The final figure is a detailed study of the possible growth of a population of cells that has stem cells, progenitors and differentiated cells (model of figure 3a of the main text with $n = 1$). Here again we assume that stem cells divide with probability a_s, progenitors with probability a_p, and differentiated cells die off with probability $1/2$. Our table shows all the different regimes of growth of the average number of cells S_t and P_t. In particular, we recover the polynomial regime when $a_s = a_p = 1/2$. We can also study the situation where we include different time scales for the different divisions. For instance, we can choose that stem cells divide at every time step, progenitors at every p time steps, and differentiated cells at every d time steps (in [2] the authors chose $p = 4$ and $d = 14$). Here we will consider the polynomial regime meaning we set $a_s = a_p = 1/2$ and we will show that we still have the same polynomial behavior even with the additional time scales. If we look at the average number of progenitor cells, considering that the process $P \to [(P + P) \text{ or } (D + D)]$ produces on average one progenitor cell, the increase of the average number of progenitors is only due to the division of stem cells into progenitors $S \to P + P$. Therefore, we will still have $P_t \sim t$ regardless of the addition of the time scale p. Now looking at differentiated cells, the same reasoning shows that there is an influx of differentiated cells from the division of progenitors which makes the population grow as $D_t \sim t^2$. The loss of differentiated cells from the death process does not change this polynomial growth. Therefore, we see that overall the total number of cells will still grow as $S_t + P_t + D_t \sim t^2$. This justifies the approach given in the main text of replacing all the different time scales by a single effective time scale.

For more details on the theory of branching processes and its applications to our models see [3], [4] and [5].

0.7. Adding noise to the parameters of the system

In this last section, we look at how the behavior of our system changes if we add some noise on the parameter a controlling cell division. We will work with the simple two-branch model of figure 1b of the main text. We first consider the case of uncorrelated noise. Let A be a random variable with law given by some probability density f on $[0,1]$ centered around $1/2$. The only requirements on f are therefore $\int_{a=0}^{1} f(a)da = 1$ and $f(a) = f(1-a)$. For each branching event (one division or death), we now choose the fate of the cell in the following manner: we first pick the parameter a according to the law of the random variable A, and then we decided the fate of the cell considering that it will divide with probability a and die with probability $1-a$. Let Z be the random variable corresponding to the number of descendents a cell gives birth to during a branching event. In our case we have:

$$E[Z] = 2\int_{a=0}^{1} f(a)ada = 1$$

$$E[Z^2] = 4\int_{a=0}^{1} f(a)ada = 2$$

Therefore we obtain $E[Z] = Var[Z] = 1$ which is the exact same result as for the case of $a = 1/2$ without any added noise. If we repeat this procedure for each branching event i with a variable A_i independent of all other A_j, then the average number of cells in the system remains unchanged and all the expressions of figure 4 still hold in this case.

Let us now take a quick look at the case of correlated noise. In this case it is possible to get completely different behaviors for the system. We consider again the simple branching model but where for half the time steps $a = 1/2 - \epsilon$ and for the other half $a = 1/2 + \epsilon$. We call N_0 the starting number of cells and N_t the average number of cells at time t. Here we have:

$$N_t = N_0(1+2\epsilon)^{\frac{t}{2}}(1-2\epsilon)^{\frac{t}{2}}$$

For $\epsilon \ll 1$, we get:

$$N_t \approx N_0 e^{-2\epsilon^2 t}$$

We apply this result to the method given in section III of the main text. We will choose $t = 50$ as this is roughly the number of time steps we need to apply our method (see figure 3c of the main text) and we choose $\epsilon = 0.05$ meaning $a = 0.5 \pm 0.05$. With these values we get $e^{-2\epsilon^2 t} \approx 0.778$ which is an error of roughly 20% on the number of cells. Considering that our method distinguishes well between the different types of growth it should still be possible to apply our results in this case. But for larger ϵ or longer t the behavior of the system can be completely different.

References

[1] Marek Kimmel and David E. Axelrod. *Branching Processes in Biology*. Springer, 2002.
[2] Gregory Driessens, Benjamin Beck, Amelie Caauwe, Benjamin D. Simons, and Cedric Blanpain. Defining the mode of tumour growth by clonal analysis. *Nature*, 488:527–531, 2012.
[3] Theodore E. Harris. *The Theory of Branching Processes*. Springer-Verlag, 1964.
[4] Tibor Antal and P. L. Krapivsky. Exact solution of a two-type branching process: Clone size distribution in cell division kinetics. *Journal of Statistical Mechanic*, 2010.
[5] Tibor Antal and P. L. Krapivsky. Exact solution of a two-type branching process: Models of tumor progression. *Journal of Statistical Mechanic*, 2011.