A minimum of two distinct heritable factors are required to explain correlation structures in proliferating lymphocytes.

John F. Markham[1], Cameron J. Wellard[2], Edwin D. Hawkins[3], Ken R. Duffy[4], Philip D. Hodgkin[2]

[1] Victorian Research Laboratory, National ICT Australia, Level 2, Building 193, Department of Electrical and Electronic Engineering, The University of Melbourne, Victoria 3010, Australia

[2] Immunology Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Victoria 3050, Australia

[3] Immune Signalling Laboratory, Cancer Immunology Research Program, Peter MacCallum Cancer Institute, St Andrews Place, East Melbourne, Victoria 3002, Australia

[4] Hamilton Institute, National University of Ireland Maynooth, Ireland

Keywords

Abstract

During the adaptive immune response, lymphocyte populations undergo a characteristic three phase process: expansion through a series of cell divisions; cessation of expansion; and, finally, most of the accumulated lymphocytes die by apoptosis. The data used, thus far, to inform understanding of these processes, both in vitro and in vivo, is taken from flow cytometry experiments. One significant drawback of flow cytometry is that individual cells cannot be tracked, so that it is not possible to investigate inter-dependencies in the fate of cells within a family tree. This deficit in experimental information has recently been overcome by Hawkins et al. (2009) who report on time lapse microscopy experiments in which B-cells were stimulated through the TLR9 receptor. Cells stimulated in this way do not aggregate, so that data regarding family trees can be recorded. In this article we further investigate the Hawkins et al. (2009) data. Our conclusions are striking: in order to explain the familial correlation structure in division times, death times and propensity to divide, a minimum of two distinct heritable factors are necessary. As the data shows that two distinct factors are necessary, we develop a stochastic model that has two heritable factors and demonstrate that it can reproduce the key features of the data. This model shows that two heritable factors are sufficient. These deductions have a clear impact upon biological understanding of the adaptive immune response. They also necessitate changes to the fundamental premises behind the tools developed by statisticians to draw deductions from flow cytometry data. Finally, they affect the mathematical modelling paradigms that are used to study these systems, as these are widely developed based on assumptions of cellular independence that are not appropriate.

Introduction

The reciprocal cellular processes of division and apoptosis combine to regulate biological

processes ranging from patterning body and tissue shape, to regulation and maintenance of the numbers of red blood cells, platelets, monocytes and lymphocytes in the blood. As a result of the ubiquity of this mechanism there is tremendous general interest in the regulation and simultaneous control of division and death. Investigators, from as early as the 1950s, have used film and microscopy to observe and measure the kinetics of cell division *in vitro* (Absher & Cristofalo 1984; Collyn-D'Hooghe 1977; Dawson et al. 1965; Minor & Smith 1974; Powell 1955). These studies, on a variety of cell types, all report that inter-mitotic division times show significant variation within clones of growing cells. Both quantitative and qualitative explanations were given to describe this variation. The influential Smith and Martin model proposed that variation originated from a stochastic regulator operating in an "A state" (assumed to be G1) that governed entry into a deterministic B phase (S, G2, and M) of the cell cycle (Smith & Martin 1973). Alternatively, size models implicated imprecise inheritance of cellular components regulating growth and replication as being responsible for differences in division times (Clifford & Sudbury 1972; Tyson & Diekmann 1986). The source of the inter-divisional variation or its significance is still not known. Cells undergoing apoptosis also show variation in times to die that are consistent with a stochastic internal process that is at least partly the result of a balance of anti- and pro-apoptotic molecules (Hawkins et al. 2007; Spencer et al. 2009). Similarly, little is understood about how control of division and apoptosis is related and how this relation impacts on control of cell populations in an immune response.

An excellent system for studying complex population shaping by regulated division and death is the adaptive immune response mounted by both T and B lymphocytes. At its core is the clonal expansion of lymphocytes of given specificity due to the appearance of antigen. During this response, B and T-cell populations undergo a characteristic three phase process: expansion through a series of cell divisions; cessation of expansion; and, finally, most of the accumulated lymphocytes die by apoptosis. Advances in flow cytometry and the discovery of non-interfering fluorescent dyes that act as cell labels have enabled the collection of experimental data on the kinetics of lymphocyte division progression and cell survival, e.g. (Lyons & Parish 1994; Parish 1999). These techniques yield high quality information at the level of populations. For example, use of the fluorescent dye carboxyfluorescein succinimidyl ester (CFSE) can provide a time-course for the number of live and dead lymphocytes and the fraction of cells that have undergone any given number of cell divisions. This data has strongly influenced immunological understanding. It has inspired statisticians to develop methodologies to study flow cytometry data, e.g. (Hyrien & Zand 2008), and provided information on which modelers have based their paradigms e.g.(Ganusov et al. 2005; Gett & Hodgkin 2000; Hawkins et al. 2007; Leon et al. 2004).

Data from these experiments are not, however, without their limitations. One significant drawback of flow cytometry data is that individual cells cannot be tracked, so that it is not possible to investigate dependencies in the fate of cells within a family tree. In the absence of this information, biologists, statisticians and modelers assume that all cells act as independent entities. This deficit in experimental information has recently been overcome by Hawkins et al. (2009) who report on time lapse microscopy experiments in

which B-cells were stimulated through the TLR9 receptor. Cells stimulated in this way undergo the usual population dynamics, dividing for 2-6 generations, but do not aggregate, so that extensive data regarding family trees can be observed and recorded.

In this article we detail a further investigation of the Hawkins et al. (2009) data. In order to explain the familial correlation structure in division times, death times and propensity to divide, a minimum of two distinct heritable factors are necessary. One factor regulates the propensity for a cell to divide and, if it does so, the time at which it divides. The other factor relates the propensity for cell division and the time taken to apoptosis. We then develop a stochastic model that has two heritable factors and demonstrate that it can reproduce the key features of the data. Thus the data shows that two distinct factors are necessary and the model shows that two are also sufficient. These deductions have important implications for mathematical modelling paradigms that are used to study these systems.

Results

The B Cell Data Set

Hawkins et al. (2009) has reported a data set derived from visual annotation of dividing primary naïve B lymphocytes stimulated using the Toll Like receptor-9 (TLR-9) ligand, CpG. The initial populations of cells and their progeny were cultured on Terasaki plates and followed for 120 hours using time lapse microscopy. Images were taken of the cells in 7 of the wells in each plate at a frequency of one per two minutes. Cell division was judged manually and cell death was judged by manual observation of propidium iodide uptake as a result of loss of membrane integrity upon apoptosis. Pedigrees of cells were followed from stimulation for up to 7 rounds of division, by which time nearly all cells had died. In total, 107 and 89 pedigrees were followed in two different experiments (Fam2 and Fam3 respectively) and times to die and divide for related cells recorded.

The data presented by Hawkins et al. (2009) are the first available for primary lymphocytes and the first individual cell tracking experiments to include substantial information concerning cell death times and division cessation. They noted a number of trends in the average behaviour of the cell population that gave some insight into cell operation and particularly the extent of inheritance. We first summarise these features and then report on new correlations that must be accommodated into any description of cellular inheritance. Then we present a physical model with a demonstrably minimum number of heritable factors that has the ability to reproduce these features.

Trends and Correlations in the data set

CpG-stimulated naïve B cells typically undergo a series of between one and six division rounds. The time to first division takes approximately 35 h while the more rapid subsequent divisions average 10 h, although the mean time increases by approximately 10-15% in the later division rounds. As noted for many other cell types, the times to divide are highly variable and, when plotted as a frequency histogram, follow a right skewed distribution. Hawkins et al. (2009) also noted a high degree of correlation in siblings' division times. Other reported features included the phenomenon of division

destiny where cells ceased to both grow and divide after 2-6 division rounds. These cells eventually died with the times to die also highly variable, with the mean time decreasing by approximately 25-35% in the later division rounds. The division destiny of progeny was heritable and strongly dependent on the original founder cell which Hawkins et al. (2009) illustrated using a heatmap to display division destiny of cell pedigrees. This effect comes about because the fate of siblings cells (that is, whether they divide or die) is highly correlated. Figure 1 presents a new quantitative representation of this relationship. The fate of siblings is broken down per division. In early divisions it is almost always observed that both siblings divide, while in later divisions it is almost always the case that neither sibling divides. Only in the middle phase of the response do we find siblings having different responses and even then this is in less than 20% of cases.

Figure 2 presents the correlations in division times for siblings and first cousins for one experimental set of results labelled Fam2 (the other data is qualitatively similar). Each is positively correlated (Figure 3 E and F). In order to check that this correlation is not simply due to the dependence of time to divide on number of divisions we looked at the correlations in subpopulations of cells which had undergone an identical number of divisions and found the same result (data not shown). Interestingly, the correlation between sibling times to divide is particularly strong at earlier division times as can be seen in Figure 2 A, where if we exclude siblings whose division times sum to less than 20 hours, the Pearson correlation coefficient (Sokal & Rohlf 1995), r, falls from 0.71 to 0.23.

Propensity to divide of related cells is strongly correlated

In the following discussion we use the term 'propensity to divide' to describe the likelihood of a cell to divide. In this experiment four different cell outcomes are observed: cells can be observed to undergo division or death, cells can be lost from view (around 17%) and a small number (2.5%) reach the end of the experiment alive. We assume that after sufficient time has elapsed, all cells will undergo one of two fates, division or death. We measure the correlation of fates of sibling cells by assigning the number 1 to division and 0 to death and measuring Pearson's r for these numbers. So, for example if siblings always had the same fate (that is if one divided then the other always divided or vice versa) then they would have $r = 1$. Conversely if the fate of sibling cells was independent then they would be uncorrelated and have $r = 0$. After doing so, we find that according to this method cell fate is strongly correlated between siblings and equal to 0.81 (0.76,0.86) for Fam2 and 0.87 (0.82,0.91) for Fam3 (95% confidence intervals in brackets). It is also a heritable property as demonstrated by the correlation between cousins' propensity to divide and also the clonal property whereby all cells in a clone loose their impetus to divide after approximately the same number of divisions (Hawkins 2009).

Propensity to divide is correlated to both time to divide *and* time to die

In Hawkins et al. (2009) it was shown that a heritable factor both increases the propensity to divide and shortens the time to divide. Here, we find a correlation between propensity to divide and time to *die*. Figure 4 A illustrates this by showing that cells whose siblings

divide, tend to die later than cells whose siblings die. These data lead us to conclude that a common factor influences both time to die and propensity to divide. The observation that time to die is correlated between siblings and cousins (Figure 4 B,C) suggests that such a factor is heritable.

One common factor is not sufficient to describe the data

If the putative common factor that regulates propensity to divide and time to divide is the same as the common factor that regulates propensity to divide and time to die then one might expect to observe a consistent negative correlation between time to die and time to divide for related cells. We looked for this in two places. First of all, there is a small subset of sibling cells which undergo different fates. Data from such siblings (Figure 5 A and Figure 3 G) shows that there is a small positive correlation although, as mentioned, the number of sibling cells with uneven fates is small. Secondly, we looked at the relationship between mother time to divide and daughter time to die (Figure 5 B) and found no significant correlation (see also Figure 3 H). This is strongly suggestive that at least two independent heritable factors are necessary to explain the data.

Modelling division times

Having established that at least two heritable factors are necessary to explain the data, we now demonstrate that two are sufficient for a mathematical model to reproduce the most significant features of the data. The features of the data can be divided into three categories: statistics describing division times, statistics describing death times and those describing fate determination. Our approach will be to work through each category in turn, developing a minimal model that can describe all the features in each category. At the end of the process we will have a unified, minimal model that can describe the relevant features of the data using two heritable factors. Our test for sufficiency will be to identify the important criteria in the various aspects of measured cell behaviour and to show how our two factor model can satisfy each one.

We start by looking at division times. Based on the above discussion we seek a model that can reproduce the following features of the data:

1) Right skewed distribution with a minimum division time of approximately 6 hours.
2) A trend of increasing t_{divide} as a function of generations.
3) Correlated t_{divide} for siblings and inheritance of t_{divide} from mother cells.
4) Correlation of t_{divide} for siblings being stronger for pairs of cells that divide earlier.

We found that models that divided the cell cycle into a series of steps with deterministic and exponential waiting times, such as the Smith-Martin transition probability model (Smith & Martin 1973) had difficulty reproducing the experimentally observed strong correlations at early division times (data not shown). In contrast we will show that a development of the modelling framework first proposed as Castor's G_1 Rate Model (Castor 1980) and Cooper's Continuum Model (Cooper 1982) can be adapted to re-create all of the desired properties listed above. While other models might be possible, we present here a detailed elaboration of a modified rate model to illustrate a projected

underlying biological mechanism.

A modified rate-based model

The G_1 rate model (Castor 1980) introduced the idea that the time taken for sibling cells to pass through G_1 is correlated and that this can be modelled by distributing the rate of passage as a bivariate normal distribution. Our reason for adopting this distribution, as will be revealed, is that it reproduces the observation that siblings that divide early are more highly correlated than those that divide late. Our approach will be to generalise the distribution so that it can be applied beyond sibling correlations and explain correlations between different generations of cells. Castor's model also contains a stochastic mechanism to explain the passage of cells through a second ($S+ G_2+M$) phase. We find this to be unnecessary for our purposes and replace it with a fixed time which we call $t_{\min}$. Hence we write the following form to describe the division time of a B lymphocyte:

$$t = t_{\min} + \frac{1}{r}.$$

If r is distributed normally and with a positive lower bound (justified below on physical grounds) then this gives a right skewed distribution with some minimum division time, $t_{\min}$, assumed to be constant for all cells. The correlation between time to divide and propensity to divide suggests that the quantity r is somehow associated with the ability to enter into division. Consequently we adopt a simple physical interpretation for r due to Cooper (1982) and postulate that r is proportional to the rate of synthesis of an initiating factor within each cell which, upon reaching a threshold level, triggers initiation of cell division (Figure 6 B). Events subsequent to this trigger can be thought of as taking time $t_{\min}$. If the concentration of this initiating factor is f then we can write

$$f(t) = mrt$$

where m is a constant that converts to units of concentration and t is the time since division. For the purpose of notational convenience in the equations and discussion that follow we set $m = 1$ and refer only to r, which has units of inverse time. Since r represents a rate of synthesis it must be positive and since the observed t has an upper bound r must be bounded from below. We speculate that r corresponds to an ensemble of contributing elements such as enzymes involved in signalling cascades and transcription factors regulating expression of essential proteins for growth. While the contributions from such elements may fluctuate over the course of the cell cycle, in order to keep our model as simple as possible, we take r to be constant during the accumulation of the initiating factor, f, but allow it to fluctuate at other times (that is between when f initiates division and when division actually occurs).

In order to introduce the correlations between the division times of two siblings and between the mother and daughter cells we must describe the manner in which each new cell takes on its value of r. It is clear from our conclusions above that there is a degree of sharing between siblings that is inherited from the mother, and that the inherited level is predominantly dictating division times of the two siblings, given the high level of correlation. We note that there is less correlation between mothers and daughters than there is between siblings suggesting that subsequent to division time being decided upon, but prior to division occurring, the value of r in the mother undergoes fluctuations which

are then passed on to both daughter cells. This is in contrast to fluctuations in r that occur subsequently in each daughter cell which contribute to differences in sibling division times. The process is illustrated in Figure 6 A. The point in the cell cycle where division time is decided is marked with an X. Fluctuations in r beyond this point do not affect division time for the cell, only its daughters. The value of r available at X is equal to the sum of three parts: (i) the amount that was available to the mother at X, r_m, (divided in two since it is split equally between daughters), plus (ii) variations in r that occurred in the mother after this point, r_{m-d} (also divided in two) plus (iii) independent variation in r in each daughter cell up to this point, r_{d_1} and r_{d_2}. We can write this formally as follows:

$$t_1 = t_{\min} + \frac{1}{\frac{1}{2}(r_m + r_{m-d}) + r_{d_1}}$$

$$t_2 = t_{\min} + \frac{1}{\frac{1}{2}(r_m + r_{m-d}) + r_{d_2}}$$

where t_1 and t_2 are the division times for sibling cells and r_m is the portion of r that

contributed to the division time of the mother cell. That is, $t_m = t_{\min} + \dfrac{1}{r_m}$.

where t_m is the division time for the mother cell and r_m is responsible for the inherited component of division time since it contributes to the division times of both mother and daughters.

r_{m-d} is a normally distributed random number, $r_{m-d} \sim N(\mu_{m-d}^{division}, \sigma_{m-d}^{division})$. It gives the difference between mother and daughters which is common to both siblings.

r_{d_1} and r_{d_2} are normally distributed random numbers with zero mean, generated independently for each cell according to $r_{d_1}, r_{d_2} \sim N(0, \sigma_d^{division})$. These are responsible for the difference between sibling division times.

For the initial generation of cells we choose r_m to be normally distributed according to $r_m \sim N(\mu_m^{division}, \sigma_m^{division})$. On rare occasions total r can be very small or negative, resulting in unphysical division times. To prevent this, the distribution is truncated (see supplementary information for details).

We observe that cells with a large $r_m + r_{m-d}$ tend to divide earlier. For such cells the noise from r_{d_1} and r_{d_2} will be proportionately less, hence they will be more correlated, satisfying our initial criterion that siblings dividing earlier be more highly correlated. The fact that siblings are more correlated than mother-daughter pairs suggests that $\sigma_d^{division} < \sigma_{m-d}^{division}$ (see Table 1). In other words most of the noise on r (and hence the physical quantity that it represents) is picked up between when division time is decided and when division actually occurs. One can speculate on the source of the noise, but suffice it to say that if the physical quantities that determine r are produced and subject to imperfect regulation then one would expect it to accumulate fluctuations over time (Sigal et al. 2006).

If $\sigma_{m-d}^{division}$ and $\sigma_d^{division}$ are tuned to give agreement with correlations between mother – daughter and sibling-sibling correlations then we can predict the observed cousin-cousin correlations (Figure 3 D). This demonstrates that the model for division times captures salient features of the data.

Modelling death times

The above modified rate-based model recreated the key correlations in division times transmitted through generations. As noted, correlations, albeit weaker, are also found in death times through generations. Using the same approach, we seek the simplest model that can reproduce the key features of the data relating to the inheritance of death times. These features are as follows:

1) Right skewed distribution.
2) A trend to decreasing t_{die} in later generations.
3) Correlated t_{die} for siblings and inheritance as demonstrated by correlation of cousin death times.
4) Unlike t_{divide}, t_{die} isn't strongly correlated at small values.
5) t_{divide} is independent of t_{die}. While Figure 5 A and B show some correlation between division and death times, the magnitude is small and, the signs are conflicting. This suggests that the two may be modelled as independent processes.

Here again we postulate that components making up the survival machinery of the cell are partly inherited from the mother, and partly made independently and *de novo*, in each new cell. We propose as the simplest case, and in absence of further information, that the quantity of the factor controlling time taken to apoptosis is directly proportional to the time to die of the cell. Thus, we chose a minimal generalisation of the Cyton model mechanism for death (Hawkins et al. 2007).

$$t_1^{death} = t_m^{death} k_{d_1}$$

$$t_2^{death} = t_m^{death} k_{d_2}$$

where t_m^{death} is the death time carried by the mother cell, which we identify as proportional to the level of our second common factor. Clearly, if this was passed on, the mother did not die. For undivided cells we assume a lognormal distribution in the population of the factor and therefore, lognormally distributed times to die as advocated in (Hawkins et al. 2007), that is

$$\log\left(t_m^{death}\right) \sim N(\mu_{t_m}^{death}, \sigma_{t_m}^{death}).$$

k_{d_1} and k_{d_2} are components which produce independent variations for each daughter and are distributed according to $\log(k_{d_1}), \log\left(k_{d_2}\right) \sim N(\mu_{k_{d_n}}^{death}, \sigma_{k_{d_n}}^{death})$

The parameter $\mu_{k_{d_n}}^{death}$ was chosen to give the correct trend in death times with generation while $\sigma_{k_{d_n}}^{death}$ was selected to match the correlation between siblings that is observed in the data (see Figure 3 E). In contrast to division times we note that we require no analogue to r_{m-d} suggesting that for time to die, all of the variation in t_{die} subsequent to division arises independently in the daughter cells.

Linking fate determination to division time

In developing a model to explain the connection between fate determination and division time we need to keep in mind the growth data from (Hawkins 2009) that suggests that cell fate is determined at, or soon after, cell division. The early determination of cell fate, plus the high correlation of cell fate as previously described here, suggests that perhaps only inherited material from the mother, $r_m + r_{m-d}$, need be used. Even so, approximately 10% of cells undergo different fates, so a further random component specific to each sibling is required. One option would be to include r_{d_1} and r_{d_2}, that is select cell fate based on the total value of r. But this would lead to a sharp cut off in the distributions for division time, which is something the data does not support. We propose a stochastic process acts on cells to produce uneven fates. Looking at the division times for cells whose sibling died (Hawkins 2009) suggests that this mechanism acts only on cells with long division times, that is when cells are on the cusp of being able to divide. To summarise, if $r_m + r_{m-d}$ is more than r_{high}, then both siblings divide. If $r_m + r_{m-d}$ is less than r_{low}, then both siblings die. If $r_m + r_{m-d}$ is in between r_{low} and r_{high}, then each cell has a stochastic outcome dependent on value of r that it has associated with it. In this intermediate region we propose that the probability of division increases linearly with $r_m + r_{m-d}$ to effect a smooth transition. Thus, we model the probability of division of each cell as:

$$\Pr(divides) = \begin{cases} 0, & r_m + r_{m-d} < r_{low} \\ \dfrac{r_m + r_{m-d} - r_{low}}{r_{high} - r_{low}}, & r_{low} \leq r_m + r_{m-d} \leq r_{high} \\ 1, & r_m + r_{m-d} > r_{high} \end{cases}$$

Supplementary Figure 1 B, C shows that this model gives the correct qualitative relationship between propensity to divide and t_{divide}. Furthermore, because only the inherited component and not the individual components of r (r_{d_1} and r_{d_2}) are used to decide whether a cell can divide, the propensity of sisters to divide can have a similar correlation to time to divide, as is observed in the data. Because only half of r is passed onto the daughter cells at division and because r_{m-d} can, on average, be negative the average r of a population of cells is depleted over successive generations. As this depletion occurs the distribution of r for the overall cell population will pass through the region between r_{low} and r_{high}. As it does so the proportion of cells dividing and the correlation between fates for sisters both have the correct qualitative form. That is, early in the response most cells divide; uneven cell fate is most likely to occur mid-response and late in the response, most cells die (see Figure 6 C,D). Finally, the distribution of r in the founder population and the subsequent preservation of relative levels in descendents that arises from the proposed mechanism leads to the strong founder effect for division destiny (Hawkins 2009).

Linking fate determination to death time

The model has many of the features sought, however, as it presently stands, it will not

show any dependence between times to die and propensity to divide as shown in Figure 4. One reasonable way to include this is to make the probability of division dependent on a functional combination of t_{divide} and t_{die} (or components thereof). Unfortunately this tends to lead to unphysical t_{divide} and t_{die}. Removing such unphysical subpopulations by fiat leads to unintended correlations between the two times (data not shown). To avoid these two problems we instead make a minor change to the stochastic cutoff procedure above to incorporate the inherited component of death time, t_m^{death}. The fate of cells with intermediate values of $r_m + r_{m-d}$, that is, where $r_{low} < r_m + r_{m-d} < r_{high}$, now depends on the value of this second common factor. Cells with this intermediate amount of r can be thought of as being on the cusp of being assigned to either death or division. Such cells are sensitive to a second signal to decide cell fate; a signal (in the form of a high t_m^{death}) that causes them to both die. Alternatively, if such cells are not sufficiently committed to death (by having a low t_m^{death}) then they are assigned different fates as before. Formally,

$$\Pr(divides) = \begin{cases} 0, & r_m + r_{m-d} < r_{low} \\ \dfrac{r_m + r_{m-d} - r_{low}}{r_{high} - r_{low}}, & r_{low} \leq r_m + r_{m-d} \leq r_{high} \ and \ t_m^{death} \cdot k_{m-d} \geq t_{\min}^{death} \\ 0, & r_{low} \leq r_m + r_{m-d} \leq r_{high} \ and \ t_m^{death} \cdot k_{m-d} < t_{\min}^{death} \\ 1, & r_m + r_{m-d} > r_{high} \end{cases} .$$

We find that setting
$$t_{\min}^{death} = \mu_{t_m}^{death} - \sigma_{t_m}^{death}$$
gives us the correct qualitative dependence of time to die on propensity to divide (Supplementary Figure 1A) while preserving the dependence of time to divide on propensity to divide (Supplementary Figure 1 B,C). It also points to the existence of a subpopulation of cells on the verge of not being able to progress further through divisions, which are sensitised to other signals to trigger fate selection.

Discussion

The recent data of Hawkins et al. (2009) recorded the correlation in times to divide and die in B cells following stimulation with CpG that leads to proliferation, eventual cessation after a varying number of division rounds, and then death. Striking familial correlations were observed that we reasoned could provide a unique insight into the source of shared and randomised components of cell fate in this system. To facilitate this goal we sought a minimal model that could reproduce the important features of the data. We found that a minimum of two different heritable factors were necessary to explain the correlation structure in the data. We then showed that two heritable factors were sufficient by constructing a model postulating one heritable factor that controls a cell's time to divide and another which regulates time to apoptosis. In our model both factors play a role in determining cell fate. We find the data is consistent with a mechanism

where the value of the first factor varies the rate of accumulation of downstream cellular mediators(s) that trigger cell division when a threshold is reached. Stochastic variation in parts of this mechanism, either in the level of the inherited factor, the level of accumulated mediators or selection of the threshold level for triggering division contributes to deciding cell fate (either division or death). While we make no presumptions about the particular physical mechanism involved, if we assume the second factor also modulates this stochastic fate selection then we can correctly describe the regulation and correlation between all three aspects of cell number regulation without the need for any further heritable components.

One of the striking features of the Hawkins et al. (2009) data is the extreme correlation of propensity to divide. On average this is more strongly correlated than any other quantity between siblings. In order for propensity to divide to be more strongly correlated than time to divide, it had to be derived from the common component influencing time to divide inherited from the mother cell. This feature did not include the additional random component of time to divide that contributed independently to division times of each sibling. In other words, time to divide picked up more randomness at or after division whereas division and propensity to divide did not. This is consistent with the hypothesis that cell fate is decided at, or soon after, division as was suggested by examination of cell size in (Hawkins 2009).

The tracking of cell lineages was undertaken by Hawkins et al. (2009) to highlight the source and nature of variation in lymphocyte regulation. Numerous prior models have been proposed to describe lymphocyte proliferation and survival, although few accommodate such strong lineage affiliations as revealed in this new data set. Furthermore, most models interleave cell division and death by assuming an age-independent time to die that is inconsistent with the pattern of death observed in these data. An earlier paper by Hawkins et al. (2007) proposed the cyton model based on the hypothesis that times to divide and die were independent, and acted in competition, with each being clocked from their last division and following some skewed right probability. Here, our minimal model also assumes that division and death are clocked from mitosis. However, in contrast to the cyton model, we assume cells decide their fate, either division or death, early after a division, and that time to the chosen fate, is then regulated. This simpler model is possible because we recorded no instance of cells dying during a growth phase leading to a cell division. Rather, only cells that lost the impetus to grow went on to die. The cyton model is capable of reproducing this behaviour by having a distinct time to die distribution for cells that have undergone division destiny (Subramanian et al. 2008). The generality of this and other methods for incorporating death in useful biologically relevant mechanisms will only become apparent when additional regulated cell systems are followed in a similar manner to the CpG stimulated B cells studied here.

We do not rule out the possibility that there may exist other common factors than the ones proposed. Nor do we rule out the possibility that a model with more degrees of freedom might give outcomes that agree more closely with observations. For example, the model could be extended to give better agreement with correlations between distantly related cells in a pedigree. However in order to explain a subset of the observed

correlations in our data we already require a significant number of degrees of freedom; each correlation needs to be parameterised, as does division linked behaviour. The decision on how far to go down the path of increasing model complexity to fit data is based on whether doing so adds insight into the system or utility. In this case formulating a model to describe the operation of two common factors leads to the discovery that one of the common factors only acted on cells that were sensitised to respond. That is the common factor for death could only affect a subset of the total cell population. Beyond the insight gained, a question that remains unanswered is "do we need to use these multivariate models to analyse population experiments?" In responses that are limited to relatively few division cycles, existing univariate models are sufficient for the purposes of reproducing the mean population sizes. However, as the number of division cycles increases, the effect of correlations in division time between parents and their progeny on the mean population dynamics increases, and it becomes necessary to use a model that accounts for this correlation (Wellard et al. 2009). Furthermore, studies using branching process analyses (Crump & Mode 1969; Duffy & Subramanian 2009; Wellard et al. 2009) suggest that correlation in time to divide and in cell fate necessarily leads to increased variability of total cell numbers. For simulation of small populations of cells this can impact on the results. For example fluctuations in the numbers of a small clone of lymphocytes could result in its extinction. A study using a model that incorrectly implements correlation between cell fates would be unable to capture this behaviour. The conclusion to be drawn from this is that the choice of model depends on the application. Utility would suggest that for systems of thousand of cells the simplest univariate model as measured by ease of solution and goodness of fit should be used. For clones of tens of cells, as can exist at the beginning of an immune response, the model needs to capture correlated behaviour and the solution method needs to be able to able to calculate the fluctuations about the expected mean behaviour.

Finally, our modelling approach leads to testable biological hypotheses and suggests directions for future investigation. Our study suggests that it would be fruitful to search for cell surface, cytoplasmic or nuclear proteins diluting and varying from generation to generation that are involved in triggering both division and death.   This could be achieved by proteomic analysis, which makes no assumptions, or a candidate investigation of likely cell cycle and cell death regulators. As the expression of putative factors appears to be required primarily in the first division, and less so in subsequent divisions, high throughput sequencing of RNA (Mortazavi et al. 2008) might be used to compare RNA expression levels for cohorts of cells from consecutive divisions to identify candidate transcripts with these features.   Once the diluting elements controlling the division and death times are identified, expression as fluorescent tagged fusion molecules would allow further time lapse microscopy experiments to correlate and monitor the stochastic inheritance and re-synthesis in each generation predicted here. The existence of heritable factors regulating the propensity to divide with times to divide and die raises the prospect that there may exist more heritable factors, changing with division, that regulate other aspects of the immune response. It has been shown that differentiation decisions for both B lymphocytes and T lymphocytes alter with successive division rounds (Bird et al. 1998; Gett & Hodgkin 1998; Hodgkin et al. 1996). We speculate that a similar quantitative approach applied to following alternative fates may

be able to provide further insight into regulatory mechanisms of the immune response and the control of the rapid emergence of cellular heterogeneity.

Tables

parameter	value	parameter	value
$\mu_m^{division}$	0.9	$\mu_{t_m}^{death}$	40
$\sigma_m^{division}$	0.02	$\sigma_{t_m}^{death}$	15
$\mu_{m-d}^{division}$	-0.03	$\mu_{k_{d_n}}^{death}$	0.9
$\sigma_{m-d}^{division}$	0.04	$\sigma_{k_{d_n}}^{death}$	0.3
$\sigma_d^{division}$	0.02	$t_{threshold}$	10
r_{high}	1/12	t_{max}	25
r_{low}	1/22	$n_{InitialCells}$	40
t_{fixed}	4		

Table 1 Quantities used to generate simulated data from multivariate model

Acknowledgements

References

Absher, M. & Cristofalo, V. J. 1984 Analysis of cell division by time-lapse cinematographic studies of hydrocortisone-treated embryonic lung fibroblasts. *J Cell Physiol* **119**, 315-9.

Bird, J. J., Brown, D. R., Mullen, A. C., Moskowitz, N. H., Mahowald, M. A., Sider, J. R., Gajewski, T. F., Wang, C. R. & Reiner, S. L. 1998 Helper T cell

differentiation is controlled by the cell cycle. *Immunity* **9**, 229-37.

Castor, L. N. 1980 A G1 rate model accounts for cell-cycle kinetics attributed to 'transition probability'. *Nature* **287**, 857-9.

Clifford, P. & Sudbury, A. 1972 The Linear Cell-Size-Dependent Branching Process. *Journal of Applied Probability* **9**, 687-696.

Collyn-D'Hooghe, M., Valleron, A.-J. & Malaise, E. P. 1977 Time-lapse cinematography studies of cell. cycle and mitosis duration. *Expl Cell Res.* **106**, 405-407.

Cooper, S. 1982 The continuum model: statistical implications. *J Theor Biol* **94**, 783-800.

Crump, K. S. & Mode, C. J. 1969 An age-dependent branching process with correlations among sister cells. *J. Appl. Prob.* **6**, 205-210.

Dawson, K. B., Madoc-Jones, H. & Field, E. O. 1965 Variations in the Generation Times of a Strain of Rat Sarcoma Cells in Culture. *Exp Cell Res* **38**, 75-84.

Duffy, K. R. & Subramanian, V. G. 2009 On the impact of correlation between collaterally consanguineous cells on lymphocyte population dynamics. *J Math Biol* **59**, 255-85.

Ganusov, V. V., Pilyugin, S. S., de Boer, R. J., Murali-Krishna, K., Ahmed, R. & Antia, R. 2005 Quantifying cell turnover using CFSE data. *J Immunol Methods* **298**, 183-200.

Gett, A. V. & Hodgkin, P. D. 1998 Cell division regulates the T cell cytokine repertoire, revealing a mechanism underlying immune class regulation. *Proc Natl Acad Sci U S A* **95**, 9488-93.

Gett, A. V. & Hodgkin, P. D. 2000 A cellular calculus for signal integration by T cells. *Nat Immunol* **1**, 239-44.

Hawkins, E. D., Markham, J. F., McGuinness, L.P. and Hodgkin, L. P. 2009 A single-cell pedigree analysis of alternative stochastic lymphocyte fates. *Proc Natl Acad Sci U S A*.

Hawkins, E. D., Turner, M. L., Dowling, M. R., van Gend, C. & Hodgkin, P. D. 2007 A model of immune regulation as a consequence of randomized lymphocyte division and death times. *Proc Natl Acad Sci U S A* **104**, 5032-7.

Hodgkin, P. D., Lee, J. H. & Lyons, A. B. 1996 B cell differentiation and isotype switching is related to division cycle number. *J Exp Med* **184**, 277-81.

Hyrien, O. & Zand, M. S. 2008 A Mixture Model with Dependent Observations for the Analysis of CFSE-Labeling Experiments. *J American Statistical Association* **103**, 222-239.

Leon, K., Faro, J. & Carneiro, J. 2004 A general mathematical framework to model generation structure in a population of asynchronously dividing cells. *J Theor Biol* **229**, 455-76.

Lyons, A. B. & Parish, C. R. 1994 Determination of lymphocyte division by flow cytometry. *J Immunol Methods* **171**, 131-7.

Minor, P. D. & Smith, J. A. 1974 Explanation of degree of correlation of sibling generation times in animal cells. *Nature* **248**, 241-3.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. 2008 Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-8.

Parish, C. R. 1999 Fluorescent dyes for lymphocyte migration and proliferation studies. *Immunol Cell Biol* **77**, 499-508.

Powell, E. 1955 Some features of the generation times of individual bacteria. *Biometrika*

42, 16-44.

Sigal, A., Milo, R., Cohen, A., Geva-Zatorsky, N., Klein, Y., Alaluf, I., Swerdlin, N., Perzov, N., Danon, T., Liron, Y., Raveh, T., Carpenter, A. E., Lahav, G. & Alon, U. 2006 Dynamic proteomics in individual human cells uncovers widespread cell-cycle dependence of nuclear proteins. *Nat Methods* **3**, 525-31.

Smith, J. A. & Martin, L. 1973 Do cells cycle? *Proc Natl Acad Sci U S A* **70**, 1263-7.

Sokal, R. R. & Rohlf, F. J. 1995 *Biometry: the principles and practice of statistics in biological research*: W. H. Freeman and Co.: New York.

Spencer, S. L., Gaudet, S., Albeck, J. G., Burke, J. M. & Sorger, P. K. 2009 Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature* **459**, 428-32.

Subramanian, V. G., Duffy, K. R., Turner, M. L. & Hodgkin, P. D. 2008 Determining the expected variability of immune responses using the cyton model. *J Math Biol* **56**, 861-92.

Tyson, J. J. & Diekmann, O. 1986 Sloppy size control of the cell division cycle. *J Theor Biol* **118**, 405-26.

Wellard, C. J., Markham, J. F., Hawkins, E. D. & Hodgkin, P. D. 2009 Correlations and fluctuations in the proliferation of B cells: A branching process approach. *Submitted for publication.*

Figure captions

Figure 1 Trends in cell fate broken down per division for experiments fam2 and fam3. In both cases the proportion of siblings undergoing different fates is a maximum mid-response.

Figure 2 Correlation of times to divide for related cells (fam2). (A) Times to divide for siblings are more highly correlated than for first cousins (B). Siblings whose division times sum to more than 20 hours (above and to the right of the red line in (A)) are less correlated (r=0.23) than the population as a whole (r=0.71).

Figure 3 Measured correlation (Pearson r) from experiment compared with simulation. Illustration shows the correlations being measured by heavy dashed lines. Figures within brackets show the 95% confidence interval obtained using a bootstrapping method as described in supplementary information.

Figure 4 Correlation of times to die for related cells (fam2) (A) When both siblings die, they tend to do so earlier than cells whose sibling divide (p<0.0001 for median value, see supplementary information for details). (B) Times to die for sisters are more highly correlated than for (C) cousins.
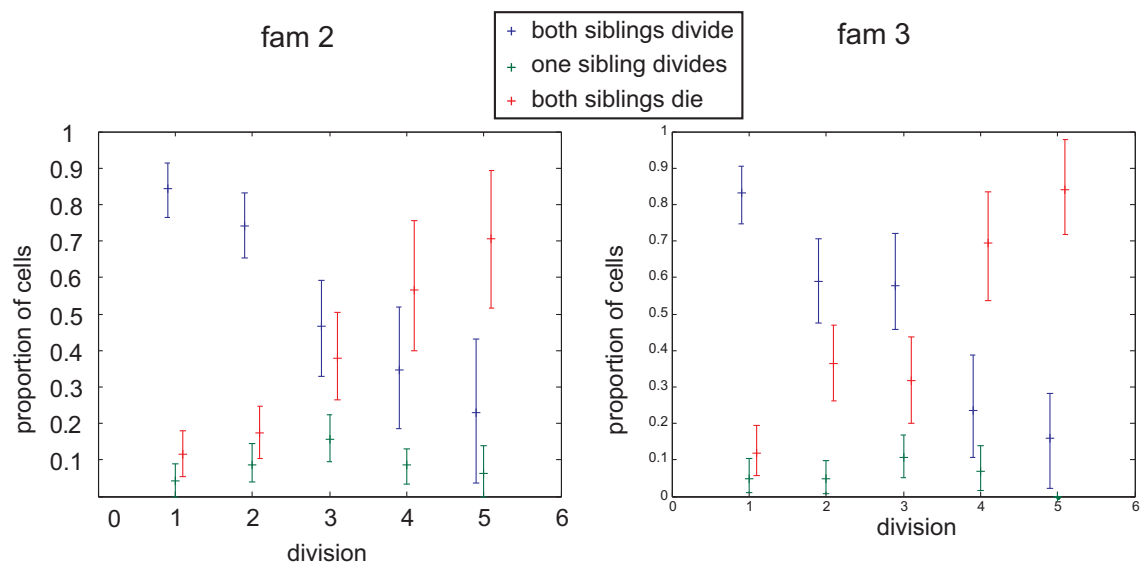
Figure 5 Correlation between times to divide and die (fam2). (A) Sisters which undergo different fates show a small amount of positive correlation between their respective times to die and divide. (B) There is no significant correlation (refer Figure 3 G) between when
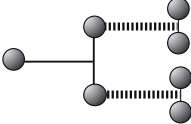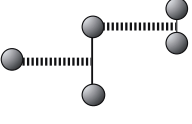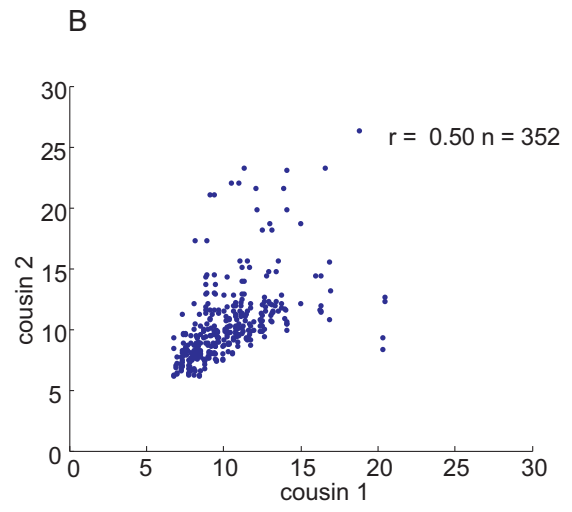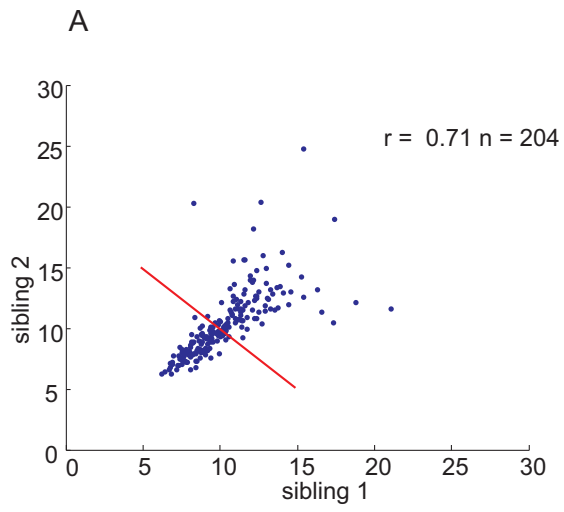
a cell divides and when its offspring die.

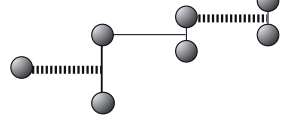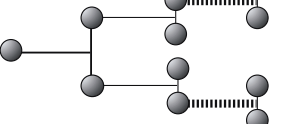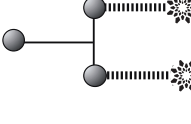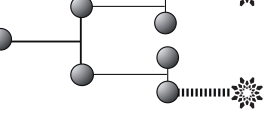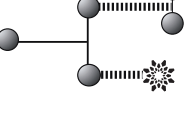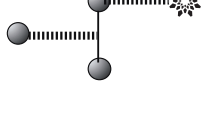*Figure 6 (A) Time line showing mechanism for inheritance of division time. Division time is determined by the value of r, shown here as r_X, at the times marked X. For the mother cell this is r_m, and so its division time is given by $t_m = t_{\min} + \dfrac{1}{r_m}$. Between this time and division occurring, the mother cell increases its value of r by an amount, r_{m-d}. At division, the total value of r, $r_m + r_{m-d}$, is assumed to be split evenly between the daughter cells. Some time between division and the point marked X, each daughter cell acquires a further independent contribution to r called r_{d_n} where $n = 1, 2$. As with the mother cell, the daughters now hav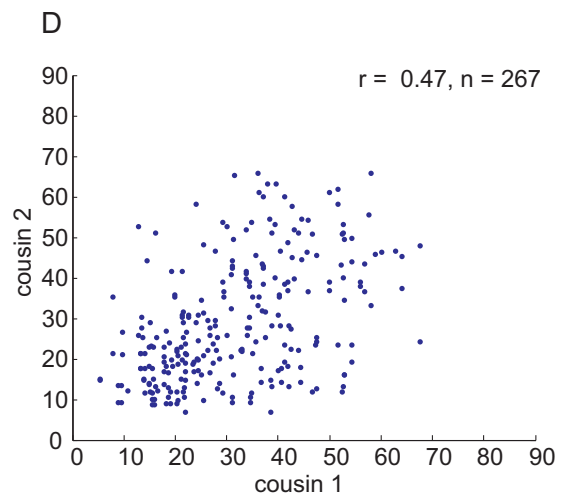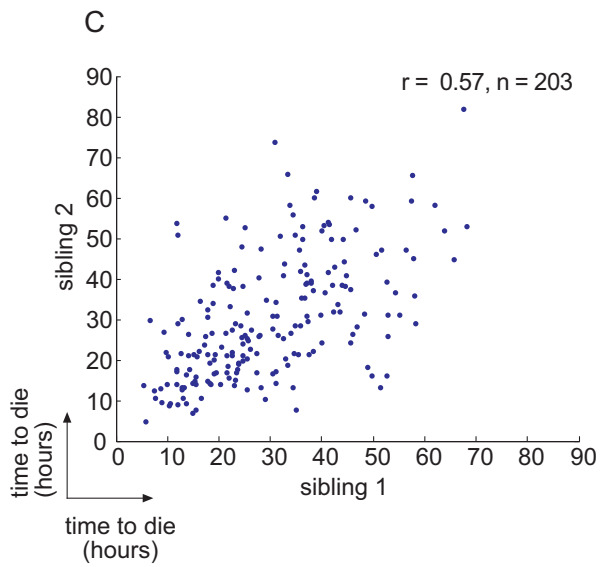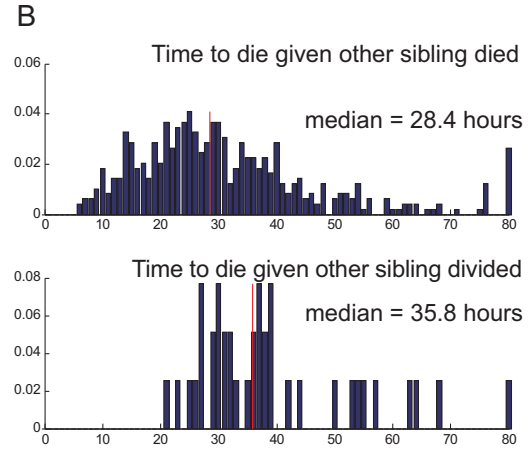e divisio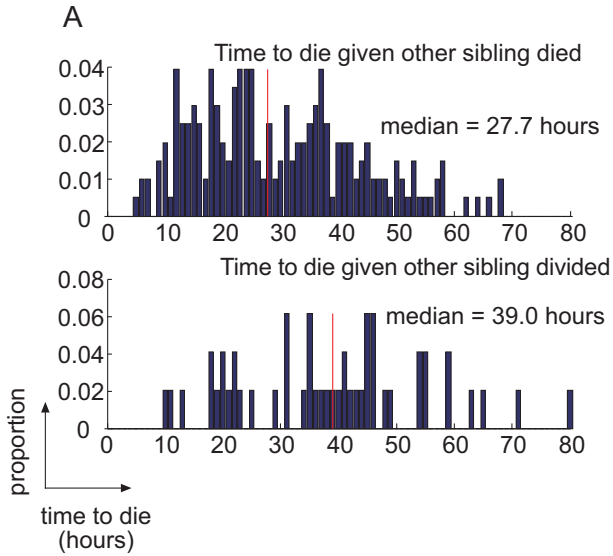n times determined by their value of r at X which is now $t_{d_n} = t_{\min} + \dfrac{1}{\frac{1}{2}\left(r_m + r_{m-d}\right) + r_{d_n}}$*
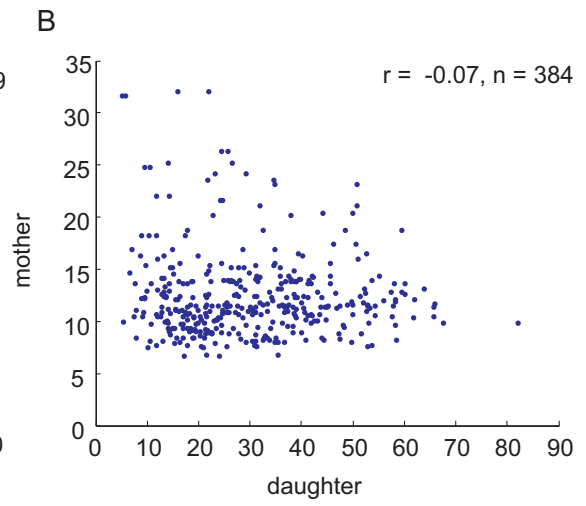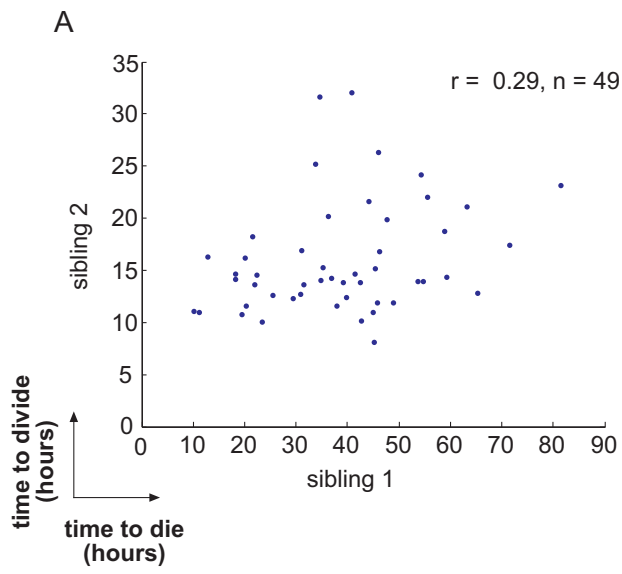
(B) Right skewed distribution arises from a rate-and-comparator model for replication. The quantity f is accumulated according to $f(t) = mrt$, where m is some constant of proportionality with units of f and r is a rate with units inverse time. Upon reaching a threshold value, $f_{threshold}$, division machinery is initiated. At this point $t_{threshold} = \dfrac{f_{threshold}}{mr}$ and so there is an inverse relationship between the value of r and the variable component of division time, $t_{threshold}$. If r is distributed normally according to $N(r)$ as shown in the vertically oriented distribution then $t_{threshold}$ is distributed with an inverse-normal pdf, drawn horizontally. The shaded areas show a cohort of cells with a particular range of values r and $t_{threshold}$. (C) The value of r at first division is high and so nearly all cells divide. As cells go through consecutive division cycles the quantity r gradually diminishes until they are unable to divide further. (D) Mid-response, the value of r passes through a narrow region in which stochastic fate selection can occur. Here, a second factor controlling cell death is also able to influence the probability of division.
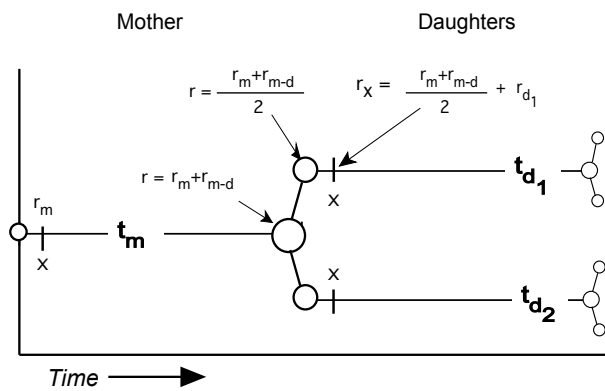
A

B

	A $r^{division}_{s\text{-}s}$	B $r^{division}_{m\text{-}d}$	C $r^{division}_{gm\text{-}gd}$	D $r^{division}_{cous\text{-}cous}$
Fam 1	0.71 (0.57, 0.85)	0.51 (0.32, 0.71)	0.39 (0.16, 0.59)	0.43 (0.21, 0.67)
Fam 2	0.81 (0.68, 0.90)	0.56 (0.42, 0.70)	0.23 (0.02, 0.47)	0.58 (0.43, 0.74)
Sim	*0.77 (0.70, 0.84)*	*0.59 (0.49, 0.68)*	*0.17 (0.00, 0.34)*	*0.46 (0.28, 0.62)*

	E $r^{death}_{s\text{-}s}$	F $r^{death}_{cous\text{-}cous}$	G $r^{division\text{-}death}_{s\text{-}s}$	H $r^{division\text{-}death}_{m\text{-}d}$
Fam 1	0.59 (0.47, 0.70)	0.48 (0.34, 0.59)	0.30 (0.08, 0.52)	-0.08 (-0.21, 0.06)
Fam 2	0.66 (0.55, 0.76)	0.57 (0.39, 0.70)	0.25 (0.01, 0.49)	-0.18 (-0.30, -0.05)
Sim	*0.75 (0.65, 0.85)*	*0.53 (0.36, 0.71)*	*0.01 (-0.52, 0.61)*	*0.09 (-0.02 0.22)*

A

Time to die given other sibling died

median = 27.7 hours

Time to die given other sibling divided

median = 39.0 hours

proportion

time to die
(hours)

B

Time to die given other sibling died

median = 28.4 hours

Time to die given other sibling divided

median = 35.8 hours

C

r = 0.57, n = 203

sibling 2

time to die
(hours)

sibling 1

time to die
(hours)

D

r = 0.47, n = 267

cousin 2

cousin 1

A

Mother   Daughters

$r = \dfrac{r_m + r_{m-d}}{2}$ $r_x = \dfrac{r_m + r_{m-d}}{2} + r_{d_1}$

$r = r_m + r_{m-d}$

r_m

t_m

t_{d_1}

t_{d_2}

x

Time

B

Time to trigger division

t_{min}

Division

f_{thresh}

r_x

r_{high} r_{low}

Divide   Divide or Die   Die

f

Time

C

Generation: 0   1   2   3

r

$< r_{low}$ - no division

Time   Division   Division   Division

D

2nd factor influence
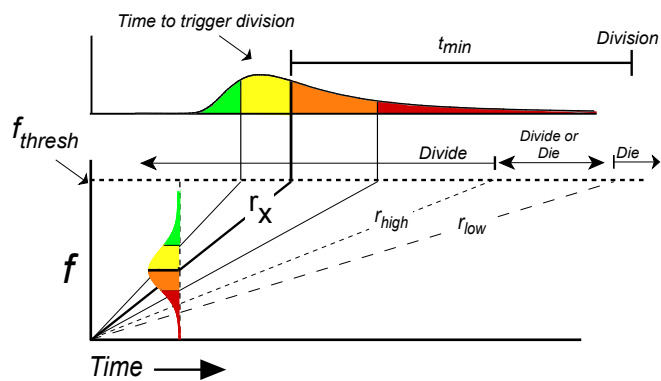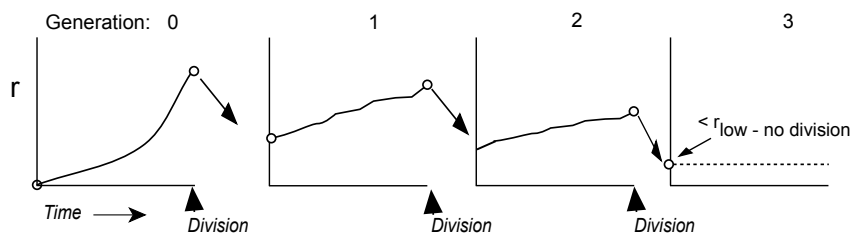
Pr(div)

1

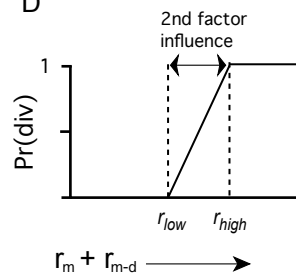r_{low} r_{high}

$r_m + r_{m-d}$

Supplementary Information

Bounding division time

With the scheme described it is still possible that a few cells with small or even negative r will be assigned to division, leading to negative or large t_{divide}. We correct for this by using a logistic function to compress values at small r so that it is bounded from below (and hence t_{divide} is bounded from above). Before using r to find t_{divide} we apply the following mapping:

$$r_{out} = \begin{cases} \dfrac{1}{k_{out}} \cdot \dfrac{1}{1+e^{-k_{in}(r_{in}-a)}} + b, & r_{in} < a \\ r_{in}, & r_{in} \geq a \end{cases}$$

with

$$a = \frac{1}{t_{threshold}}$$

$$b = \frac{1}{t_{max}}$$

$$k_{out} = \frac{1}{2 \cdot (a-b)}$$

$$k_{in} = 4 \cdot k_{out}$$

where $t_{threshold}$ is the point at which r_{out} begins to deviate from r_{in}, t_{max} is effectively a cutoff on the maximum allowed time to divide

a and b are chosen so that $\dfrac{dr_{out}}{dr_{in}}$ is continuous. This logistic function is applied

selectively to r when it is smaller than $\dfrac{1}{t_{threshold}}$.

Agent based simulations of the proposed model

Our models were implemented using an agent-based approach with the parameters in Table 2. Simulations were repeated 500 times and results with 95% confidence intervals plotted in Table 1.

Approximation for loss

During the time lapse microscopy experiments some cells were lost during tracking (Hawkins 2009). Mostly this occurred soon after division. In our simulations we have approximated loss by eliminating the same proportion of cells in each division as were lost in the experiments. Cells are modelled as being lost at division. No attempt is made

to simulate the distribution of times at which cells are lost.

Confidence intervals

In all scatter plots, the reported correlations were calculated using the function *corr()* from MATLAB Version 7.5.0.342 (R2007b) from the Mathworks company.

Confidence intervals on experimental quantities in all figures and tables unless stated otherwise are at the 95% level and computed using 500 bootstrap replicants formed by selecting *pedigrees* of cells. As in (Hawkins 2009) it is assumed that pedigrees are independently and identically distributed.

The p-value in Figure 4 was calculated using a permutation test drawn from 10^6 re-sampled data sets.
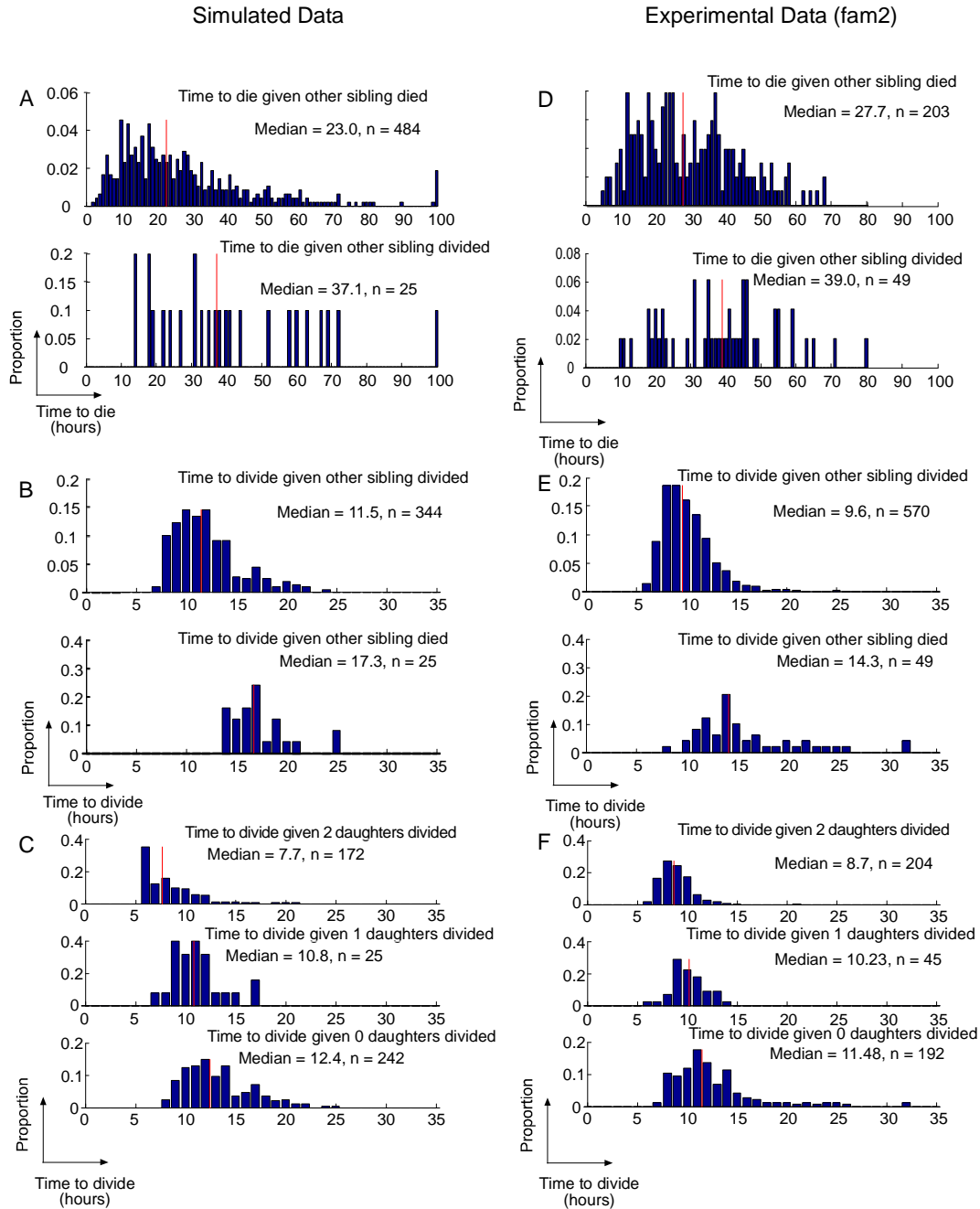
Supplementary Figure 1 Simulated data generated using a model with two common factors reproduces the correct relationship between propensity to divide and times to die and divide. (A) Cells whose siblings die tend to die earlier than cells whose siblings divide. (B) Cells whose siblings divide tend to divide earlier than cells whose siblings die. (C) Cells divide earlier when more of their daughters divide. (D), (E) and (F) show the corresponding experimental data for fam2. (E) and (F) reproduced from (Hawkins 2009), © PNAS 2009.