Why are cell populations maintained via multiple compartments?

Flavia Feliciangeli[1,2], Hanan Dreiwi[1], Martín López-García[1], Mario Castro Ponce[3], Carmen Molina-París[1,4], and Grant Lythe[1]

[1] School of Mathematics, University of Leeds, Leeds LS2 9JT, UK.
[2] Systems Pharmacology and Medicine, Bayer AG, Leverkusen, 51368, Germany.
[3] Instituto de Investigación Tecnológica, Universidad Pontificia Comillas, Madrid, Spain.
[4] T-6, Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM 87545, USA.

10th October 2022

Abstract

We consider the maintenance of "product" cell populations from "progenitor" cells via a sequence of one or more cell types, or compartments, where each cell's fate is chosen stochastically. If there is only one compartment then large amplification, that is, a large ratio of product cells to progenitors comes with disadvantages. The product cell population is dominated by large families (cells descended from the same progenitor) and many generations separate, on average, product cells from progenitors. These disadvantages are avoided using suitably-constructed sequences of compartments: the amplification factor of a sequence is the product of the amplification factors of each compartment, while the average number of generations is a sum over contributions from each compartment. Passing through multiple compartments is, in fact, an efficient way to maintain a product cell population from a small flux of progenitors, avoiding excessive clonality and minimising the number of rounds of division *en route*. We use division, exit and death rates, estimated from measurements of single-positive thymocytes, to choose illustrative parameter values in the single-compartment case. We also consider a five-compartment model of thymocyte differentiation, from double negative precursors to single-positive product cells.

1 Introduction

Cell populations in organs and tissues are continuously replenished. There are many biological systems in which a small flux of progenitor cells continuously replenishes large populations of "product" cells via a structured developmental journey through a sequence of intermediate cell types [1–3]. Each cell type is referred to as a "compartment", whether or not it corresponds to a physical location. In different contexts, product cells may be termed "mature", "exhausted", "fully differentiated" or "effector" cells [4–6]. We model such systems, assuming that cells in each compartment may die, divide or "transit" to the next compartment, according to probabilistic rules. Only cells that reach the end of the sequence are called product cells. The set of product cells descended from a single progenitor is called a family. Theoretical and experimental arguments suggest that variability of family sizes is unavoidable if the fates of individual cells are subject to chance [7–10].

The dynamics of cellular developmental pathways is studied using recently-developed heritable labels, where individual progenitor haematopoietic and immune cells are tagged and their progeny counted [9–12]. Different experimental definitions of what constitutes a compartment are adopted: most often, human or mouse cells are classified by the abundance of one or more types of molecules on their surface, measured using flow cytometry. For example, in a study of the specific CD8$^+$ T-cell response to persistent *Toxoplasma gondii* infection, the surface markers CXCR3 and KLRG1 were used to identify an intermediate T-cell subset between memory and effector cells [13].

Maturation and selection of T cells in the thymus takes place via a sequence of cellular phenotypes, from bone-marrow progenitors to single-positive (SP4 or SP8) thymocytes [14–17], leading, in the case of an adult mouse, to about one million T cells per day exiting the thymus [18, 19]. In an adaptive immune response, naive antigen-specific T-cell populations expand dramatically. The numbers and phenotypes of descendants of individual naive T cells are highly variable, but the magnitude of the total response is reproducible when the output of many families is combined [9,10,20]. Variability of family sizes is confirmed by direct time-lapse observations *in vitro* [8].

Hundreds of billions of blood cells are replaced every day in a typical adult, all descended from small numbers of haematopoietic stem cells (HSCs) [21–23]. HSCs produce multipotent progenitor cells (MPPs) [2, 10,24] through a hierarchy of cellular states [25]: more primitive HSC1s and more mature HSC2s, followed by MPP1, MPP2 and MPP3 cells. Low rates of division of cells in early compartments of a lineage is conjectured to reduce the risk that potentially cancerous mutations accumulate [26–28]. Increased risk of T-cell acute lymphoblastic leukaemia [29, 30] is indeed found if the early compartments of the usual thymic sequence are absent [31, 32].

Here, we examine the amplification of a small flux of progenitor cells to continuously replenish a product cell population from a theoretical perspective, based on stochastic rules governing the fates of individual cells. We calculate the probability distributions of the number of product cells per progenitor cell, and of the number of rounds of division that separates them. Our particular focus is on how these distributions depend on the number of compartments. Every cell in each compartment undergoes one of three fates: the cell may divide, die or make a transition to the next compartment [33–36]. The "transition" event, corresponding to cell differentiation in many biological contexts, is called "exit" for short. The balance of probabilities between fates depends on the compartment but each cell in a compartment chooses its fate independently. In this sense our scheme is simpler than models that include interaction and competition between cells [37, 38]. A consequence of our assumption of independence is that a cell's division probability must be less than one half (otherwise the mean number of cells that descend from it would be infinite).

We analyse the possible descendants of one progenitor cell, families of cells that journey through the sequence of compartments. The number of cells from one family that become product cells is the random variable $\mathbf{R}$. To model the case where a small input flux of progenitors replenishes a larger product population, the mean of $\mathbf{R}$ will be large. In Section 2 we find the probability distribution of $\mathbf{R}$ as the ultimate state of a multitype branching process [39]. The mean number of product cells per progenitor, $\mathbb{E}(\mathbf{R})$, is denoted N. If there is a constant mean influx, ϕ, of progenitor cells, then there is a constant mean outflux, $N\phi$, of product cells. The single-compartment case is illustrated in Figure 1. It may be termed "direct differentiation" because only one such event is needed to convert a progenitor cell to a product cell. We note that the product cell population (red circles) consists of cells that become product cells at different times. Similarly, the solid blue circles in Figure 1 represent cells that are born, and may die, at different times. In this single-compartment scheme, large values of N are always associated with a high degree of clonality. Excessive "clonality", where the variation in family size, from one progenitor to another, causes the population of product cells to be dominated by a few large families, may increase the risk of cancerous mutations becoming established in the population [40, 41]. For example, the mean of $\mathbf{R}$ is equal to 10 if ten percent of progenitors yield 100 product cells, and the remainder yield none. One of our main results is that large values of N are possible without excessive clonality when the number of compartments, C, is greater than one, as illustrated in Figure 2.

The ability of product cells to perform their function may be negatively affected by the number of rounds of cell division that separates them from their progenitor, because every round of division brings with it a risk of mutation [42, 43]. For this reason, as well as identifying an individual cell by the compartment it belongs to, $c = 1, \ldots, C$, we label it by generation, $n = 0, 1, \ldots$. The progenitor cell is said to be in generation 0. Whenever a cell in generation n divides, the result is two cells in generation $n+1$ [44, 45]. From this point of view, the population of product cells is heterogeneous because it is made up of cells of different generations (Figure 3), cells with different "replicative histories" [23] or "replicative ages" [46]. Our analysis centres on the random variable $\mathbf{G}$, defined to be the generation number of a randomly-selected product cell.

The paper is organised as follows. Sections 2, 3 and 4 consist of the main theoretical results and a set of remarks. In Section 2, we analyse the case $C = 1$. Explicit expressions for the distribution of family sizes are obtained via the probability generating function. In Section 3, we consider sequences of compartments: cells may make a transition from compartment c to compartment $c+1$, for $c = 1, \ldots, C-1$. We treat the structured journey of development from a single progenitor cell to a population of product cells as a realisation of a multitype branching process [47, 48]. By contrast, we note that discretised age-structured models [49] are different from sequences of compartments because birth events produce new individuals in the first compartment only. Since we are interested in the ultimate fate of the system, we proceed as in the theory of discrete-time branching processes, by defining relationships between random variables using probability generating functions. For instance, the probability generating function of the number of cells that exit the final compartment, descended from one progenitor cell, is given as a composition of probability generating functions. Note that, while mean quantities can also be obtained by solving linear systems of ordinary differential equations [50–54], the full distribution of $\mathbf{R}$ is encoded in its probability generating function. The product cell population, classified into generations, is examined in Section 4. In particular, we consider the random variable $\mathbf{G}$: its mean value, D, and its distribution (as encoded in its probability generating function). In Section 5, we generalise our considerations to include a fourth type of event: asymmetric division; that is, a division event that leaves one daughter cell in the same compartment that the mother cell divided and the other daughter cell exits the compartment. The appendices provide additional details, not included in the main body of the manuscript. In particular, the recursion relations that we use to generate the probability that k cells exit from one or two compartments are given in Appendix A; the variance of the random variable $\mathbf{R}$, which is proportional to $N^{2+1/C}$ when N is large, is calculated in Appendix B; and the generalisation of our methods to include asymmetric division is presented in Appendix C.

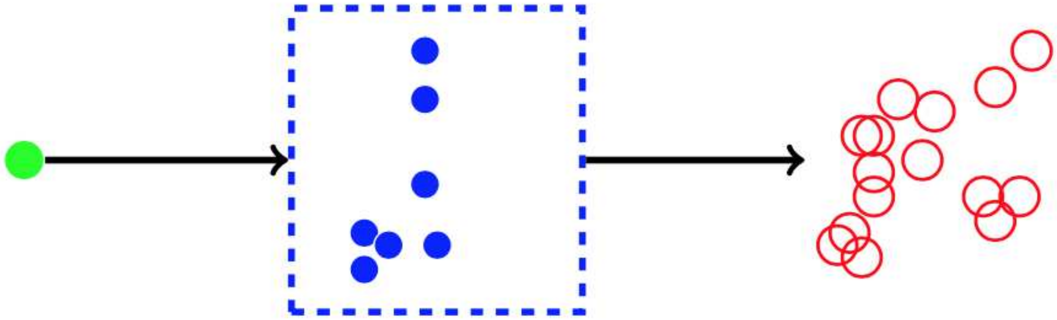

Figure 1: The one-compartment system. A single progenitor cell (shown on the left, green) is the founder of the population. In the compartment (represented by the dashed box), each cell (shown as a blue filled circle), independently, may die, divide, or "exit". An exit event is the differentiation of a cell to product cell type (shown as a red empty circle). The random variable $\mathbf{R}$ is the number of product cells when no cells remain in the compartment. We count the product cells as a cumulative total and do not consider any death or division events of product cells. The quantity $N = \mathbb{E}(\mathbf{R})$ is the "amplification factor": the mean number of product cells per progenitor.

2 How many cells exit a compartment?

The case of one compartment is illustrated in Figure 1. Three types of single-cell events contribute to the creation of a family of product cells from a single progenitor: individual cells may divide, die or transit (or differentiate) to a different cell type, or compartment. Our assumption is that every cell in a given compartment follows the same rules, independently, which is a fundamental assumption in branching processes [55,56]. Here, we restrict ourselves to counting cells, ignoring both inter-event times and the total time taken for progeny to disappear from all intermediate compartments and exit from the last one.

Analyses based on ordinary differential equations can calculate mean quantities, such as the mean number of product cells per progenitor. We, instead, calculate full distributions using first-step arguments and the probability generating function. The full distribution is of particular relevance in experiments where only

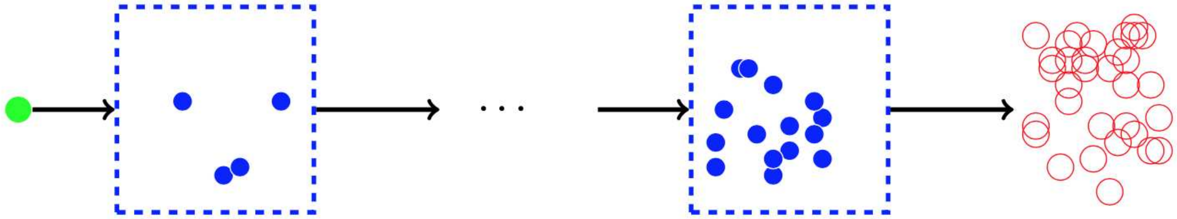

Figure 2: The multiple-compartment system. A single progenitor cell (shown on the left, green) is the founder of the population. Each cell in compartment c, independently, may die, divide or transit from compartment c to compartment $c+1$, where $c = 1, \dots, C-1$. Cells that exit compartment C are product cells (shown in red). The overall amplification factor N is the mean number of product cells per progenitor, which is the product of the amplification factors in each compartment.

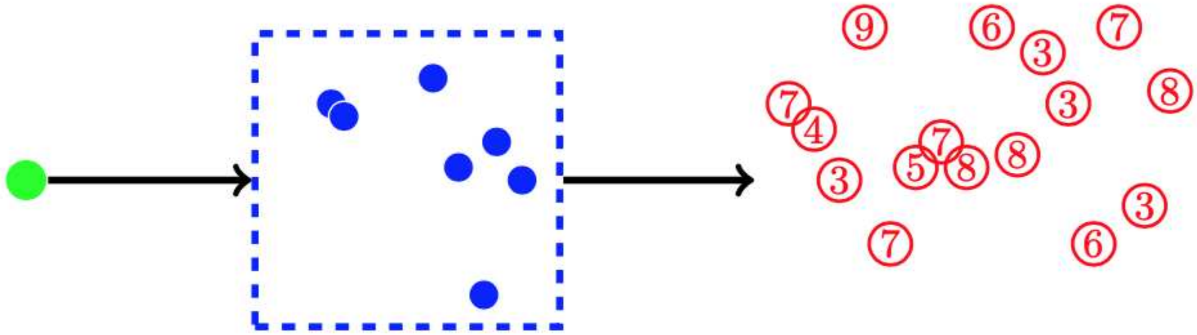

Figure 3: We classify the set of product (red) cells according to generation (number of divisions from the progenitor cell). The progenitor cell is said to be in generation 0. Whenever a cell in generation n divides, the result is two daughter cells in generation $n+1$. The final state of the process is a population of red cells, each having made the transition at a different time and each with its own generation number. The case $C = 1$ is illustrated here. If $C > 1$ then the mean number of divisions in the product population is the sum of the mean numbers of divisions in each compartment.

a finite number of families can be tracked. When the rules at the level of a single cell are stochastic, some progenitors do not yield any product cells, while some found large families.

In this Section we analyse the case of one compartment, $C = 1$. Each cell in the compartment, independently, may die, divide, or exit the compartment, with respective probabilities $p_{\rm d}$, $p_{\rm b}$ and $p_{\rm e}$, where $p_{\rm d} + p_{\rm b} + p_{\rm e} = 1$. We assume that

$$p_{\rm d} + p_{\rm e} > p_{\rm b}, \tag{H1}$$

so that extinction is the ultimate fate of the population of (blue) cells in the compartment. Exit has the same effect as death on the population in the compartment because exited cells play no further part in the dynamics of that compartment. Although the ultimate fate of the system is not affected by the inter-event time distributions, it is useful to keep in mind some examples that satisfy the assumptions that every cell, independently, dies, divides, or exits with probabilities $p_{\rm d}$, $p_{\rm b}$ and $p_{\rm e}$, repectively.

- A continuous-time birth-death-migration Markov process with exponential waiting times, where the probabilities $p_{\rm b}$, $p_{\rm d}$, and $p_{\rm e}$ are related to the rates of death, division and exit (*i.e.,* migration) , μ, λ

and ν, respectively, by

$$p_{\rm d} = \frac{\mu}{\mu+\nu+\lambda}, \quad p_{\rm b} = \frac{\lambda}{\mu+\nu+\lambda}, \quad p_{\rm e} = \frac{\nu}{\mu+\nu+\lambda}. \tag{1}$$

Sawicka *et al.* [14] estimated μ, λ and ν for SP4 and SP8 thymocytes based on experimental data [57]. The estimated division rates were $\lambda_4 = 0.181$ day^{-1} and $\lambda_8 = 0.085$ day^{-1}; death rates $\mu_4 = 0.040$ day^{-1} and $\mu_8 = 0.110$ day^{-1}; and exit rates $\nu_4 = 0.231$ day^{-1} and $\nu_8 = 0.152$ day^{-1}, respectively for SP4 and SP8 (see Section 3.3, Table 2 of Ref. [14]).

- A population in which each cell is assigned three independent random variables: a death time $\tau_{\rm d}$, a division time $\tau_{\rm b}$, and a differentiation time $\tau_{\rm e}$. The fate of the cell is whichever is the minimum of the three times [8, 58]. Then, probabilities can be defined as follows

$$p_{\rm d} = \mathbb{P}\left(\tau_{\rm d} < \tau_{\rm b} \text{ and } \tau_{\rm d} < \tau_{\rm e}\right), \quad p_{\rm b} = \mathbb{P}\left(\tau_{\rm b} < \tau_{\rm d} \text{ and } \tau_{\rm b} < \tau_{\rm e}\right), \quad \text{and} \quad p_{\rm e} = \mathbb{P}\left(\tau_{\rm e} < \tau_{\rm b} \text{ and } \tau_{\rm e} < \tau_{\rm d}\right).$$

We note that (1) holds in the case where the probability densities of $\tau_{\rm d}, \tau_{\rm b}$, and $\tau_{\rm e}$ are exponential.

The random variable $\mathbf{R}$ is the total number of product cells, starting from a single progenitor cell. Let us define q_k as follows:

$$q_k = \mathbb{P}\left(\mathbf{R} = k\right), \quad k = 0, 1, 2, \ldots. \tag{2}$$

We make use of the following argument based on the first event that occurs in the compartment. If the first event is cell division, then the two daughter cells, independently, follow the same rules as their mother cell. Therefore, q_0 satisfies the quadratic equation

$$q_0 = p_{\rm d} + p_{\rm b}\, q_0^2. \tag{3}$$

We can read (3) as a sum over the three possible first events, making use of the law of total probability:

$$\sum_{s\in\{\rm d,e,b\}} p_s\, \mathbb{P}\left(\mathbf{R} = 0 \,|\, \text{first event is } s\right) = p_{\rm d}1 + p_{\rm e}0 + p_{\rm b}q_0^2.$$

Because q_0 is a probability, we take the solution of (3) in the interval $[0, 1]$, given by

$$q_0 = \frac{1-\Delta}{2p_{\rm b}} = \frac{2p_{\rm d}}{1+\Delta}, \quad \text{where} \quad \Delta^2 = 1 - 4p_{\rm d}p_{\rm b}. \tag{4}$$

Similarly, the mean of $\mathbf{R}$ can be written as

$$N = \mathbb{E}(\mathbf{R}) = \sum_{s\in\{\rm d,e,b\}} p_s\, \mathbb{E}(\mathbf{R} \,|\, \text{first event is } s) = p_{\rm d}0 + p_{\rm e}1 + p_{\rm b}2N, \tag{5}$$

so

$$N = \frac{p_{\rm e}}{1-2p_{\rm b}}. \tag{6}$$

The condition (H1), which is equivalent to $2p_{\rm b} < 1$, assures that N is finite. We also observe that $p_{\rm b}$ must be close to $\frac{1}{2}$ for N to be large.

The probability q_1 satisfies an equation similar to (3):

$$q_1 = p_{\rm e} + p_{\rm b}\, 2q_0q_1. \tag{7}$$

Thus, we have $q_1 = \frac{p_{\rm e}}{\Delta}$. We may find further q_k (for $k \geq 2$) making use of the relationship

$$q_k = p_{\rm b}\left(q_kq_0 + q_{k-1}q_1 + \cdots + q_1q_{k-1} + q_0q_k\right), \quad \text{so} \quad q_k = \frac{p_{\rm b}}{\Delta}\sum_{j=1}^{k-1} q_j\, q_{k-j}, \quad k \geq 2. \tag{8}$$

However, it is more convenient to consider the probability generating function of the random variable $\mathbf{R}$, defined as

$$\phi(z) = \mathbb{E}(z^{\mathbf{R}}) = q_0 + q_1 z + q_2 z^2 + \dots . \tag{9}$$

The probability generating function, like q_0, satisfies a quadratic equation [59, 60]:

$$\phi(z) = \sum_{s\in\{\mathrm{d,e,b}\}} p_s \, \mathbb{E}(z^{\mathbf{R}} \mid \text{first event is } s) = p_{\mathrm{d}} z^0 + p_{\mathrm{e}} z^1 + p_{\mathrm{b}} \phi^2(z).$$

Thus, taking the sign of the square root that yields $\phi(1) = 1$, we obtain

$$\phi(z) = \frac{1 - (1 - 4p_{\mathrm{b}}p_{\mathrm{d}} - 4p_{\mathrm{b}}p_{\mathrm{e}}z)^{1/2}}{2p_{\mathrm{b}}}. \tag{10}$$

Using either (10) or (8), we find

$$q_k = \left(\frac{p_{\mathrm{b}}}{\Delta}\right)^{k-1} \left(\frac{p_{\mathrm{e}}}{\Delta}\right)^{k} c_{k-1}, \quad k \geq 1, \tag{11}$$

where $c_0 = 1$ and for $k \geq 1$, we have

$$c_k = \frac{(2k)!}{k!(k+1)!}.$$

The c_k are known as the Catalan numbers [61]. Examples of q_k are shown in Figure 4 for two different choices of p_{b} and p_{e}. With the estimates of Sawicka *et al.* [14], $N \simeq 2.57$ (for SP4 thymocytes) and $N \simeq 0.86$ (for SP8 thymocytes).

The distribution (11) of the random variable $\mathbf{R}$ is not one of the well-known distributions, such as Poisson or geometric. We therefore provide some remarks on its properties.

Remark 2.1 Given any two of p_{d}, p_{b}, and p_{e}, we can recover the third using $p_{\mathrm{d}} + p_{\mathrm{b}} + p_{\mathrm{e}} = 1$. In fact, we may parametrise the compartment in terms of any two, linearly independent, combinations of p_{d}, p_{b} and p_{e}. We will, on occasions, use N itself along with p_{d}. That is, using $N = \frac{p_{\mathrm{e}}}{1-2p_{\mathrm{b}}}$, we can write

$$p_{\mathrm{b}} = \frac{N - 1 + p_{\mathrm{d}}}{2N - 1}, \quad \text{and} \quad p_{\mathrm{e}} = \frac{N(1 - 2p_{\mathrm{d}})}{2N - 1}. \tag{12}$$

Remark 2.2 The variance, V, of $\mathbf{R}$ is given by

$$V = \phi''(1) + N - N^2 = \frac{2p_{\mathrm{b}}}{p_{\mathrm{e}}} N^3 + N - N^2, \tag{13}$$

which can be rewritten as

$$V = \frac{2}{1 - 2p_{\mathrm{d}}}(N - 1 + p_{\mathrm{d}})N^2 + N - N^2. \tag{14}$$

Thus, the standard deviation of $\mathbf{R}$ is proportional to $N^{3/2}$ as $N \to +\infty$.

Remark 2.3 It is convenient to generate values of q_k, ($k \geq 1$), via the recursion relation

$$q_{k+1} = \frac{2k-1}{k+1} \frac{2p_{\mathrm{b}}p_{\mathrm{e}}}{1 - 4p_{\mathrm{b}}p_{\mathrm{d}}} q_k. \tag{15}$$

Remark 2.4 We note that [62]

$$q_k < \frac{p_{\mathrm{e}}}{\sqrt{\pi}\Delta} \, \gamma_1^{k-1} \, k^{-3/2}, \quad k \geq 1, \tag{16}$$

where we have introduced

$$\gamma_1 = \frac{4p_{\mathrm{b}}p_{\mathrm{e}}}{1 - 4p_{\mathrm{b}}p_{\mathrm{d}}}. \tag{17}$$

If $N \gg 1$, then we have $\gamma_1 \simeq 1 - \frac{1-2p_{\mathrm{d}}}{4N^2}$.

Remark 2.5 The factor $k^{-3/2}$ in (16) can be understood [63–65] as resulting from the square-root singularity in the probability generating function (10) rearranged as follows:

$$2p_{\mathrm{b}}\phi(z) = 1 - \Delta(1-\gamma_1 z)^{1/2}. \tag{18}$$

Remark 2.6 The right-hand side of (16) is the asymptotic form of q_k as $k \to +\infty$ [62]. That is, we have

$$\log\left(\frac{q_{k+1}}{q_k}\right) \simeq \log\gamma_1 - \frac{3}{2}\log\left(1+\frac{1}{k}\right),$$

when $k \gg 1$. If, in addition, $N \gg 1$ then we can write

$$\log\left(\frac{q_{k+1}}{q_k}\right) \simeq -\frac{1-2p_{\mathrm{d}}}{4N^2} - \frac{3}{2}\frac{1}{k}. \tag{19}$$

The decrease in q_k as a function of k is primarily due to the factor $k^{-3/2}$, when $(1-2p_{\mathrm{d}})k < 6N^2$; thereafter, it is due to the factor γ_1^k (see Figure 5). We may summarise the behaviour of q_k as having two régimes: it is first governed by the power law when k is small enough that $\gamma_1^k \simeq 1$, then by the geometric term at values of k greater than $6N^2/(1-2p_{\mathrm{d}})$.

Remark 2.7 In a population of cells made up of multiple realisations of $\mathbf{R}$, we can also understand the dominance of large families of cells by evaluating k_{50}, the lowest value of k such that half of the cells are part of a family of fewer than k cells. That is,

$$\frac{N}{2} < \sum_{k=1}^{k_{50}} kq_k.$$

Using (16), $kq_k < \frac{p_{\mathrm{e}}}{\sqrt{\pi}\Delta}\frac{1}{\sqrt{k}}$, so we can write

$$\begin{aligned} \frac{\sqrt{\pi}\Delta}{2p_{\mathrm{e}}}N &< \sum_{k=1}^{k_{50}} \frac{1}{\sqrt{k}}, \\ \frac{\sqrt{\pi}\Delta}{2p_{\mathrm{e}}}N &< 2\sqrt{k_{50}}, \\ k_{50} &> \frac{\pi\Delta^2}{16p_{\mathrm{e}}^2}N^2. \end{aligned} \tag{20}$$

Assuming $N > 1$ and using (12), we conclude

$$k_{50} > \frac{\pi}{16}\frac{\Delta^2}{(1-2p_{\mathrm{d}})^2}(2N-1)^2. \tag{21}$$

The factor $\frac{\Delta^2}{(1-2p_{\mathrm{d}})^2}$ is an increasing function of p_{d}. In summary, for a given value of N, k_{50} is minimised by setting $p_{\mathrm{d}} = 0$. An analytical bound on this minimum is $k_{50} > \frac{\pi}{16}(2N-1)^2$. Some numerical examples are: when $N = 10$ and $p_{\mathrm{d}} = 0$, $k_{50} = 83$ and the analytical bound (21) is $k_{50} > 71$; when $N = 10^2$ and $p_{\mathrm{d}} = 0$, $k_{50} = 9,009$ and the bound is $k_{50} > 7,775$.

3 How many cells exit a sequence of compartments?

We now consider the case where there are C compartments before the final population of product cells. The random variable $\mathbf{R}$ is the number of product cells, descended from one cell in the first compartment. That is, there are C "transition" or "differentiation" events between the progenitor and the product phenotype. The case $C = 1$ was analysed in Section 2. The case $C \geq 2$ is illustrated in Figure 2.

Each cell, independently, may die, divide, or make a transition from its current compartment to the next, with probabilities

$$p_{\mathrm{d}}(c), \quad p_{\mathrm{b}}(c), \quad \text{and} \quad p_{\mathrm{e}}(c),$$

where $p_{\mathrm{d}}(c) + p_{\mathrm{b}}(c) + p_{\mathrm{e}}(c) = 1$ for each c, with $c = 1, \ldots, C$. The condition (H1), that guarantees a finite number of product cells, is imposed in each compartment:

$$p_{\mathrm{d}}(c) + p_{\mathrm{e}}(c) > p_{\mathrm{b}}(c), \quad \text{for each } c, \quad \text{with} \quad c = 1, \ldots, C.$$

The quantity $N_c = \dfrac{p_{\mathrm{e}}(c)}{1 - 2p_{\mathrm{b}}(c)}$ is the mean number of cells exiting compartment c for each cell that makes a transition to that compartment (from compartment $c-1$). If $\mathbf{R}_c$ is the number of cells exiting compartment c, descended from one cell in compartment c, then the probability generating function of $\mathbf{R}_c$ is

$$\phi_c(z) = \frac{1 - \left[\Delta_c^2 - 4p_{\mathrm{b}}(c)p_{\mathrm{e}}(c)z\right]^{1/2}}{2p_{\mathrm{b}}(c)}, \quad c = 1, \ldots, C, \tag{22}$$

with $\Delta_c^2 = 1 - 4p_{\mathrm{d}}(c)p_{\mathrm{b}}(c)$. We can write $N_c = \mathbb{E}(\mathbf{R}_c) = \phi_c'(1)$.

We seek $Q_k(C)$, the probability that the number of product cells, descended from a single progenitor via C intermediate compartments, is equal to k. We can write

$$Q_k(C) = \mathbb{P}(\mathbf{R} = k), \quad k = 0, 1, 2, \ldots. \tag{23}$$

The probability generating function of $\mathbf{R}$ is given by

$$\Phi_C(z) = \mathbb{E}(z^{\mathbf{R}}) = Q_0(C) + zQ_1(C) + z^2Q_2(C) + \cdots. \tag{24}$$

If $C = 1$ (there is only one compartment) then we recover the results of Section 1. That is, $Q_k(1) = q_k$ and $\Phi_1(z) = \phi_1(z)$. If $C = 2$, we may write

$$\mathbf{R} = \sum_{i=1}^{\mathbf{R}_1} \mathbf{R}_{2,i}, \tag{25}$$

where the $\mathbf{R}_{2,i}$ are identical and independent random variables with the same distribution as $\mathbf{R}_2$. Using (25), we find [55, 56, 60]:

$$\Phi_2(z) = \phi_1(\phi_2(z)). \tag{26}$$

In general, we have

$$\Phi_C(z) = \phi_1(\phi_2(\cdots\phi_C(z))). \tag{27}$$

We maintain the notation that $\mathbf{R}$ is the number of product cells, N the mean and V the variance of $\mathbf{R}$. The overall amplification factor is then given by

$$N = \prod_{c=1}^{C} N_c. \tag{28}$$

Remark 3.1 The definition (24) relates the probability generating function to a set of probabilities. Different algorithms exist for extracting numerical values of the probabilities in situations where the probability generating function is known [66]. Because we have found it convenient to generate values of $Q_k(C)$ using a recursion relation similar to (15), we show how to obtain such relations in Appendix A.

Remark 3.2 An interesting feature of the distribution of $\mathbf{R}$ is the universality of its large-k behaviour:

$$Q_k(C) \propto \gamma_C^k \, k^{-3/2}, \quad \text{as} \quad k \to +\infty. \tag{29}$$

We may determine γ_C by locating the square-root singularity of $\Phi_C(z)$ [63–65]. We find that $\gamma_1 = 4p_{\mathrm{b}}(1)p_{\mathrm{e}}(1)/\Delta_1^2$ and γ_2 satisfies $4p_{\mathrm{b}}(1)p_{\mathrm{e}}(1)\phi_2(\gamma_2^{-1}) = \Delta_1^2$.

We define

$$\chi_C(z) = \phi_2(\phi_3(\cdots \phi_C(z))), \tag{30}$$

so that (18) is generalised to

$$[1 - 2p_{\mathrm{b}}(1)\Phi_C(z)]^2 = \Delta_1^2[1 - \gamma_1 \chi_C(z)]. \tag{31}$$

We expand around $z = 1$, making use of the fact that $\chi(1) = 1$ and $\chi'(1) = N/N_1$, to obtain

$$1 - \gamma_1 \chi_C(z) \simeq 1 - \gamma_1[1 - (1-z)N/N_1] = [\gamma_1(N/N_1 - 1) + 1] \left(1 - \frac{\gamma_1 N/N_1}{\gamma_1(N/N_1 - 1) + 1} z\right).$$

We are then able to identify

$$\gamma_C = \left(1 + \frac{1 - \gamma_1}{\gamma_1 N/N_1}\right)^{-1}. \tag{32}$$

If $N_1, N \gg 1$ then $1 - \gamma_1 \simeq \frac{1}{4N_1^2}$ and we can approximate γ_C by the following expression

$$\gamma_C \simeq 1 - \frac{1 - 2p_{\mathrm{d}}(1)}{4N_1 N}. \tag{33}$$

Remark 3.3 If $C > 2$, we may make further progress with some assumptions to reduce the number of parameters. For example, consider the case where N_c is independent of c and $p_{\mathrm{d}}(c) = 0$ in each compartment. Then

- the variance of $\mathbf{R}$ is proportional to $N^{2+\frac{1}{C}}$ as $N \to +\infty$ (for details, see Appendix B), and
- the constant γ_C can be written as follows

$$\gamma_C = 1 - \frac{1}{4}\frac{1}{N^{1+1/C}} + \frac{1}{16}\frac{1}{N^{2(1+1/C)}} + \cdots . \tag{34}$$

Figure 6 shows $k^{3/2}Q_k(C)$ as a function of k, with parameters chosen as just described above. In all three cases shown, the mean number of product cells, N, is equal to 25 and N_c is independent of c. We shall see, below, that this choice of parameters is optimal from the perspective of minimising the mean number of divisions per cell. The fact that the most efficient arrangement of compartments is found when each has the same amplification factor does not rule out different dynamics in different compartments. Indeed, a common scenario in cell biology is each compartment has faster rates than its predecessor [67–69].

Remark 3.4 One effect of the presence of multiple compartments can be understood by comparison with the k_{50} values in Remark 2.7 (for a single compartment). If $C = 2$, $N = 10$ and $p_{\mathrm{d}} = 0$, then the k_{50} value is 33; if $C = 2$, $N = 100$ and $p_{\mathrm{d}} = 0$, it is 1010. The corresponding k_{50} values when $C = 3$ are 25 and 528, for $N = 10$ and $N = 100$, respectively.

4 The population of exiting cells: how many divisions?

The progenitor cell is in generation 0. Daughter cells of the progenitor cell are said to be in generation 1. Daughter cells of a cell in generation n are in generation $n + 1$. In this way, the product cell population is classified by generation number, which is the number of divisions that separates a cell from the progenitor, or the depth of the cell in the tree that begins with the progenitor [70]. In Sections 2 and 3, we calculated the distribution of $\mathbf{R}$, the number of product cells per progenitor, its mean and variance. In this Section, we derive the probability generating function of the random variable $\mathbf{G}$, the generation number of a randomly-selected product cell.

4.1 Classifying cells by generation: a single compartment

To define the random variable $\mathbf{G}$, we begin with two simple random variables, $\mathbf{U}$ and $\mathbf{V}$, with state space $\{0,2\}$ and $\{0,1\}$, respectively, and such that

$$\mathbb{P}(\mathbf{U}=0)=1-p_{\mathrm{b}}, \quad \mathbb{P}(\mathbf{U}=2)=p_{\mathrm{b}}, \quad \text{and} \quad \mathbb{P}(\mathbf{V}=0)=1-p_{\mathrm{e}}, \quad \mathbb{P}(\mathbf{V}=1)=p_{\mathrm{e}}.$$

We recall the random variables of a discrete-time branching process [55, 56, 71]. Let us introduce $\mathbf{Z}_0=1$ and

$$\mathbf{Z}_{n+1}=\sum_{i=1}^{\mathbf{Z}_n}\mathbf{U}_i, \quad n=0,1,2,\ldots, \tag{35}$$

where, for each i, $\mathbf{U}_i$ is an independent copy of $\mathbf{U}$. $\mathbf{Z}_n$ is the number of cells in generation n, whatever their fate, and each $\mathbf{U}_i$ is the number of daughter cells from one cell. Here, we also need to define

$$\mathbf{Y}_n=\sum_{i=1}^{\mathbf{Z}_n}\mathbf{V}_i, \quad n=0,1,2,\ldots, \tag{36}$$

where each $\mathbf{V}_i$ is an independent copy of $\mathbf{V}$. $\mathbf{Y}_n$ is the number of product cells in generation n. The random variables $\mathbf{R}$ and $\mathbf{G}$ are defined via

$$\mathbf{R}=\sum_{n=0}^{+\infty}\mathbf{Y}_n, \quad \text{and} \quad \mathbb{P}(\mathbf{G}=n)=\frac{1}{N}\mathbb{E}(\mathbf{Y}_n). \tag{37}$$

One realisation of the process is shown in Figure 7.

The mean values of $\mathbf{Y}_n$ are given by

$$\mathbb{E}(\mathbf{Y}_n)=p_{\mathrm{e}}\mathbb{E}(\mathbf{Z}_n)=p_{\mathrm{e}}(2p_{\mathrm{b}})^n. \tag{38}$$

The condition (H1) is equivalent to $2p_{\mathrm{b}}<1$. Hence, as $n\to+\infty$, $\mathbb{E}(\mathbf{Z}_n)\to 0$ and $\mathbb{E}(\mathbf{Y}_n)\to 0$.

Recall that the average number of product cells is $N=\dfrac{p_{\mathrm{e}}}{1-2p_{\mathrm{b}}}$. The average generation number in the product cell population is given by

$$D=\mathbb{E}(\mathbf{G})=\frac{p_{\mathrm{e}}}{N}\sum_{n=1}^{+\infty}n(2p_{\mathrm{b}})^n=\frac{2p_{\mathrm{b}}}{1-2p_{\mathrm{b}}}. \tag{39}$$

Using (37), we find that the variance of $\mathbf{G}$ is given by $\mathrm{var}(\mathbf{G})=D(D+1)$.

In Figure 8, N and D are displayed as functions of p_{b} and p_{d}: lines of constant N are blue and lines of constant D are red. Also shown (in green) are the estimates of Sawicka *et al.* [14]: $p_{\mathrm{b}}=0.4004$ and $p_{\mathrm{d}}=0.0885$ (SP4 thymocytes) and $p_{\mathrm{b}}=0.2449$ and $p_{\mathrm{d}}=0.3170$ (SP8 thymocytes). We note the following limits: (i) as $p_{\mathrm{b}}\to\frac{1}{2}$ with p_{d} fixed, $\dfrac{D}{N}\to\dfrac{2}{1-2p_{\mathrm{d}}}$; (ii) as $p_{\mathrm{b}}\to 0$ with p_{d} fixed, $N\to 1-p_{\mathrm{d}}$ and $D\to 0$.

Remark 4.1 As in Section 2, we make use of the freedom to express all single compartment quantities in terms of N and p_{d}. Combining (5) and (39) gives the following linear relationship between D and N:

$$D=\frac{2N-1}{1-2p_{\mathrm{d}}}-1. \tag{40}$$

Given $N>1$, the minimum possible value of D is found when $p_{\mathrm{d}}=0$:

$$D_{\min}=2(N-1). \tag{41}$$

Remark 4.2 We may express all single compartment quantities in terms of variables which can be experimentally measured, such as number of product cells and generations, N and D. In particular, we have

$$p_{\mathrm{b}}=\frac{1}{2}\frac{D}{D+1}, \quad \text{and} \quad p_{\mathrm{e}}=\frac{N}{D+1}.$$

These relationships could enable p_{b}, p_{d} and p_{e}, to be determined from experimentally-measurable quantities, $N=\mathbb{E}(\mathbf{R})$ and $D=\mathbb{E}(\mathbf{G})$ [9, 10, 20]. The corresponding variances have simple expressions: $V=\mathrm{var}(\mathbf{R})=N^2(D-1)+N$ and $\mathrm{var}(\mathbf{G})=D(D+1)$, respectively.

4.2 Classifying cells by generation: a sequence of C compartments

Cells that transit from compartment c to compartment $c+1$, with $c = 1, \dots C-1$, retain their generation number. Cells that exit compartment C are product cells. To analyse the multi-compartment system, we define the following sets of random variables, $\mathbf{Z}_n(c)$ and $\mathbf{Y}_n(c)$, as follows:

- For $n \geq 0$ and $1 \leq c \leq C$, $\mathbf{Z}_n(c)$ is the number of generation n cells in compartment c, whatever their fate. We assume that $\mathbf{Z}_0(1) = 1$.
- For $n \geq 0$ and $1 \leq c \leq C$, $\mathbf{Y}_n(c)$ is the number of generation n cells that exit compartment c. That is, $\mathbf{Y}_n(c) \leq \mathbf{Z}_n(c)$.

Then

$$\mathbf{Z}_0(c) = \mathbf{Y}_0(c-1), \quad c = 2, \dots, C.$$

To express the relationships between the random variables $\mathbf{Z}_n(c)$ and $\mathbf{Y}_n(c)$, we introduce for $1 \leq c \leq C$, the random variables $\mathbf{U}(c)$ and $\mathbf{V}(c)$, with state space $\{0, 2\}$ and $\{0, 1\}$, respectively, such that

$$\mathbb{P}(\mathbf{U}(c) = 0) = 1 - p_{\mathrm{b}}(c), \quad \mathbb{P}(\mathbf{U}(c) = 2) = p_{\mathrm{b}}(c), \quad \text{and} \quad \mathbb{P}(\mathbf{V}(c) = 0) = 1 - p_{\mathrm{e}}(c), \quad \mathbb{P}(\mathbf{V}(c) = 1) = p_{\mathrm{e}}(c).$$

The relation (35), standard in branching processes, is generalised to one that may appear in a branching process with immigration. For $c \geq 2$, we have $\mathbf{Z}_{n+1}(1) = \sum_{i=1}^{\mathbf{Z}_n(1)} \mathbf{U}_i(1)$ and

$$\mathbf{Z}_{n+1}(c) = \mathbf{Y}_{n+1}(c-1) + \sum_{i=1}^{\mathbf{Z}_n(c)} \mathbf{U}_i(c), \quad c = 2, \dots, C, \quad n = 0, 1, \dots, \tag{42}$$

and

$$\mathbf{Y}_n(c) = \sum_{i=1}^{\mathbf{Z}_n(c)} \mathbf{V}_i(c), \quad c = 1, \dots, C, \quad n = 0, 1, \dots, \tag{43}$$

The number of product cells is the number of cells exiting the final compartment:

$$\mathbf{R} = \sum_{n=0}^{+\infty} \mathbf{Y}_n(C). \tag{44}$$

A realisation of the multi-compartment process is illustrated in Figure 9. The random variable $\mathbf{G}$ is the generation number of a randomly-selected product cell:

$$\mathbb{P}(\mathbf{G} = n) = \frac{1}{N} \mathbb{E}(\mathbf{Y}_n(C)). \tag{45}$$

We consider the two mean quantities that characterise each compartment:

$$N_c = \frac{p_{\mathrm{e}}(c)}{1 - 2p_{\mathrm{b}}(c)}, \quad \text{and} \quad D_c = \frac{2p_{\mathrm{b}}(c)}{1 - 2p_{\mathrm{b}}(c)}, \quad c = 1, \dots, C. \tag{46}$$

Thus, N_c is the mean number of cells exiting compartment c, descended from a single cell in compartment c, while D_c is the average increase in the generation number in the compartment (the average number of divisions undergone). We now introduce the following probability generating functions (for details, see Appendix C.2), to keep track of the increase in generation number in compartment c, for $c = 0, 1, \dots, C$:

$$\xi_c(z) = \frac{p_{\mathrm{e}}(c)}{N_c} \sum_{n=1}^{+\infty} (2zp_{\mathrm{b}}(c))^n = \frac{1 - 2p_{\mathrm{b}}(c)}{1 - 2p_{\mathrm{b}}(c)z}. \tag{47}$$

For the whole sequence of compartments, let N be the mean number of product cells for every progenitor cell, and D be the average generation number of a product cell. Then

$$N = \mathbb{E}(\mathbf{R}) = N_1 N_2 \cdots N_C, \quad \text{and} \quad D = \mathbb{E}(\mathbf{G}) = D_1 + D_2 + \cdots + D_C. \tag{48}$$

The difference between a single compartment and a sequence of multiple compartments is already apparent if we compare $C = 1$ to $C = 2$, given the same value of N. In Figure 10 we plot the average generation number, D, as a function of the mean number of exiting cells, N. In the examples with $C = 2$, shown on the right in Figure 10, $N_1 = N_2$. The green lines show cases where there is no cell death. Given a value of N, D is lower when $C = 2$ (proportional to $\sqrt{N}$ as $N \to +\infty$) than when $C = 1$ (proportional to N as $N \to +\infty$). Figure 11 illustrates the probability distribution of $\mathbf{G}$ for different values of C with N fixed. The distribution narrows as the number of intermediate compartments increases.

Finally, the probability generating function of $\mathbf{G}$, defined as $\Xi(z) = \sum_{n=0}^{+\infty} \mathbb{P}(\mathbf{G} = n) z^n$, is given by the product

$$\Xi(z) = \xi_1(z)\xi_2(z)\cdots\xi_C(z), \tag{49}$$

where, for each $c = 1, \dots, C$, $\xi_c(z)$ has been defined in (47).

4.3 Minimising the average generation number

Since excessive "clonality" may increase the risk of cancerous mutations becoming established [40, 41], and because every round of division brings with it a risk of mutation, senescence or exhaustion [72–75], we now ask ourselves, how should a sequence of C compartments be constructed in order to yield a given amplification of progenitor to product cells, while minimising the average number of divisions? Thus, given N, we seek to minimise D, given by (48). We write (46) as follows

$$D_c = \alpha_c N_c - \beta_c, \quad \text{where} \quad \alpha_c = \frac{2}{1 - 2p_\mathrm{d}(c)}, \quad \text{and} \quad \beta_c = \frac{2 - 2p_\mathrm{d}(c)}{1 - 2p_\mathrm{d}(c)}.$$

Let us imagine that the probabilities $p_\mathrm{d}(c)$ are fixed, but the probabilities $p_\mathrm{b}(c)$ are variable. Using the Lagrange multiplier method, we impose the constraint $N = N^*$ by defining

$$L(p_\mathrm{b}(1), \dots, p_\mathrm{b}(C), \Lambda) = D - \Lambda(N - N^*) = \sum_{c=1}^{C} \frac{2p_\mathrm{b}(c)}{1 - 2p_\mathrm{b}(c)} - \Lambda\left(\prod_{c=1}^{C} \frac{1 - p_\mathrm{b}(c) - p_\mathrm{d}(c)}{1 - 2p_\mathrm{b}(c)} - N^*\right). \tag{50}$$

We make use of the partial derivatives

$$\frac{\partial L}{\partial p_\mathrm{b}(c)} = \frac{2}{(1 - p_\mathrm{b}(c))^2}\left(1 - \Lambda\frac{N^*}{\alpha_c N_c}\right), \quad c = 1, \dots, C,$$

to find the following conditions

$$\alpha_1 N_1 = \alpha_2 N_2 = \cdots = \alpha_C N_C. \tag{51}$$

We continue the analysis by defining the arithmetic and geometric means of the α_c:

$$\bar{\alpha} = \frac{1}{C}\sum_{c=1}^{C} \alpha_c, \quad \text{and} \quad \tilde{\alpha} = \left(\prod_{c=1}^{C} \alpha_c\right)^{1/C}. \tag{52}$$

Then, the optimal values of N_c have the property that

$$\alpha_c N_c = N^{1/C}\tilde{\alpha}, \quad \text{for each} \quad 1 \le c \le C. \tag{53}$$

The corresponding minimum value of D is then given by

$$D_\mathrm{min} = \sum_{c=1}^{C} (\alpha_c N_c - \beta_c) = C\left(\tilde{\alpha} N^{1/C} - \frac{1}{2}\bar{\alpha} - 1\right), \tag{54}$$

which is an increasing function of each of the $p_\mathrm{d}(c)$ for $1 \le c \le C$.

An interesting observation that can be made from the conditions (51) is that, if $p_\mathrm{d}(c)$ does not depend on c, then N_c is also independent of c. That is, if the death probability does not vary from compartment to compartment, then the optimal arrangement of division rates is such that each compartment has the same amplification factor, $N_c = N^{1/C}$. Then, we have

$$D_{\min} = \frac{2C}{1-2p_{\rm d}}\left(N^{1/C} - 1 + p_{\rm d}\right). \tag{55}$$

Given N and C, $D_{\min}$ is an increasing function of $p_{\rm d}$. We observe that $D_{\min}$ is a decreasing function of C. As $C \to +\infty$, $D_{\min} \to 2\log N$, recovering the logarithmic behaviour characteristic of binary trees [42,76].

5 Asymmetric division

A subject of recent research is the possibility of asymmetric cell division, where one daughter cell remains in the mother's compartment while the other transitions to the next compartment [9,38,46,77–83]. From the point of view of Markov processes, an asymmetric division event is unusual, in that division and change of cell type are supposed to be simultaneous. From a biological point of view, on the other hand, defining such an event may be natural: the mother's intra-cellular and cell-surface proteins will not be exactly evenly partitioned between the two daughters, who may experience different conditions during the process of cell division [84,85]. From a modelling perspective, one could imagine the constant flux of progenitor cells in our scheme as being produced by a constant pool of stem cells undergoing asymmetric division.

The mathematics of asymmetric division is accommodated, as detailed in Appendix C, by introducing a fourth type of event, asymmetric division, and its corresponding probability, $p_{\rm a}$. It is also possible to consider a fifth, where both daughter cells exit their mother's compartment at birth [76], and to incorporate "de-differentiation": cells moving backward in the hierarchy [86]. Böttcher *et al.* [46] developed a mathematical model with three types of event that all involve division: both daughter cells may remain in a compartment, both may transition, or one may remain and one transition. In this Section, we explore and apply our methods to a biological system in which asymmetric cell division may play a role: T cell development [81].

The development of thymocytes involves waves of proliferation, intertwined with differentiation, apoptosis and self-renewal to produce mature T cells, each with a unique T cell receptor (TCR). T cell development takes place in the thymus and starts with lymphoid precursor cells, lacking expression of CD4 and CD8 co-receptors, known as double-negative (DN) thymocytes. The structured journey of development of these precursor cells involves the following stages, each of them defined by the cell-surface expression of developmentally regulated markers: DN1, DN2, DN3a, DN3b, DN4, and double-positive (DP) thymocytes [87,88]. Transition from the DN1 to DN2 stage marks the initiation of gene rearrangement at the TCRβ gene locus [87]. The DN3 stage is characterised by the expression of the pre-T cell receptor (pre-TCR). It is at this stage that β-selection takes place; a checkpoint which defines the transition from the pre-selection DN3a to the post-selection DN3b stage. The DN3b population gives rise to the DN4 subset, which in turn undergoes proliferation and differentiation [88]. Further development involves the up-regulation of both CD4 and CD8 co-receptors to generate DP cells. Finally, DP cells go through gene rearrangement at the TCRα gene locus and the resulting $\alpha\beta$ TCR heterodimer then undergoes MHC-mediated selection to yield SP4 or SP8 cells.

Mammalian T cell development suggests a possible role for asymmetric cell division [81] during the β-selection stage; subsequent divisions are predominantly symmetric. Pham *et al.* experimentally studied the DN3a to SP transition and defined a deterministic mathematical model of the process [81] (see Figure 12). Cells of the first compartment, DN3a-pre, can only die or undergo asymmetric cell division [81]. Thus, cells have already divided at least once when they arrive in the second compartment, as experimentally observed. The finding of Pham *et al.* that the death rate was larger than the rate of asymmetric division at the DN3a-pre stage implies, in the context of our model, that the probability of asymmetric cell division in the first compartment, $p_{\rm a}(1)$, is smaller than $\frac{1}{2}$, with $p_{\rm a}(1) + p_{\rm d}(1) = 1$. Cells in compartments two (DN3a-post), three (DN3b), four (DN4), and five (DP) can die, divide (symmetrically) or differentiate (transition to the next compartment). We then write $p_{\rm b}(c) + p_{\rm d}(c) + p_{\rm e}(c) = 1$ for $c = 2, 3, 4, 5$. DN3 thymocytes undergo β-selection, which raises their probability of death. Accordingly, we choose $p_{\rm b}(c) < p_{\rm d}(c)$ for DN3a-post and DN3b. By contrast, DN4 and DP thymocytes are more likely to divide than to die [87,88] (see Table 1).

The analysis of Pham *et al.* was purely deterministic and therefore only considered mean numbers of cells in each compartment. In Figure 12, we show the distributions of two biologically significant random variables in our stochastic model: the number of product cells in a family founded by one progenitor and the generation

	DN3a-pre	DN3a-post	DN3b	DN4	DP
$p_{\mathrm{b}}(c)$	0	0.25	0.25	0.45	0.45
$p_{\mathrm{e}}(c)$	0	0.3	0.3	0.3	0.3
$p_{\mathrm{d}}(c)$	0.55 / 0.9	0.45	0.45	0.25	0.25
$p_{\mathrm{a}}(c)$	0.45 / 0.1	0	0	0	0
N_c	$\frac{9}{11}$ / $\frac{1}{9}$	0.6	0.6	3	3
D_c	$\frac{20}{11}$ / $\frac{10}{9}$	1	1	9	9

Table 1: Parameter values for the five-compartment thymocyte development model. For any $1 \leq c \leq 5$, $p_{\mathrm{b}}(c)$ is the probability that a cell in compartment c divides, $p_{\mathrm{d}}(c)$ is the probability that a cell in compartment c dies, $p_{\mathrm{e}}(c)$ is the probability that a cell in compartment c transitions to compartment $c+1$, and $p_{\mathrm{a}}(c)$ is the probability that a cell in compartment c undergoes an asymmetric division event, where one daughter remains in compartment c and one transits to compartment $c+1$. The values of N_c and D_c are calculated using (6) and (39), (71) and (83).

number of a cell in the product cell (here, SP) population. Two cases are shown $p_{\mathrm{a}}(1) = 0.1$ and $p_{\mathrm{a}}(1) = 0.45$. In the first, 90% of DN3a-pre cells die, so the average family size in the product population, $N = 0.36$, is smaller, on average, than in the second case, when only 55% of DN3a-pre cells die and $N = 2.651$. (These values are the product of the N_c values in Table 1.) Nevertheless, in both cases families of over 10^2 cells are not uncommon. Single-positive thymocytes are released from the thymus to the periphery, where families of cells correspond to T cell receptor clonotypes [18, 19, 89–91]. In a mouse, where division of naive T cells in the periphery is rare, the diversity of the T cell repertoire (the number of different TCRs simultaneously present) and the distribution of family sizes are determined by the distribution of family sizes at the time of release from the thymus [90–94].

The distributions of generation number $\mathbf{G}$ are also shown in Figure 12. They are relatively narrow: product cells with $\mathbf{G} > 100$ are rare. The difference between the distributions with $p_{\mathrm{a}}(1) = 0.1$ and $p_{\mathrm{a}}(1) = 0.45$ is small because, in both cases, the majority of cells that make the transition DN3a-pre to DN3a-post do so in the first generation. The mean values, $D = 21.1$ and $D = 21.9$ respectively, may be obtained by summing the values of D_c, $c = 1, \ldots, 5$ given in Table 1.

In the example we have analysed in this section, the intermediate compartments have a rationale related to TCR selection that is independent of family sizes and the distribution of generation numbers: we may conclude nature has made a virtue of the necessity of passing through multiple stages. However, intermediate compartments are also found in other cellular replenishment systems without an obvious independent reason.

6 Conclusion

Cells of the same phenotype are often thought of as belonging to a compartment, which may correspond to a spatial location, a biological function, or simply a set of cell-surface attributes which can be measured with flow cytometry. In many circumstances, a population of "product" cells performing a specific role is maintained, via a sequence of compartments, from a much smaller progenitor population. Why are multiple such compartments so often observed rather than a simpler one-step differentiation from progenitor to product cell? Using theoretical arguments, we show why such schemes are advantageous. In our model, individual cells in a compartment may die or divide (in the compartment), or transition to the next compartment, meaning that they change phenotype or "differentiate". Our mathematical approach is based on two fundamental biological (or empirical) observations: amplification (from progenitor cell to product cell populations) and stochasticity (of the fate of individual cells). Thus, we assume that each cell in a given compartment, independently, chooses one of the available fates according to a shared set of probabilities: p_{b}, p_{e} and p_{d} are the probabilities of division, transition and death, respectively. When a cell divides, its daughter cells, independently, follow the same rules as their mother. Hence, all population properties are deduced from a complete understanding of the possible progeny of a single progenitor. Furthermore, the population of product cells is the sum of families, each founded by a single progenitor cell. We do not consider inter-event times. Rather, each realisation is a sequence of events that ultimately results in extinction of the progeny in the pre-product compartment or compartments, with only product cells surviving. We construct sequences of C compartments, where cells may transit from compartment c to compartment $c+1$, with $c = 1, \ldots, C-1$. Given an overall amplification factor, N, the dominance of large families of cells in the product cell population decreases as C increases. Using probability generating functions, we find $Q_k(C)$, the probability that the number of product cells, descended from a single progenitor via C intermediate compartments, is equal to k. When k is large, $Q_k(C) \propto \gamma_C^k \, k^{-3/2}$, with $\gamma_C < 1$.

Our model deals in probabilities, which we relate to two important quantities, N and D, that can be measured in some experiments. The first, N, is the average number of product cells descended from a single progenitor, which can be measured if the progenitor cell is given a heritable label. The second, D, is the mean generation number of the product cell population, which can be measured if progenitor cells are stained with a fluorescent dye that dilutes with division, such as cell trace CFSE or cell trace violet. A recently-developed genetic tracing technique called *DivisionRecorder* makes it possible to measure the mean number of divisions of immune cell populations up to dozens of rounds of division [20]. The analysis presented in this manuscript shows that both N and D have long-tailed distributions when there are no intermediate compartments, and it allows us to quantify the reduction of clonality and long-term division history in product cell populations as the number of compartments is increased [95].

When there is only a single compartment (that is, when progenitor cells differentiate directly into product cells) the mean number of product cells per progenitor is related to an individual cell's division and exit probabilities by $N = \frac{p_{\mathrm{e}}}{1-2p_{\mathrm{b}}}$ and the mean generation number in the product cell population is given by $D = \frac{2p_{\mathrm{b}}}{1-2p_{\mathrm{b}}}$. Thus, large values of N, found when the value of p_{b} is less than but close to $\frac{1}{2}$, lead to large values of D. The presence of intermediate compartments is advantageous from this point of view: the mean generation number, D, decreases as C increases. Given N, the minimum value of D, found when p_{d} is zero, is given by $D_{\min} = 2C(N^{1/C} - 1)$. Whatever the value of p_{d}, the most efficient arrangement of compartments is found when each has the same amplification factor.

Our theoretical analyses are found in Section 2 for a single compartment, Section 3 for a sequence of compartments, and Section 4 for the number of divisions in the compartmental system. We find that a sequence of compartments achieves the amplification of progenitor to product cells required in tissue organization and homeostasis while avoiding excessive clonality and minimising the average number of divisions. Section 5 applies our methods to the structured development journey of thymocytes, where we generalise our considerations to include asymmetric division; that is, a division event that leaves one daughter cell in the same compartment that the mother cell divided and the other daughter cell exits the compartment. Additional details have been provided in the appendices: the recursion relations to obtain the probability that k cells exit from one or two compartments are given in Appendix A; the variance of the random variable $\mathbf{R}$ is calculated in Appendix B; and the generalisation of our methods to include asymmetric division is presented in Appendix C.

Author contributions

All authors contributed to research design. F.F., C.M.P. and G.L. performed theoretical modeling. F.F. and G.L. performed computer simulations. F.F., C.M.P. and G.L. wrote the first draft of the manuscript. All authors wrote and reviewed the final version of the manuscript.

Acknowledgements

This manuscript has been reviewed at Los Alamos National Laboratory and assigned the report number LA-UR-21-2563.

Data accessibility

Python codes to perform Gillespie simulations to generate Figure 6 (Qkhist.py), Figure 11 (Gdist04.py), and Figure 12 (RGdist06.py) are available at `https://doi.org/10.5281/zenodo.7181108`.

Ethics

This article does not present research with ethical considerations.

Funding statement

This work has been supported by the European Commission through the Marie Skłodowska-Curie Action (H2020-MSCA-ITN-2017), Innovative Training Network Quantitative T cell Immunology and Immunotherapy (QuanTII), project number 764698 (FF, CMP and GL). We acknowledge the support from the School of Mathematics of the University of Leeds to HD (Daphne Jackson Fellow and visiting post-doctoral fellow). This work has been partially funded by grant PID2019-106339GB-I00 from MCIN/AEI/10.13039/501100011033 (MC). Research presented in this article was supported by the Laboratory Directed Research and Development program of Los Alamos National Laboratory under project number 20220754ER awarded to C.M.P.

References

[1] Katrin Busch, Kay Klapproth, Melania Barile, Michael Flossdorf, Tim Holland-Letz, Susan M Schlenner, Michael Reth, Thomas Höfer, and Hans-Reimer Rodewald. Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. *Nature*, 518(7540):542–546, 2015.

[2] Thomas Höfer, Melania Barile, and Michael Flossdorf. Stem-cell dynamics and lineage topology from *in vivo* fate mapping in the hematopoietic system. *Current opinion in biotechnology*, 39:150–156, 2016.

[3] Catherine M Sawai, Sonja Babovic, Samik Upadhaya, David JHF Knapp, Yonit Lavin, Colleen M Lau, Anton Goloborodko, Jue Feng, Joji Fujisaki, Lei Ding, et al. Hematopoietic stem cells are the major source of multilineage hematopoiesis in adult animals. *Immunity*, 45(3):597–609, 2016.

[4] V. Thomas-Vaslin, H.K. Altes, R.J. de Boer, and D. Klatzmann. Comprehensive assessment and mathematical modeling of T cell population dynamics and homeostasis. *Journal of Immunology*, 180(4):2240, 2008.

[5] Matthew D Johnston, Carina M Edwards, Walter F Bodmer, Philip K Maini, and S Jonathan Chapman. Mathematical modeling of cell population dynamics in the colonic crypt and in colorectal cancer. *Proceedings of the National Academy of Sciences*, 104(10):4008–4013, 2007.

[6] Philippe A Robert, Heike Kunze-Schumacher, Victor Greiff, and Andreas Krueger. Modeling the dynamics of T-cell development in the thymus. *Entropy*, 23(4):437, 2021.

[7] James E Till, Ernest A McCulloch, and Louis Siminovitch. A stochastic model of stem cell proliferation, based on the growth of spleen colony-forming cells. *Proceedings of the National Academy of Sciences of the United States of America*, 51(1):29, 1964.

[8] Ken R Duffy and Philip D Hodgkin. Intracellular competition for fates in the immune system. *Trends in cell biology*, 22(9):457–464, 2012.

[9] Carmen Gerlach, Jan C Rohr, Leïla Perié, Nienke van Rooij, Jeroen WJ van Heijst, Arno Velds, Jos Urbanus, Shalin H Naik, Heinz Jacobs, Joost B Beltman, Rob J. de Boer, and Ton N. M. Schumacher. Heterogeneous differentiation patterns of individual CD8+ T cells. *Science*, 340(6132):635–639, 2013.

[10] Leïla Perié, Philip D Hodgkin, Shalin H Naik, Ton N Schumacher, Rob J de Boer, and Ken R Duffy. Determining lineage pathways from cellular barcoding experiments. *Cell reports*, 6(4):617–624, 2014.

[11] Veit R Buchholz, Ton NM Schumacher, and Dirk H Busch. T cell fate at the single-cell level. *Annual review of immunology*, 34:65–92, 2016.

[12] Alexander S Miles, Philip D Hodgkin, and Ken R Duffy. Inferring differentiation order in adaptive immune responses from population-level data. In *Mathematical, Computational and Experimental T Cell Immunology*, pages 133–149. Springer, 2021.

[13] H Hamlet Chu, Shiao-Wei Chan, John Paul Gosling, Nicolas Blanchard, Alexandra Tsitsiklis, Grant Lythe, Nilabh Shastri, Carmen Molina-París, and Ellen A Robey. Continuous effector CD8$^+$ T cell production in a controlled persistent infection is sustained by a proliferative intermediate population. *Immunity*, 45(1):159–171, 2016.

[14] Maria Sawicka, Gretta L Stritesky, Joseph Reynolds, Niloufar Abourashchi, Grant Lythe, Carmen Molina-París, and Kristin A Hogquist. From pre-DP, post-DP, SP4, and SP8 thymocyte cell counts to a dynamical model of cortical and medullary selection. *Frontiers in Immunology*, 5, 2014.

[15] Andreas Krueger, Natalia Zietara, and Marcin Łyszkiewicz. T cell development by the numbers. *Trends in immunology*, 38(2):128–139, 2017.

[16] Charles Sinclair, Iren Bains, Andrew J Yates, and Benedict Seddon. Asymmetric thymocyte death underlies the CD4:CD8 T-cell ratio in the adaptive immune system. *Proceedings of the National Academy of Sciences*, 110(31):E2905–E2914, 2013.

[17] Andrew Yates. Theories and quantification of thymic selection. *Frontiers in immunology*, 5:13, 2014.

[18] Ineke den Braber, Tendai Mugwagwa, Nienke Vrisekoop, Liset Westera, Ramona Mögling, Anne Bregje de Boer, Neeltje Willems, Elise HR Schrijver, Gerrit Spierenburg, Koos Gaiser, Erik Mul, Sigrid A. Otto, An F.C. Ruiter, Mariette T. Ackermans, Frank Miedema, José A.M. Borghans, Rob J. de Boer, and Kiki Tesselaar. Maintenance of peripheral naive T cells is sustained by thymus output in mice but not humans. *Immunity*, 36(2):288–297, 2012.

[19] Thea Hogan, Graeme Gossel, Andrew J Yates, and Benedict Seddon. Temporal fate mapping reveals age-linked heterogeneity in naive T lymphocytes in mice. *Proceedings of the National Academy of Sciences*, 112(50):E6917–E6926, 2015.

[20] Kaspar Bresser, Lianne Kok, Arpit C Swain, Lisa A King, Laura Jacobs, Tom S Weber, Leïla Perié, Ken R Duffy, Rob J de Boer, Ferenc A Scheeren, et al. Replicative history marks transcriptional and functional disparity in the CD8+ T cell memory pool. *Nature Immunology*, pages 1–11, 2022.

[21] Janis L Abkowitz, Daniela Golinelli, David E Harrison, and Peter Guttorp. In vivo kinetics of murine hemopoietic stem cells. *Blood*, 96(10):3399–3405, 2000.

[22] Ron Sender and Ron Milo. The distribution of cellular turnover in the human body. *Nature medicine*, 27(1):45–48, 2021.

[23] Jason Cosgrove, Lucie SP Hustin, Rob J de Boer, and Leïla Perié. Hematopoiesis in numbers. *Trends in immunology*, 42(12):1100–1112, 2021.

[24] Nils B Becker, Matthias Günther, Congxin Li, Adrien Jolly, and Thomas Höfer. Stem cell homeostasis by integral feedback through the niche. *Journal of Theoretical Biology*, 481:100–109, 2019.

[25] Thomas Hofer, Hans-Reimer Rodewald, Katrin Busch, Ann-Kathrin Fanti, Alessandro Greco, Xi Wang, Qin Zhang, Melania Barile, Hideyuki Oguro, and Sean J Morrison. Hematopoietic stem cells self-renew symmetrically or gradually proceed to differentiation. *bioRxiv*, 2020.

[26] Cristian Tomasetti and Bert Vogelstein. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, 347(6217):78–81, 2015.

[27] Cristian Tomasetti, Lu Li, and Bert Vogelstein. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*, 355(6331):1330–1334, 2017.

[28] Robert L Bowman, Lambert Busque, and Ross L Levine. Clonal hematopoiesis and evolution to hematopoietic malignancies. *Cell stem cell*, 22(2):157–170, 2018.

[29] Vera C Martins, Eliana Ruggiero, Susan M Schlenner, Vikas Madan, Manfred Schmidt, Pamela J Fink, Christof von Kalle, and Hans-Reimer Rodewald. Thymus-autonomous T cell development in the absence of progenitor import. *Journal of Experimental Medicine*, 209(8):1409–1417, 2012.

[30] Luna Ballesteros-Arias, Joana G Silva, Rafael A Paiva, Belén Carbonetto, Pedro Faísca, and Vera C Martins. T cell acute lymphoblastic leukemia as a consequence of thymus autonomy. *Journal of Immunology*, 202(4):1137–1144, 2019.

[31] Laetitia Peaudecerf, Sara Lemos, Alessia Galgano, Gerald Krenn, Florence Vasseur, James P Di Santo, Sophie Ezine, and Benedita Rocha. Thymocytes may persist and differentiate without any input from bone marrow progenitors. *Journal of Experimental Medicine*, 209(8):1401–1408, 2012.

[32] Thomas Boehm. Self-renewal of thymocytes in the absence of competitive precursor replenishment. *Journal of Experimental Medicine*, 209(8):1397–1400, 2012.

[33] Makio Ogawa. Differentiation and proliferation of hematopoietic stem cells. *Blood*, 81:2844–2853, 1993.

[34] Janis L Abkowitz, Sandra N Catlin, and Peter Guttorp. Evidence that hematopoiesis may be a stochastic process in vivo. *Nature medicine*, 2(2):190–197, 1996.

[35] Tannishtha Reya, Sean J Morrison, Michael F Clarke, and Irving L Weissman. Stem cells, cancer, and cancer stem cells. *nature*, 414(6859):105–111, 2001.

[36] Jason Xu, Yiwen Wang, Peter Guttorp, and Janis L Abkowitz. Visualizing hematopoiesis as a stochastic process. *Blood advances*, 2(20):2637–2645, 2018.

[37] Ingo Roeder, Matthias Horn, Ingmar Glauche, Andreas Hochhaus, Martin C Mueller, and Markus Loeffler. Dynamic modeling of imatinib-treated chronic myeloid leukemia: functional insights and clinical implications. *Nature medicine*, 12(10):1181–1184, 2006.

[38] Dániel Grajzel, Imre Derényi, and Gergely J Szöllősi. A compartment size-dependent selective threshold limits mutation accumulation in hierarchical tissues. *Proceedings of the National Academy of Sciences*, 117(3):1606–1611, 2020.

[39] Tamar Tak, Giulio Prevedello, Gaël Simon, Noémie Paillon, Camélia Benlabiod, Caroline Marty, Isabelle Plo, Ken R Duffy, and Leïla Perié. HSPCs display within-family homogeneity in differentiation and proliferation despite population heterogeneity. *Elife*, 10:e60624, 2021.

[40] JS Wainscoat and MF Fey. Assessment of clonality in human tumors: a review. *Cancer Research*, 50(5):1355–1360, 1990.

[41] Anne-Marie Lyne, Lucie Laplane, and Leïla Perié. To portray clonal evolution in blood cancer, count your stem cells. *Blood*, 137(14):1862–1870, 2021.

[42] H EM Kay. How many cell-generations? *The Lancet*, 286(7409):418–419, 1965.

[43] Benjamin Werner, David Dingli, and Arne Traulsen. A deterministic model for the occurrence and dynamics of multiple mutations in hierarchically organized tissues. *Journal of The Royal Society Interface*, 10(85):20130349, 2013.

[44] ML Samuels. Distribution of the branching-process population among generations. *Journal of Applied Probability*, 8(4):655–667, 1971.

[45] Gianfelice Meli, Tom S Weber, and Ken R Duffy. Sample path properties of the average generation of a bellman–harris process. *Journal of mathematical biology*, 79(2):673–704, 2019.

[46] Marvin A Böttcher, David Dingli, Benjamin Werner, and Arne Traulsen. Replicative cellular age distributions in compartmentalized tissues. *Journal of The Royal Society Interface*, 15(145):20180272, 2018.

[47] Tibor Antal and PL Krapivsky. Exact solution of a two-type branching process: models of tumor progression. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(08):P08018, 2011.

[48] Salvador E Luria and Max Delbrück. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28(6):491, 1943.

[49] Hal Caswell. *Matrix population models*, volume 1. Sinauer Sunderland, MA, 2000.

[50] Kenneth Zierler. A critique of compartmental analysis. *Annual review of biophysics and bioengineering*, 10(1):531–562, 1981.

[51] Caroline Colijn and Michael C Mackey. A mathematical model of hematopoiesis—i. periodic chronic myelogenous leukemia. *Journal of Theoretical Biology*, 237(2):117–132, 2005.

[52] Franziska Michor, Timothy P Hughes, Yoh Iwasa, Susan Branford, Neil P Shah, Charles L Sawyers, and Martin A Nowak. Dynamics of chronic myeloid leukaemia. *Nature*, 435(7046):1267–1270, 2005.

[53] Zakary L Whichard, Casim A Sarkar, Marek Kimmel, and Seth J Corey. Hematopoiesis and its disorders: a systems biology approach. *Blood*, 115(12):2339–2347, 2010.

[54] Benjamin Werner, David Dingli, Tom Lenaerts, Jorge M Pacheco, and Arne Traulsen. Dynamics of mutant cells in hierarchical organized tissues. *PLoS Computational Biology*, 7(12):e1002290, 2011.

[55] T. E. Harris. *The theory of branching processes.* Springer-Verlag, Berlin, 1963.

[56] M. Kimmel and D. E. Axelrod. *Branching Processes in Biology.* Springer, 2002.

[57] Gretta L Stritesky, Yan Xing, Jami R Erickson, Lokesh A Kalekar, Xiaodan Wang, Daniel L Mueller, Stephen C Jameson, and Kristin A Hogquist. Murine thymic selection quantified using a unique method to capture deleted T cells. *Proceedings of the National Academy of Sciences*, 110(12):4679–4684, 2013.

[58] JM Marchingo, G Prevedello, A Kan, S Heinzel, PD Hodgkin, and KR Duffy. T-cell stimuli independently sum to regulate an inherited clonal division fate. *Nature Communications*, 7(1):1–12, 2016.

[59] J. Michael Steel. *Stochastic Calculus and Financial Applications.* Springer, 2001.

[60] Herbert S Wilf. *generatingfunctionology.* CRC press, 2005.

[61] David Singmaster. An elementary evaluation of the Catalan numbers. *The American Mathematical Monthly*, 85(5):366–368, 1978.

[62] Ronald D Dutton and Robert C Brigham. Computationally efficient bounds for the Catalan numbers. *European Journal of Combinatorics*, 7(3):211–213, 1986.

[63] Donald E Knuth and Herbert S Wilf. A short proof of Darboux's lemma. *Applied Mathematics Letters*, 2(4):iii–iv, 1989.

[64] Daniel H Greene and Donald E Knuth. *Mathematics for the Analysis of Algorithms.* Birkhauser, Boston-Basel-Stuttgart,, 1990.

[65] Philippe Flajolet and Robert Sedgewick. *Analytic combinatorics.* Cambridge University Press, 2009.

[66] James P Gleeson, Jonathan A Ward, Kevin P O'sullivan, and William T Lee. Competition-induced criticality in a model of meme popularity. *Physical Review Letters*, 112(4):048701, 2014.

[67] Susan M Kaech, E John Wherry, and Rafi Ahmed. Effector and memory T-cell differentiation: implications for vaccine development. *Nature Reviews Immunology*, 2(4):251–262, 2002.

[68] Jhagvaral Hasbold, Lynn M Corcoran, David M Tarlinton, Stuart G Tangye, and Philip D Hodgkin. Evidence from the generation of immunoglobulin G–secreting cells that stochastic mechanisms regulate lymphocyte differentiation. *Nature immunology*, 5(1):55–63, 2004.

[69] Thomas Höfer and Hans-Reimer Rodewald. Differentiation-based model of hematopoietic stem cell functions and lineage pathways. *Blood, The Journal of the American Society of Hematology*, 132(11):1106–1113, 2018.

[70] Ken R Duffy, Gianfelice Meli, and Seva Shneer. The variance of the average depth of a pure birth process converges to 7. *Statistics and Probability Letters*, 150:88–93, 2019.

[71] D. Stirzaker. *Stochastic processes and models.* Oxford University Press, 2005.

[72] Linda Partridge and David Gems. Mechanisms of aging: public or private? *Nature Reviews Genetics*, 3(3):165–175, 2002.

[73] Nicole F Mathon and Alison C Lloyd. Cell senescence and cancer. *Nature Reviews Cancer*, 1(3):203–213, 2001.

[74] Mary Philip and Andrea Schietinger. CD8+ T cell differentiation and dysfunction in cancer. *Nature Reviews Immunology*, pages 1–15, 2021.

[75] Arne N Akbar and Sian M Henson. Are senescence and exhaustion intertwined or unrelated processes that compromise immunity? *Nature Reviews Immunology*, 11(4):289–295, 2011.

[76] Imre Derényi and Gergely J Szöllősi. Hierarchical tissue organization as a general mechanism to limit the accumulation of somatic mutations. *Nature Communications*, 8(1):1–8, 2017.

[77] Benjamin Werner, Fabian Beier, Sebastian Hummel, Stefan Balabanov, Lisa Lassay, Thorsten Orlikowsky, David Dingli, Tim H Brümmendorf, and Arne Traulsen. Reconstructing the in vivo dynamics of hematopoietic stem cells from telomere length distributions. *eLife*, 4:e08687, 2015.

[78] Jienian Yang, Maksim V Plikus, and Natalia L Komarova. The role of symmetric stem cell divisions in tissue homeostasis. *PLoS Computational Biology*, 11(12):e1004629, 2015.

[79] Thomas Stiehl and Anna Marciniak-Czochra. Stem cell self-renewal in regeneration and cancer: insights from mathematical modeling. *Current Opinion in Systems Biology*, 5:112–120, 2017.

[80] Leili Shahriyari and Natalia L Komarova. Symmetric vs. asymmetric stem cell divisions: an adaptation against cancer? *PLoS ONE*, 8(10):e76195, 2013.

[81] Kim Pham, Raz Shimoni, Mirren Charnley, Mandy J Ludford-Menting, Edwin D Hawkins, Kelly Ramsbottom, Jane Oliaro, David Izon, Stephen B Ting, Joseph Reynolds, et al. Asymmetric cell division during T cell development controls downstream fate. *Journal of Cell Biology*, 210(6):933–950, 2015.

[82] Melania Barile, Katrin Busch, Ann-Kathrin Fanti, Alessandro Greco, Xi Wang, Hideyuki Oguro, Qin Zhang, Sean J Morrison, Hans-Reimer Rodewald, and Thomas Höfer. Hematopoietic stem cells self-renew symmetrically or gradually proceed to differentiation. *Available at SSRN 3787896*, 2020.

[83] Michael Flossdorf, Jens Rössler, Veit R Buchholz, Dirk H Busch, and Thomas Höfer. CD8+ T cell diversification by asymmetric cell division. *Nature immunology*, 16(9):891–893, 2015.

[84] John T Chang, Vikram R Palanivel, Ichiko Kinjyo, Felix Schambach, Andrew M Intlekofer, Arnob Banerjee, Sarah A Longworth, Kristine E Vinup, Paul Mrass, Jane Oliaro, et al. Asymmetric T lymphocyte division in the initiation of adaptive immune responses. *Science*, 315(5819):1687–1691, 2007.

[85] Mariana Borsa, Isabel Barnstorf, Nicolas S Baumann, Katharina Pallmer, Alexander Yermanos, Fabienne Gräbnitz, Niculò Barandun, Annika Hausmann, Ioana Sandu, Yves Barral, et al. Modulation of asymmetric cell division as a mechanism to boost CD8+ T cell memory. *Science immunology*, 4(34), 2019.

[86] Da Zhou, Yue Luo, David Dingli, and Arne Traulsen. The invasion of de-differentiating cancer cells into hierarchical tissues. *PLoS Computational Biology*, 15(7):e1007167, 2019.

[87] Maria Ciofani and Juan Carlos Zúñiga-Pflücker. Determining $\gamma\delta$ versus $\alpha\beta$ T cell development. *Nature Reviews Immunology*, 10(9):657–663, 2010.

[88] Tessa Crompton, Susan V Outram, and Ariadne L Hager-Theodorides. Sonic hedgehog signalling in T-cell development and activation. *Nature Reviews Immunology*, 7(9):726–735, 2007.

[89] Benedict Seddon and Andrew J Yates. The natural history of naive T cells from birth to maturity. *Immunological Reviews*, 285(1):218–232, 2018.

[90] Grant Lythe, Robin E Callard, Rollo L Hoare, and Carmen Molina-París. How many TCR clonotypes does a body maintain? *Journal of Theoretical Biology*, 389:214–224, 2016.

[91] Pedro Gonçalves, Marco Ferrarini, Carmen Molina-Paris, Grant Lythe, Florence Vasseur, Annik Lim, Benedita Rocha, and Orly Azogui. A new mechanism shapes the naïve CD8+ T cell repertoire: The selection for full diversity. *Molecular Immunology*, 85:66–80, 2017.

[92] Grant Lythe and Carmen Molina-París. Some deterministic and stochastic mathematical models of naïve T-cell homeostasis. *Immunological reviews*, 285(1):206–217, 2018.

[93] Jonathan Desponds, Thierry Mora, and Aleksandra M Walczak. Fluctuating fitness shapes the clone-size distribution of immune repertoires. *Proceedings of the National Academy of Sciences*, 113(2):274–279, 2016.

[94] Peter C de Greef, Theres Oakes, Bram Gerritsen, Mazlina Ismail, James M Heather, Rutger Hermsen, Benjamin Chain, and Rob J de Boer. The naive T-cell receptor repertoire has an extremely broad distribution of clone sizes. *Elife*, 9:e49900, 2020.

[95] Madhura Mukhopadhyay. Reporting T cell proliferation. *Nature Methods*, 19(5):521–521, 2022.

A Recursion relation for the probabilities $Q_k(C)$

In principle, the whole distribution of a random variable can be obtained once its probability generating function is known. In practice, an algorithm is required to compute the numerical values of the desired probabilities [66]. Here, we describe equations that we have used, relating the probability that the random variable $\mathbf{R}$ is equal to k to the probability that it is equal to $k-1$, in the simplest case ($C=1$), and to $k-1$ and $k-2$ in other cases ($C=2$).

A.1 Recursion relation: a single compartment

In the case $C=1$, we rewrite (10) as $2p_{\mathrm{b}}\phi(z) = 1 - w(z)$, where $w^2(z) = 1 - 4p_{\mathrm{b}}p_{\mathrm{d}} - 4p_{\mathrm{b}}p_{\mathrm{e}}z$. We now compute the first derivative of $\phi(z)$. One can show that $w(z)\phi'(z) = p_{\mathrm{e}}$ and that $\phi(z)$ satisfies the following differential equation

$$w^2(z)\phi'(z) + 2p_{\mathrm{e}}p_{\mathrm{b}}\phi(z) - p_{\mathrm{e}} = 0. \tag{56}$$

Inserting $\phi(z) = \sum_{k=0}^{+\infty} q_k z^k$ in (56), and matching terms proportional to z^k yields the recursion relation (15).

A.2 Recursion relation: two compartments

We next consider the case $C=2$. In what follows we obtain a differential equation for $\Phi_2(z)$ of the form

$$T(z)\Phi_2''(z) + R(z)\Phi_2'(z) + S(z)\left(1 - 2p_{\mathrm{b}}(1)\Phi_2(z)\right) = 0, \tag{57}$$

with $T(z)$, $R(z)$ and $S(z)$ polynomials in z (with real coefficients), and given by

$$T(z) = t_0 + t_1 z + t_2 z^2, \quad R(z) = r_0 + r_1 z, \quad \text{and} \quad S(z) = s_0.$$

Then, given that $\Phi_2(z) = \phi_1(\phi_2(z)) = \sum_{k=0}^{+\infty} Q_k z^k$,

$$t_0(k+2)(k+1)Q_{k+2} + [t_1k^2 + (r_0+t_1)k + r_0]Q_{k+1} + [t_2k^2 + (r_1-t_2)k + s_0]Q_k = 0. \tag{58}$$

Let us write $\Delta_c^2 = 1 - 4p_{\mathrm{d}}(c)p_{\mathrm{b}}(c)$ and $w_c^2(z) = \Delta_c^2 - 4p_{\mathrm{b}}(c)p_{\mathrm{e}}(c)z$, for $c = 1, 2$. We have $2p_{\mathrm{b}}(1)\Phi_2(z) = 1 - w_1(\phi_2(z))$ and

$$\Phi_2'(z) = \frac{p_{\mathrm{e}}(1)}{w1}\phi_2'(z) = \frac{p_{\mathrm{e}}(1)p_{\mathrm{e}}(2)}{w_1w_2}, \tag{59}$$

where w_1 is shorthand for $w_1(\phi_2(z))$, and w_2 is shorthand for $w_2(z)$. Now, we compute the second derivative of $\Phi_2(z)$:

$$\Phi_2''(z) = \frac{2p_{\mathrm{e}}(1)p_{\mathrm{e}}^2(2)}{w_1^3w_2^3}\left[p_{\mathrm{b}}(1)p_{\mathrm{e}}(1)w_2 + p_{\mathrm{b}}(2)w_1^2\right]. \tag{60}$$

Multiplying through by $w_1^3w_2^3$, we can write

$$2p_{\mathrm{e}}(1)p_{\mathrm{e}}^2(2)\left[p_{\mathrm{b}}(2)w_1^2 + p_{\mathrm{b}}(1)p_{\mathrm{e}}(1)w_2\right]T(z) + p_{\mathrm{e}}(1)p_{\mathrm{e}}(2)w_2^2w_1^2R(z) + w_2^3w_1^4S(z) = 0. \tag{61}$$

We make use of the fact that $1 - 2p_{\mathrm{b}}(1)\Phi_2(z) = w_1$ and that $w_1^2 = \Delta_1^2 - \kappa + \kappa w_2$, where $\kappa = 2p_{\mathrm{e}}(1)\frac{p_{\mathrm{b}}(1)}{p_{\mathrm{b}}(2)}$. Equating terms proportional to w_2^2, w_2^3, w_2^4 and w_2^5, we find

$$T(z) = T_2w_2^2 + T_4w_2^4, \quad R(z) = R_0 + R_2w_2^2, \quad \text{and} \quad S(z) = s_0 = 2p_{\mathrm{b}}(1)p_{\mathrm{e}}^2(1)p_{\mathrm{e}}^2(2), \tag{62}$$

where

$$T_2 = -(\Delta_1^2-\kappa)^2, \quad T_4 = \kappa^2, \quad R_0 = -2p_{\mathrm{b}}(2)p_{\mathrm{e}}(2)T_2, \quad R_2 = -4p_{\mathrm{b}}(2)p_{\mathrm{e}}(2)T_4, \quad \text{and} \quad s_0 = 2p_{\mathrm{b}}(1)p_{\mathrm{e}}^2(1)p_{\mathrm{e}}^2(2).$$

Making use of (62), we obtain

$$t_0 = \Delta_2^2T_2 + \Delta_2^4T_4, \quad t_1 = -4p_{\mathrm{b}}(2)p_{\mathrm{e}}(2)T_2 - 8p_{\mathrm{b}}(2)p_{\mathrm{e}}(2)\Delta_2^2T_4, \quad t_2 = 16p_{\mathrm{b}}^2(2)p_{\mathrm{e}}^2(2)T_4, \quad r_0 = \frac{1}{2}t_1, \quad \text{and} \quad r_1 = t_2.$$

The general two-compartment recursion relation (58) is thus given by

$$\left[\kappa^2\Delta_2^4 - (\Delta_1^2-\kappa)^2\Delta_2^4\right](k+1)(k+2)Q_{k+2} - p_{\mathrm{b}}(2)p_{\mathrm{e}}(2)[2\kappa^2\Delta_2^2 - (\Delta_1^2-\kappa)^2](2k+1)(2k+2)Q_{k+1} \tag{63}$$
$$+p_{\mathrm{b}}^2(2)p_{\mathrm{e}}^2(2)\kappa^2(16k^2-1)Q_k = 0. \tag{64}$$

If $p_{\mathrm{d}}(1) = p_{\mathrm{d}}(2) = 0$, then $\Delta_1 = \Delta_2 = 1$ and (64) takes the simpler form

$$(2\kappa-1)(k+1)(k+2)Q_{k+2} - p_{\mathrm{b}}(2)p_{\mathrm{e}}(2)(\kappa^2+2\kappa-1)(2k+1)(2k+2)Q_{k+1} + p_{\mathrm{b}}^2(2)p_{\mathrm{e}}^2(2)\kappa^2(16k^2-1)Q_k = 0. \tag{65}$$

B The variance of the distribution of family sizes

The distributions of family sizes that we have found have a pattern where the factor $k^{-3/2}$ appears. One consequence of this behaviour is that the relationship between the mean and variance is different from that found in well-known distributions such as the Poisson distribution.

With C compartments, the probability generating function of $\mathbf{R}$ is given by (27), and the variance of $\mathbf{R}$ is given by

$$V = \Phi_C''(1) + N - N^2. \tag{66}$$

We make use of (30), to write $\Phi_C'(z) = \phi_1'(\chi_C(z))\chi_C'(z)$ and $\Phi_C''(1) = \phi_1''(1)\left(\chi_C'(1)\right)^2 + \phi_1'(1)\chi_C''(1)$, where $\Phi_C'(z) = \frac{\mathrm{d}}{\mathrm{d}z}\Phi_C(z)$.

We next assume that $\phi_c(z) = \phi(z)$, $c = 1, \ldots, C$. Then, one can show that

$$\phi'(1) = N^{1/C}, \quad \phi''(1) = 2\frac{p_{\mathrm{b}}}{p_{\mathrm{e}}}N^{3/C}, \quad \text{and} \quad \Phi_C''(1) = 2\frac{p_{\mathrm{b}}}{p_{\mathrm{e}}}\left[N^{3/C}N^{2(1-1/C)}\right] + N^{1/C}\chi_C''(1).$$

We find

$$\Phi_1''(1) = 2\frac{p_\mathrm{b}}{p_\mathrm{e}}N^3, \quad \Phi_2''(1) = 2\frac{p_\mathrm{b}}{p_\mathrm{e}}\left(N^{5/2} + N^2\right), \quad \Phi_3''(1) = 2\frac{p_\mathrm{b}}{p_\mathrm{e}}\left(N^{7/3} + N^2 + N^{5/3}\right), \ldots.$$

That is, we have

$$\Phi_C''(1) = 2\frac{p_\mathrm{b}}{p_\mathrm{e}}N^{2+1/C}\sum_{c=0}^{C-1} N^{-c/C}. \tag{67}$$

The variance of $\mathbf{R}$ is proportional to $N^{2+\frac{1}{C}}$ in the limit $N \to +\infty$ (see Figure 13).

C Compartment analysis in the case of asymmetric division

In an asymmetric division event, one daughter cell transits to the next compartment and the other remains in the compartment. Each cell, independently, may die, divide, undergo asymmetric division, or transit to the next compartment, with probabilities

$$p_\mathrm{d}, \quad p_\mathrm{b}, \quad p_\mathrm{a}, \quad \text{and} \quad p_\mathrm{e},$$

where $p_\mathrm{d} + p_\mathrm{b} + p_\mathrm{e} + p_\mathrm{a} = 1$. The analogue of (H1), guaranteeing extinction in the compartment, is

$$2p_\mathrm{b} + p_\mathrm{a} < 1. \tag{Ha}$$

C.1 Family sizes

Proceeding to the calculation of the q_k as in Section 2, we find that (4) still holds, but (7) and (8) are replaced by $\Delta q_1 = p_\mathrm{e} + p_\mathrm{a} q_0$ and

$$q_k = \frac{p_\mathrm{b}}{\Delta}\sum_{j=1}^{k-1} q_j q_{k-j} + \frac{p_\mathrm{a}}{\Delta}q_{k-1}, \quad k \geq 2. \tag{68}$$

The probability generating function of $\mathbf{R}$ when $C = 1$, denoted by $\psi(z)$, satisfies

$$\psi(z) = p_\mathrm{d} + p_\mathrm{e} z + p_\mathrm{a} z\psi(z) + p_\mathrm{b}\psi^2(z). \tag{69}$$

The solution is given by

$$\psi(z) = \frac{1 - p_\mathrm{a} z - [(1 - p_\mathrm{a} z)^2 - 4p_\mathrm{b} p_\mathrm{d} - 4p_\mathrm{b} p_\mathrm{e} z]^{1/2}}{2p_\mathrm{b}}. \tag{70}$$

Figure 14 compares q_k in this case (asymmetric case) with that of symmetric division only ($p_\mathrm{a} = 0$).

Thus, in the case of asymmetric division, and for $C = 1$, we have

$$N = \frac{p_\mathrm{e} + p_\mathrm{a}}{1 - 2p_\mathrm{b} - p_\mathrm{a}}. \tag{71}$$

Remark C.1 If $q_k = \mathbb{P}(\mathbf{R} = k)$ then, for $k \geq 2$,

$$\begin{aligned} q_k &= \frac{\Delta}{p_b}\left(\frac{2p_b q_1 + p_a}{2\Delta}\right)^k \sum_{j=0}^{\lfloor k/2 \rfloor} c_{k-j-1}\binom{k-j}{j}\left(\frac{-2p_a^2\Delta}{(2p_a + 4p_b p_e)(2p_b q_1 + p_a)}\right)^j \\ &= \frac{\Delta}{p_b}\left(\frac{2p_b q_1 + p_a}{2\Delta}\right)^k \sum_{j=0}^{\lfloor k/2 \rfloor} \frac{1}{k-j}\binom{2k-2j-1}{k-j}\binom{k-j}{j}\left(\frac{-2p_a^2\Delta}{(2p_a + 4p_b p_e)(2p_b q_1 + p_a)}\right)^j. \end{aligned}$$

Remark C.2 It is convenient to generate q_k via a recursion relation. Following the approach described in Appendix A, we rewrite (70) as

$$2p_\mathrm{b}\psi(z) = 1 - p_\mathrm{a} z - w_a(z), \quad \text{where} \quad w_a^2(z) = \Delta - (2p_\mathrm{a} + 4p_\mathrm{b} p_\mathrm{e})z + p_\mathrm{a}^2 z^2. \tag{72}$$

Thus, $\psi(z)$ satisfies the following differential equation

$$w_a^2(z)\psi'(z) + w_a'(z)w_a(z)\psi(z) + \zeta(z) = 0, \tag{73}$$

with $\zeta(z) = (p_{\mathrm{a}}^2 - p_{\mathrm{a}} + 2p_{\mathrm{a}}p_{\mathrm{b}}p_{\mathrm{e}} - 4p_{\mathrm{b}}p_{\mathrm{e}})z + \Delta^2 - p_{\mathrm{a}} - 2p_{\mathrm{b}}p_{\mathrm{e}}$. Matching terms proportional to z^k leads to the following recursion relation:

$$\Delta^2(k+2)q_{k+2} = (2k+1)(p_{\mathrm{a}} + 2p_{\mathrm{b}}p_{\mathrm{e}})q_{k+1} - (k-1)p_{\mathrm{a}}^2 q_k. \tag{74}$$

We note that in the asymmetric case, even for $C = 1$, the recursion relation is of second order. This is due to the fact that $w_a^2(z)$ is a polynomial of order two in z.

Remark C.3 As $k \to +\infty$, we obtain the following behaviour

$$q_k \propto \gamma_a^k k^{-3/2}, \tag{75}$$

where γ_a satisfies the equation

$$(1 - 4p_{\mathrm{b}}p_{\mathrm{d}})\gamma_a^2 - (2p_{\mathrm{a}} + 4p_{\mathrm{b}}p_{\mathrm{e}})\gamma_a + p_{\mathrm{a}}^2 = 0. \tag{76}$$

Remark C.4 We now consider the case $C = 2$, with two non-identical compartments, *i.e.,* $\psi_1(z) \neq \psi_2(z)$. Let us introduce

$$\Psi_2(z) = \psi_1(\psi_2(z)), \tag{77}$$

and

$$w_{a,c}^2(z) = 1 - 4p_{\mathrm{b}}(c)p_{\mathrm{d}}(c) - [2p_{\mathrm{a}}(c) + 4p_{\mathrm{b}}(c)p_{\mathrm{e}}(c)]z + p_{\mathrm{a}}^2(c)z^2, \quad c = 1, 2. \tag{78}$$

Then, one can show that

$$2p_{\mathrm{b}}(1)\Psi_2(z) = H_1(z) - H_2(z), \tag{79}$$

where $H_1(z) = 1 - \dfrac{p_{\mathrm{a}}(1)}{2p_{\mathrm{b}}(2)} + \dfrac{p_{\mathrm{a}}(1)p_{\mathrm{a}}(2)}{2p_{\mathrm{b}}(2)} - \dfrac{p_{\mathrm{a}}(1)}{2p_{\mathrm{b}}(2)}w_{a2}(z)$, and

$$\begin{aligned} H_2(z)^2 \quad &= \quad \frac{p_{\mathrm{a}}^2(1)p_{\mathrm{a}}^2(2)}{2p_{\mathrm{b}}^2(2)}z^2 + \left(\frac{p_{\mathrm{a}}(2)}{p_{\mathrm{b}}(2)}(p_{\mathrm{a}}(1) + 2p_{\mathrm{b}}(1)p_{\mathrm{e}}(1) - \frac{p_{\mathrm{a}}^2(1)}{p_{\mathrm{b}}^2(2)}(p_{\mathrm{a}}(2) + p_{\mathrm{b}}(2)p_{\mathrm{e}}(2))\right) z \\ &+ \quad \left(\frac{p_{\mathrm{a}}(1) + 2p_{\mathrm{b}}(1)p_{\mathrm{e}}(1)}{p_{\mathrm{b}}(2)} - \frac{p_{\mathrm{a}}^2(1)(1 - p_{\mathrm{a}}(2)z)}{2p_{\mathrm{b}}^2(2)}\right) w_{a,2}(z) \\ &+ \quad \Delta^2(1) + \frac{p_{\mathrm{a}}^2(1)}{2p_{\mathrm{b}}^2(2)}(1 - 2p_{\mathrm{d}}(2)p_{\mathrm{b}}(2)) - \frac{p_{\mathrm{a}}(1) + 2p_{\mathrm{b}}(1)p_{\mathrm{e}}(1)}{p_{\mathrm{b}}(2)}. \end{aligned}$$

In this instance, for the asymmetric case with $C = 2$, and to calculate the distribution of probabilities, $Q_k(2)$, we must compute two recursion relations: one for $H_1(z)$ and a second one for $H_2(z)$. This strategy leads to a three-term recursion relation for $H_1(z)$, and a six-term recursion relation for $H_2(z)$.

C.2 Generation analysis

To define the random variable $\mathbf{G}$, we begin with three simple random variables $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$, with state spaces $\{0, 1, 2\}$, $\{0, 1\}$, and $\{0, 1\}$, respectively, where

$$\mathbb{P}(\mathbf{U} = 0) = 1 - p_{\mathrm{b}} - p_{\mathrm{a}}, \quad \mathbb{P}(\mathbf{U} = 1) = p_{\mathrm{a}}, \quad \mathbb{P}(\mathbf{U} = 2) = p_{\mathrm{b}},$$

$$\mathbb{P}(\mathbf{V} = 0) = 1 - p_{\mathrm{e}}, \quad \mathbb{P}(\mathbf{V} = 1) = p_{\mathrm{e}} \quad \text{and} \quad \mathbb{P}(\mathbf{W} = 0) = 1 - p_{\mathrm{a}}, \quad \mathbb{P}(\mathbf{W} = 1) = p_{\mathrm{a}}.$$

Let us introduce, as we did in the case of symmetric division, $\mathbf{Z}_0 = 1$ and

$$\mathbf{Z}_{n+1} = \sum_{i=1}^{\mathbf{Z}_n} \mathbf{U}_i, \quad n = 0, 1, 2, \ldots, \tag{80}$$

where, for each i, $\mathbf{U}_i$ is an independent copy of $\mathbf{U}$ (as defined above). The definition (35) still holds, but (36) is replaced by

$$\mathbf{Y}_n = \sum_{i=1}^{\mathbf{Z}_n} \mathbf{V}_i + \sum_{i=1}^{\mathbf{Z}_{n-1}} \mathbf{W}_i, \quad n = 0, 1, 2, \ldots, \tag{81}$$

where each $\mathbf{V}_i$, $\mathbf{W}_i$ are, respectively, an independent copy of $\mathbf{V}$ and $\mathbf{W}$. The mean values of $\mathbf{Y}_n$ for $n \geq 0$, which generalise (38), are given by $\mathbb{E}(\mathbf{Y}_0) = p_\mathrm{e}$ and

$$\mathbb{E}(\mathbf{Y}_n) = p_\mathrm{e}\mathbb{E}(\mathbf{Z}_n) + p_\mathrm{a}\mathbb{E}(\mathbf{Z}_{n-1}) = p_\mathrm{e}(2p_\mathrm{b} + p_\mathrm{a})^n + p_\mathrm{a}(2p_\mathrm{b} + p_\mathrm{a})^{n-1}.$$

Once again, and due to condition (Ha), in the limit $n \to +\infty$, $\mathbb{E}(\mathbf{Z}_n) \to 0$ and $\mathbb{E}(\mathbf{Y}_n) \to 0$. We are interested in obtaining the probability generating function of $\mathbf{G}$. Making use of the definition of the random variables $\mathbf{R}$ and $\mathbf{G}$, the probability generating function of $\mathbf{G}$ is given by

$$\xi(z) = \frac{1}{N}\sum_{n=0}^{+\infty} \mathbb{E}(\mathbf{Y}_n)z^n = \frac{1}{N}\left[p_\mathrm{e} + \sum_{n=1}^{+\infty} \mathbb{E}(\mathbf{Y}_n)z^n\right] = \frac{p_\mathrm{e} + p_\mathrm{a}z}{N(1 - (2p_\mathrm{b} + p_\mathrm{a})z)}. \tag{82}$$

This allows us to compute the expectation value of $\mathbf{G}$ in the asymmetric case:

$$\mathbb{E}(\mathbf{G}) = D = \frac{1}{p_\mathrm{e} + p_\mathrm{a}} \frac{p_\mathrm{a} + p_\mathrm{e}(2p_\mathrm{b} + p_\mathrm{a})}{1 - 2p_\mathrm{b} - p_\mathrm{a}}. \tag{83}$$

The variance of $\mathbf{G}$ is also computed from (82):

$$\mathrm{var}(\mathbf{G}) = \frac{2p_\mathrm{a} + p_\mathrm{e}}{p_\mathrm{e}} D(D + 1). \tag{84}$$

Remark C.5 In the case of asymmetric division, we can choose N, p_a, and p_d as the three independent parameters, so that (40) is given by

$$D = \frac{2N - 1}{1 + p_\mathrm{a} - 2p_\mathrm{d}} \left(1 + \frac{p_\mathrm{a}}{N}\right) - 1. \tag{85}$$

Figure 15, constructed using (85), summarises the effect of asymmetric division (as compared to Figure 8).

Remark C.6 We may express all single-compartment quantities in terms of N, D, and p_a, to obtain

$$p_\mathrm{b} = \frac{N[D(1 - p_\mathrm{a}) - p_\mathrm{a}] - p_\mathrm{a}}{2N(D + 1)}, \quad p_\mathrm{e} = \frac{N - p_\mathrm{a}D}{D + 1}, \quad \text{and} \quad p_\mathrm{d} = \frac{N[2 + D(1 + p_\mathrm{a}) - 2N - p_\mathrm{a}] + p_\mathrm{a}}{2N(D + 1)}.$$

Note that, if we set $p_\mathrm{a} = 0$ then all quantities simplify to the values derived in Section 4.

Remark C.7 In the case of $C > 1$ compartments, and asymmetric division, we define for $c = 1, \ldots, C$ the following random variables $\mathbf{U}(c)$, $\mathbf{V}(c)$ and $\mathbf{W}(c)$:

$$\mathbb{P}\left(\mathbf{U}(c) = 0\right) = 1 - p_\mathrm{b}(c) - p_\mathrm{a}(c), \quad \mathbb{P}\left(\mathbf{U}(c) = 1\right) = p_\mathrm{a}(c), \quad \mathbb{P}\left(\mathbf{U}(c) = 2\right) = p_\mathrm{b}(c),$$

$$\mathbb{P}\left(\mathbf{V}(c) = 0\right) = 1 - p_\mathrm{e}(c), \quad \mathbb{P}\left(\mathbf{V}(c) = 1\right) = p_\mathrm{e}(c) \quad \text{and}$$

$$\mathbb{P}\left(\mathbf{W}(c) = 0\right) = 1 - p_\mathrm{a}(c), \quad \mathbb{P}\left(\mathbf{W}(c) = 1\right) = p_\mathrm{a}(c).$$

We have in this case $\mathbf{Z}_0(1) = 1$, $\mathbf{Z}_0(c) = \mathbf{Y}_0(c - 1)$ for $c \geq 2$, and

$$\mathbf{Z}_{n+1}(c) = \mathbf{Y}_{n+1}(c - 1) + \sum_{i=1}^{\mathbf{Z}_n(c)} \mathbf{U}_i(c), \quad c = 2, \ldots, C, \quad n = 0, 1, 2, \ldots, \tag{86}$$

and

$$\mathbf{Y}_n(c) = \sum_{i=1}^{\mathbf{Z}_n(c)} \mathbf{V}_i(c) + \sum_{i=1}^{\mathbf{Z}_{n-1}(c)} \mathbf{W}_i(c), \quad c = 2, \ldots, C, \quad n = 1, 2, \ldots. \tag{87}$$

The probability generating function of $\mathbf{G}$ is given by a product of single-compartment generating functions making use of (49).

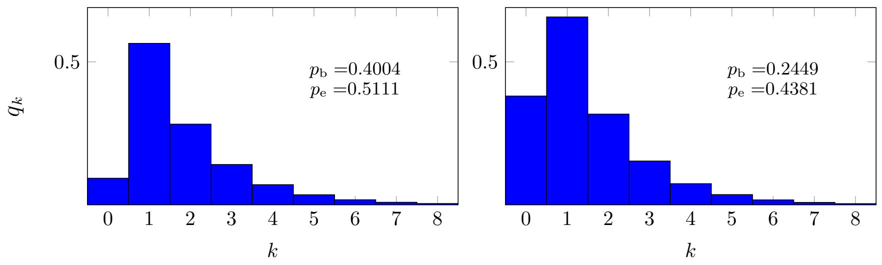

Figure 4: The quantity q_k is the probability that k cells exit a compartment, descended from one progenitor cell. Results, using (11), are shown for two different choices of p_b and p_e. On the left, we use the estimates of Sawicka *et al.* [14]: $p_\mathrm{b} = 0.4004$ and $p_\mathrm{d} = 0.0885$ for SP4 thymocytes. On the right, their estimates for SP8 thymocytes: $p_\mathrm{b} = 0.2449$ and $p_\mathrm{d} = 0.3170$.

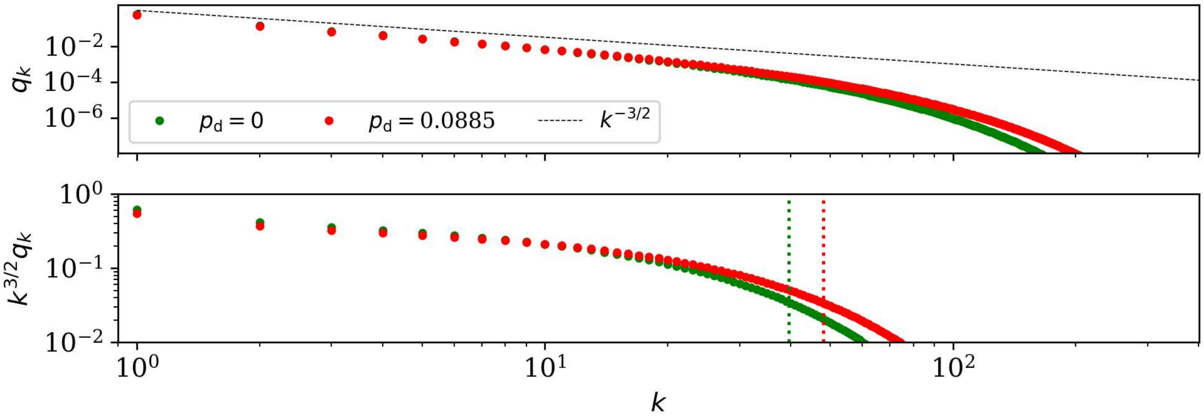

Figure 5: Top plot: the probability, q_k [using (11) and (15)], that the number of product cells is k, logarithmic scales, with and without death. The dashed line is the power law $q_k = k^{-3/2}$. Lower plot: $k^{3/2}q_k$ in the same two cases. The vertical dotted lines, at $k = 6N^2/(1-2p_\mathrm{d})$, indicate where the power law ceases to be an accurate approximation. The parameter values, calculated using (12) so that $N = 2.57$ in both cases, are $p_\mathrm{d} = 0$, $p_\mathrm{b} = 0.455$, $p_\mathrm{e} = 0.545$, and $p_\mathrm{d} = 0.0885$, $p_\mathrm{b} = 0.4004$, $p_\mathrm{e} = 0.5111$. The latter set of values corresponds to those of SP4 thymocytes, as discussed above.

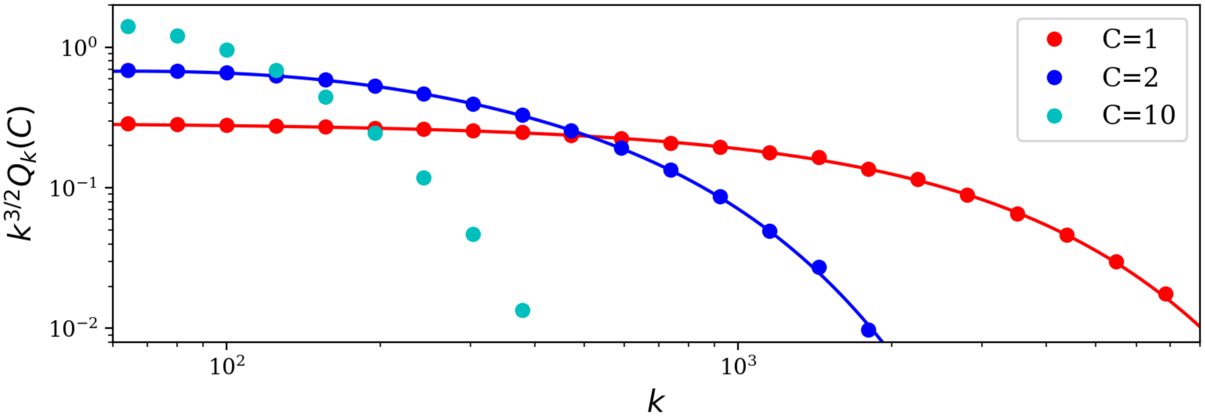

Figure 6: Plot of $k^{3/2}Q_k(C)$ as a function of k, with logarithmic scales, for $C = 1$, $C = 2$, and $C = 10$. The distribution of $\mathbf{R}$ narrows as the number of compartments increases. The solid lines are the exact results, computed using (15) and (65). The dots are averages obtained from Gillespie realisations. Parameter values, chosen using (12) with $N = 25$, are $C = 1$: $p_\mathrm{d} = 0$, $p_\mathrm{b} = 0.4898$; $C = 2$: $p_\mathrm{d}(1) = p_\mathrm{d}(2) = 0$, $p_\mathrm{b}(1) = p_\mathrm{b}(2) = 0.4444$, and $N_1 = N_2 = 5$; $C = 10$: $p_\mathrm{d}(c) = 0$, $p_\mathrm{b}(c) = 0.2158$, and $N_c = 1.38$ for each $c = 1, \ldots, 10$.

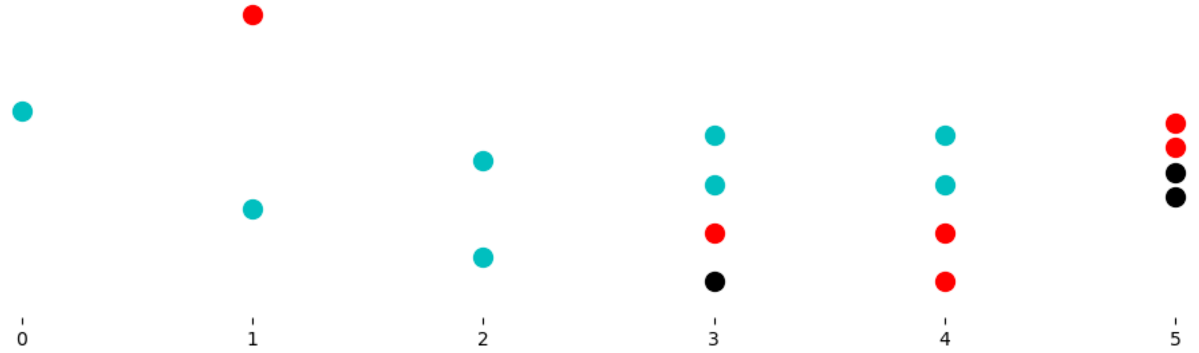

Figure 7: One realisation with $C = 1$, showing generation numbers from left to right, with $\mathbf{Z}_0 = 1$. Cyan cells divide, red cells exit, and black cells die. In this realisation $\mathbf{Y}_0 = 0$, $\mathbf{Y}_1 = 1$, $\mathbf{Y}_2 = 0$, $\mathbf{Y}_3 = 1$, $\mathbf{Y}_4 = 2$, and $\mathbf{Y}_5 = 2$. Thus, we have $\mathbf{R} = 6$. The parameter values are $p_\mathrm{b} = 0.45$ and $p_\mathrm{d} = 0.15$.

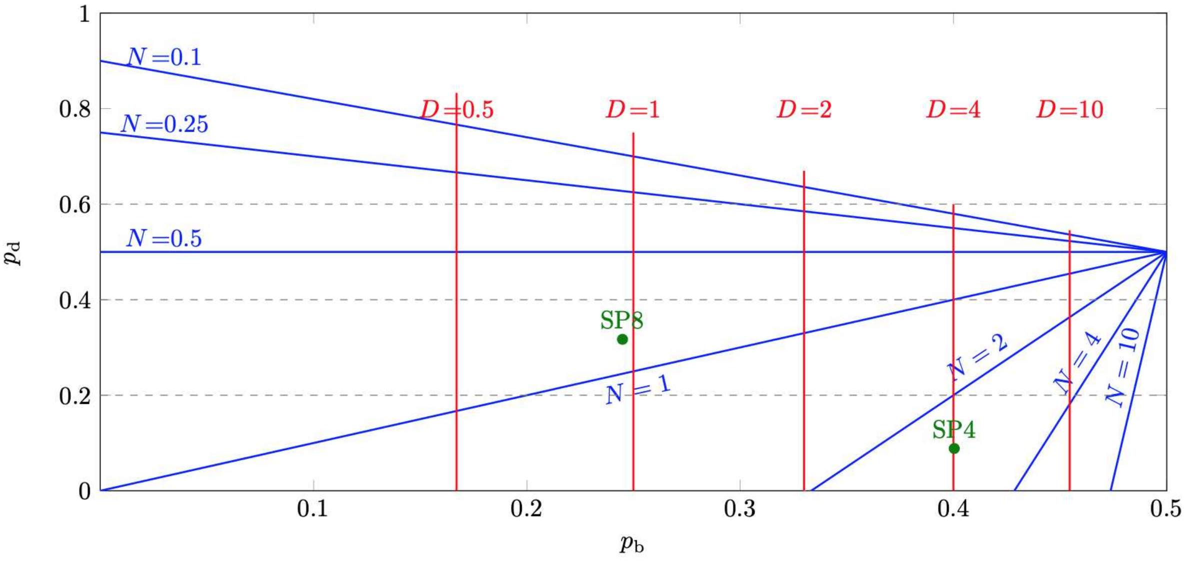

Figure 8: Lines of constant D (red) and lines of constant N (blue) in the part of the plane representing possible parameter values. The two quantities characterising the population of cells exiting a compartment, as functions of p_b and p_d, (6) and (39). Each blue line is the set of pairs $(p_\mathrm{b}, p_\mathrm{d})$ corresponding to the indicated value of N. Each red line is the set of pairs $(p_\mathrm{b}, p_\mathrm{d})$ corresponding to the indicated value of D. The triangular part of the parameter space corresponding to $N > 1$ is at bottom right. The green dots are the estimates of Sawicka *et al.* [14], for SP4 and SP8 thymocytes.

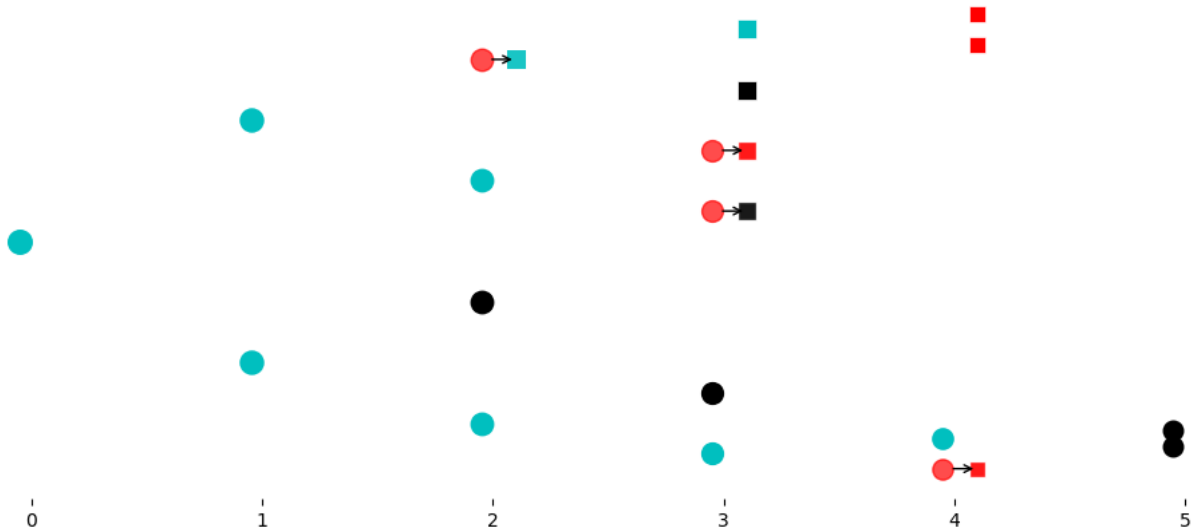

Figure 9: One realisation with $C = 2$, showing generation numbers from left to right. Cells in the first compartment are shown as circles, and cells in the second compartment as squares. Cyan cells divide, red cells exit, and black cells die. Arrows indicate a transition from the first to the second compartment. In this realisation $\mathbf{Y}_0(1) = 0$, $\mathbf{Y}_1(1) = 0$, $\mathbf{Y}_2(1) = 1$, $\mathbf{Y}_3(1) = 2$, $\mathbf{Y}_4(1) = 1$, and $\mathbf{Y}_5(1) = 0$; $\mathbf{Y}_0(2) = 0$, $\mathbf{Y}_1(2) = 0$, $\mathbf{Y}_2(2) = 0$, $\mathbf{Y}_3(2) = 1$, $\mathbf{Y}_4(2) = 3$, and $\mathbf{Y}_5(2) = 0$. Thus, we have $\mathbf{R} = 4$. The parameter values are $C = 2$, $p_{\mathrm{b}}(1) = p_{\mathrm{b}}(2) = 0.45$, and $p_{\mathrm{d}}(1) = p_{\mathrm{d}}(2) = 0.15$.

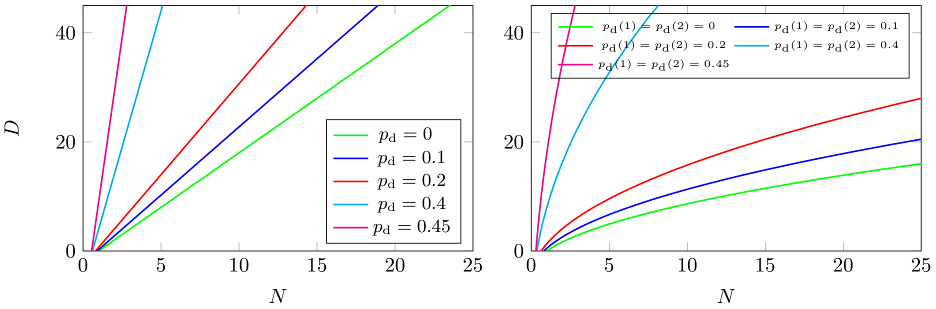

Figure 10: Average generation number of product cells, as a function of the mean number of exiting cells. Left: plot for the case $C = 1$. Right: plot for the case $C = 2$, with parameters chosen so that $N_1 = N_2$. Given a value of N, D is lower when $C = 2$ (proportional to $\sqrt{N}$ as $N \to +\infty$) than when $C = 1$ (proportional to N as $N \to +\infty$).

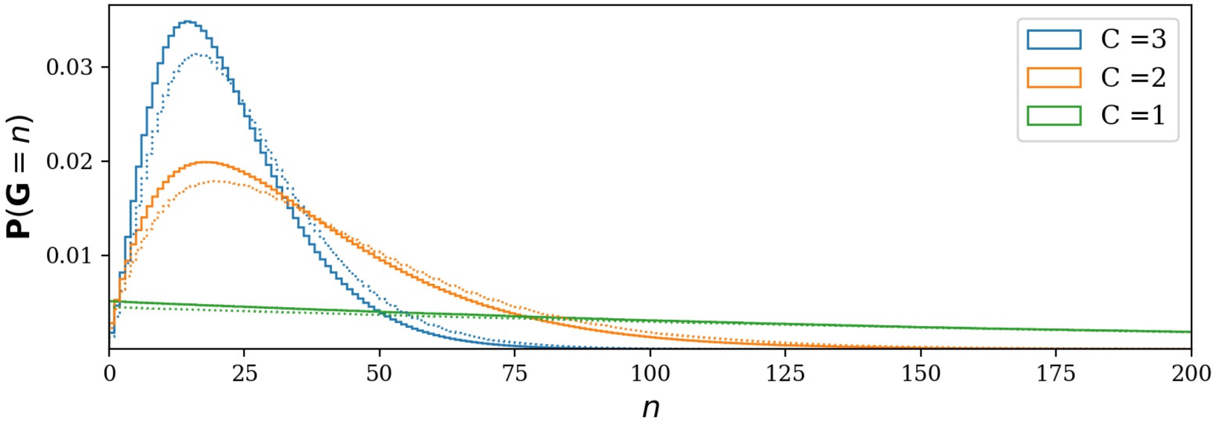

Figure 11: The probability distribution of the random variable **G**, the generation number in the product cell population. One, two and three compartments have been shown. In all cases, $N = 100$, and all compartments are identical. Solid lines correspond to $p_\mathrm{d} = 0$, and dotted lines to $p_\mathrm{d} = 0.05$.

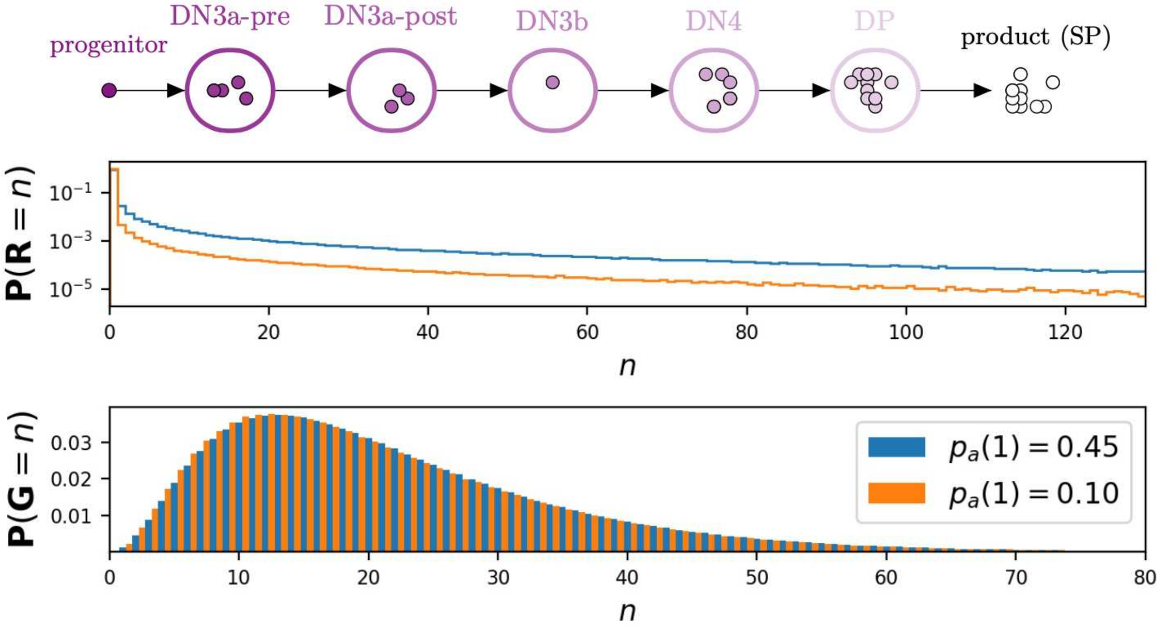

Figure 12: Top: Mathematical model of T cell development from the DN3a to the SP stage [81]. Middle and lower: Numerical results for two cases of the five-compartment thymus model. The histograms show the distributions of family sizes and of cell generation number in the population of product cells. The difference between the two cases is the first compartment, where only death and asymmetric division have non-zero probabilities. Table 1 gives the probabilities for all five compartments, and quantities derived from them.

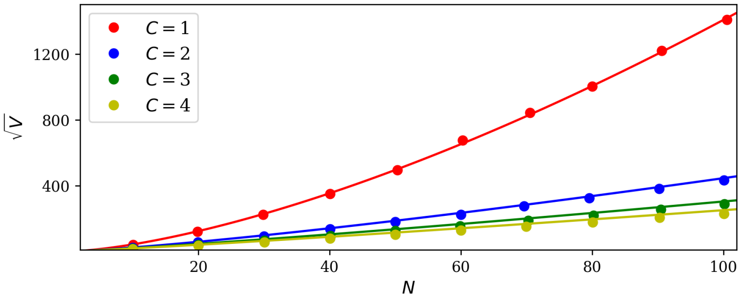

Figure 13: The standard deviation of $\mathbf{R}$ as a function of the mean of $\mathbf{R}$, N, for different values of C. The lines use the formula (66), and each line corresponds to one value of C. The dots are obtained as averages over numerical realisations. Parameter values have been chosen so that N_c is independent of c, $p_d(c) = 0$, and thus, $N_c = N^{1/c}$, $p_e(c) = 1 - p_b(c)$, and $p_b(c) = \frac{N_c - 1}{2N_c - 1}$, for all $c = 1, \ldots, C$.

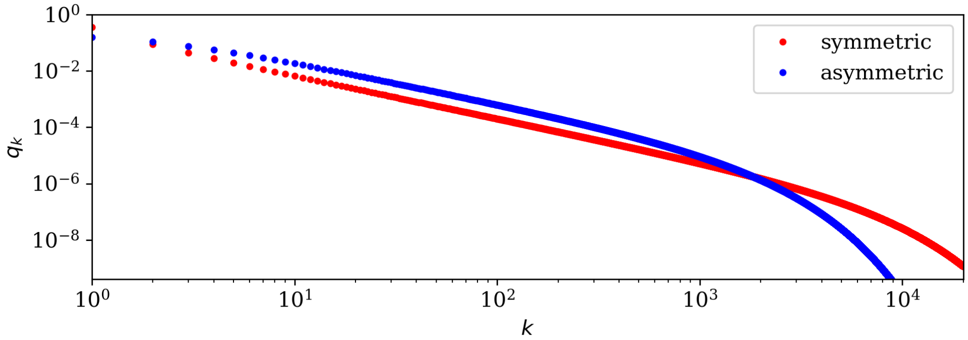

Figure 14: The distribution of $\mathbf{R}$, with and without asymmetric division, when $C = 1$. In red, the symmetric case (11), $p_a = 0$, and in blue, the purely asymmetric case, $p_e = 0$, generated using (74). In both cases we have chosen $N = 25$ and $p_d = 0.25$.

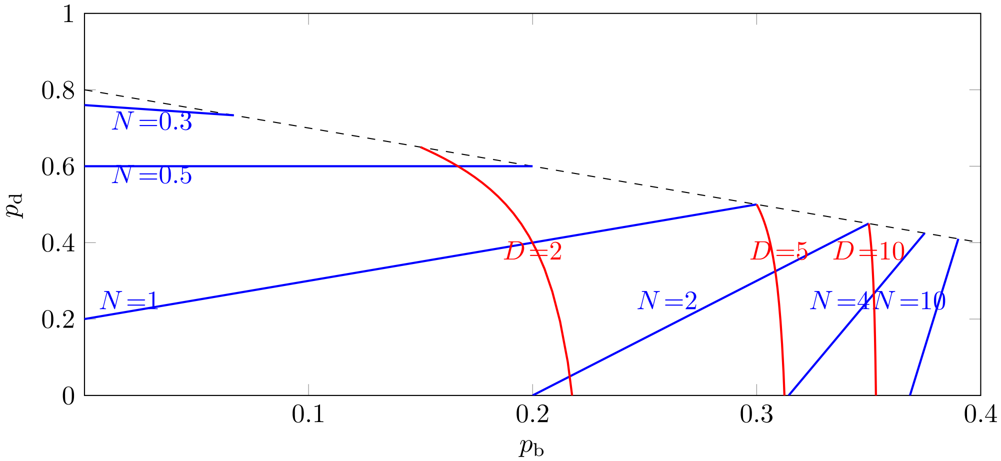

Figure 15: Lines of constant N (blue) and curves of constant D (red) in the part of the plane representing possible parameter values when $p_a = 0.2$. Each blue line is the set of pairs (p_b, p_d) corresponding to the indicated value of N. Each red curve is the set of pairs (p_b, p_d) corresponding to the indicated value of D.