Supplemental Theory for

Fate mapping of human glioblastoma reveals an invariant stem cell hierarchy

Xiaoyang Lan[1,2,3], David J. Jörg[4,5], Florence M. G. Cavalli[1,2], Long V. Nguyen[7], Robert J. Vanner[1,2,3], Laura M. Richards[12,13], Paul Guilhamon[12,13,14], Lilian Lee[1,2], Michelle Kushida[1,2], Davide Pellacani[7,8], Nicole I. Park[1,2,3], Fiona J. Coutinho[1,2,3], Heather Whetstone[1,2], Hayden J. Selvadurai[1,2], Clare Che[1,2], Betty Luu[1,2], Annaick Carles[9], Michelle Moksa[9], Naghmeh Rastegar[1,2], Renee Head[1,2], Sonam Dolma[1,2,11], Panagiotis Prinos[13,20], Michael D. Cusimano[17,18], Sunit Das[17,18], Mark Bernstein[16,18], Cheryl H. Arrowsmith[13,20], Andrew J. Mungall[10], Richard A. Moore[10], Yussane Ma[10], Marco Gallo[19], Mathieu Lupien[12,13,14], Trevor J. Pugh[12,13], Michael D. Taylor[1,2,11,15,18], Martin Hirst[9,10], Connie J. Eaves[7,8], Benjamin D. Simons[4,5,6*], and Peter B. Dirks[1,2,3,15,18*]

[1]Developmental and Stem Cell Biology Program, The Hospital for Sick Children, Toronto, Ontario M5G 0A4, Canada. [2]The Arthur and Sonia Labatt Brain Tumour Research Centre, The Hospital for Sick Children, Toronto, Ontario M5G 0A4, Canada. [3]Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A8, Canada. [4]Cavendish Laboratory, Department of Physics, J. J. Thomson Avenue, Cambridge CB3 0HE, United Kingdom. [5]The Wellcome Trust/Cancer Research UK Gurdon Institute, University of Cambridge, CB2 1QN, United Kingdom. [6]The Wellcome Trust/Medical Research Council Stem Cell Institute, University of Cambridge, United Kingdom. [7]Terry Fox Laboratory, BC Cancer Agency, 675 West 10th Avenue, Vancouver, British Columbia V5Z 1L3, Canada. [8]Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia V6T 2B5, Canada. [9]Centre for High-Throughput Biology, Department of Microbiology and Immunology, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada. [10]Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, 675 West 10th Avenue, Vancouver, British Columbia V5Z 1L3, Canada. [11]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario M5G 0A4, Canada. [12]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario M5G 2M9, Canada. [13]Department of Medical Biophysics, University of Toronto, Toronto, ON M5G 1L7, Canada. [14]Ontario Institute for Cancer Research, Toronto, Ontario M5G 0A3, Canada. [15]Division of Neurosurgery, The Hospital for Sick Children, Toronto, Ontario M5S 3E1, Canada. [16]Division of Neurosurgery, Toronto Western Hospital, Toronto, ON M5T 2S8, Canada. [17]Division of Neurosurgery, St. Michael's Hospital, Toronto, ON M5B 1W8, Canada. [18]Division of Neurosurgery, University of Toronto, Toronto, ON M5S 1A8, Canada. [19]Departments of Physiology and Pharmacology, Biochemistry and Molecular Biology, Alberta Children's Hospital Research Institute, Arnie Charbonneau Cancer Institute, University of Calgary, Calgary, AB T2N 4N1, Canada. [20]Structural Genomics Consortium, University of Toronto, Toronto, ON M5G 1L7, Canada.

*Correspondence: bds10@cam.ac.uk (B.D.S.), peter.dirks@sickkids.ca (P.B.D.).

Contents

In this Supplementary Text, we describe the quantitative analysis of clonal data obtained from serial transplantation experiments of human glioblastoma (GBM) xenografts involving lentiviral barcoding. Our strategy is to analyse the features of barcode frequency distributions to infer the underlying cell fate dynamics giving rise to the heterogeneity of clonal behavior observed in experiments. This heterogeneity could either be (i) a consequence of differential engrained or evolving fitness advantages of cells or (ii) reflecting stochastic fate choices of equipotent progenitor pools (Clayton et al., 2007; Blanpain and Simons, 2013). Here we show that the experimental data is consistent with the latter scenario and that the key features of barcode frequency distributions and correlations can be explained by a simple proliferative hierarchy with glioblastoma stem-like cells at the apex.

In Section 1, we address statistical properties of lentiviral barcoding and give estimates of the amount of uniquely labelled cells. In Section 2, we show that the experimentally obtained barcode frequencies follow a negative binomial distribution. This behavior is characteristic of a specific class of proliferative hierarchies—in Section 3, we show how such a distribution generically arises. Based on these observations, we develop a minimal model of tumor growth in Section 4 and study its predictions on tumor expansion and composition. In Section 5, we use our model to develop a simulation of the serial transplantation experiments which permits a direct comparison of our model with experiments. In Section 6, we infer plausible parameter ranges for our model on biological grounds and compare the model results of our theory with experiments. In Section 7, we use the experimentally obtained data from exome deep sequencing to probe the mutational heterogeneity of the parent tumour and as an independent window on the clonal dynamics of GBM cells.

1 Statistical properties of lentiviral barcoding

Lentiviral barcoding relies on the random infection of cells. While it entails the possibility to uniquely identify clone lineages, the randomness of the barcoding procedure may lead to the same cell acquiring multiple barcodes or to the same barcode being present in more than one cell. Since this can affect the statistical properties of the derived barcode frequency distributions, we here give an estimate for the relative amount of multiply labelled cells and barcodes present in multiple cells.

For a library consisting of N_{B} unique barcodes with a barcoding event occurring with probability p_{B}, the number n_{B} of barcodes acquired by a single cell follows the binomial distribution $Q(n_{\mathrm{B}}) = P_{\mathrm{Binomial}}(n_{\mathrm{B}}|N_{\mathrm{B}}, p_{\mathrm{B}})$, where $P_{\mathrm{Binomial}}(n|N, p) = \binom{N}{n} p^n (1-p)^{N-n}$. For large N_{B}, the distribution Q can be approximated by a Poisson distribution,

$$Q(n_{\mathrm{B}}) \simeq \frac{\nu^{n_{\mathrm{B}}}}{n_{\mathrm{B}}!} \mathrm{e}^{-\nu} , \tag{1}$$

Patient	N_{C} (10^4)	η	Q_0	Q_1	$Q_{>1}$	R_0	R_1	$R_{>1}$	Q
GBM719	12.5	37.8%	62.2%	29.5%	8.3%	74.3%	22.1%	3.6%	69.6%
GBM729	12.5	21.6%	78.4%	19.1%	2.5%	85.9%	13.1%	1.0%	84.4%
GBM735	3	37.6%	62.4%	29.4%	8.2%	93.2%	6.6%	0.2%	91.6%
GBM742	2.4	28.7%	71.3%	24.1%	4.6%	96.0%	3.9%	0.1%	95.3%
GBM743	8	17.3%	82.7%	15.7%	1.6%	92.7%	7.0%	0.3%	92.0%
GBM754	12.5	33.2%	66.8%	27.0%	6.2%	77.7%	19.6%	2.7%	74.1%

Table S1 Probabilities characterising the statistical properties of lentiviral barcoding with a library of $N_{\mathrm{B}} = 2 \times 10^5$ barcodes (L. V. Nguyen, M. Makarem, *et al.*, 2014).

where $\nu = p_{\mathrm{B}} N_{\mathrm{B}}$. Using Eq. (1), the relative amount of unlabelled cells, $Q_0 = Q(0)$, the relative amount of cells labelled with one barcode, $Q_1 = Q(1)$, and the relative amount of cells carrying more than one barcode, $Q_{>1} = \sum_{n_{\mathrm{B}}>1} Q(n_{\mathrm{B}})$, are obtained as

$$Q_0 = \mathrm{e}^{-\nu} , \qquad Q_1 = \nu \mathrm{e}^{-\nu} , \qquad Q_{>1} = 1 - (1+\nu)\mathrm{e}^{-\nu} . \tag{2}$$

The parameter ν characterizing the distribution of barcodes can be obtained from the labelling efficiency η, which denotes the relative amount of cells that bear at least one barcode, by requiring $1 - Q_0 = \eta$. This yields

$$\nu = -\ln(1-\eta) . \tag{3}$$

Conversely, we can ask the question how likely it is that the same barcode appears in multiple cells. Out of a total of N_{C} cells prepared for barcoding, the number n_{C} of cells acquiring the same barcode is distributed according to $R(n_{\mathrm{C}}) = P_{\mathrm{Binomial}}(n_{\mathrm{C}}|N_{\mathrm{C}}, p_{\mathrm{B}})$. Again, for a large number of cells N_{C}, this can be approximated by a Poisson distribution,

$$R(n_{\mathrm{C}}) \simeq \frac{\kappa^{n_{\mathrm{C}}}}{n_{\mathrm{C}}!} \mathrm{e}^{-\kappa} , \tag{4}$$

where $\kappa = p_{\mathrm{B}} N_{\mathrm{C}} = \nu N_{\mathrm{C}}/N_{\mathrm{B}}$. Analogously to Eqs. (2), we obtain the relative amount of barcodes that are present in no cell, $R_0 = R(0) = \mathrm{e}^{-\kappa}$, the relative amount of barcodes present in exactly one cell, $R_1 = R(1) = \kappa \mathrm{e}^{-\kappa}$ and the relative amount of barcodes that have been acquired by more than one cell, $R_{>1} = \sum_{n_{\mathrm{C}}>1} R(n_{\mathrm{C}}) = 1 - (1+\kappa)\mathrm{e}^{-\kappa}$.

Multiple barcoding of the same cell is unproblematic for the quantitative analysis of barcode frequency distributions—it generates copies of clones which are however subject to the same distribution of barcode frequencies. On the other hand, barcodes distributed to multiple cells lead to an effective merging of the sizes of derived clones and thus may alter the statistical properties of the barcode frequency distribution. Among the labelled cells, the relative amount of uniquely labelled cells, i.e., cells with a unique combination of one or more barcodes, is given by

$$\mathcal{Q} = \frac{1}{1 - Q_0} \sum_{n_{\mathrm{B}}=1}^{\infty} Q(n_{\mathrm{B}}) \left[(1 - p_{\mathrm{B}})^{N_{\mathrm{C}}-1} \right]^{n_{\mathrm{B}}} , \tag{5}$$

which, for $N_{\mathrm{C}} \gg 1$ and $p_{\mathrm{B}} \ll 1$, can be approximated in terms of the probabilities Q_0 and R_0 as

$$\mathcal{Q} \simeq \frac{1 - Q_0^{-R_0}}{1 - Q_0^{-1}} . \tag{6}$$

Table S1 summarizes the respective probabilities for all xenografts used in this study; a large majority of labelled cells carries a unique combination of barcodes in all xenografts.

2 Barcode frequencies follow a negative binomial distribution

To obtain a quantitative understanding of tumor growth, we analyze the distribution of barcode frequencies obtained from serial transplantation experiments. Here, we show that the distributions $p(n)$ of barcode frequencies above the detection threshold for all passages and replicate experiments follow a negative binomial distribution,

$$p(n) = \frac{1}{\mathcal{N}_0} \frac{\mathrm{e}^{-n/n_0}}{n} , \tag{7}$$

where n_0 is a characteristic barcode frequency of the respective population and $\mathcal{N}_0$ is a normalisation constant. A robust method to detect negative binomial distributions is to obtain the first incomplete moment of the distribution p, defined by

$$\mu(n) = \frac{1}{\langle n \rangle} \sum_{n'=n}^{\infty} n' p(n') , \tag{8}$$

where $\langle n \rangle = \sum_n n \ p(n)$ is the average barcode frequency. By definition, $\mu(n)$ is the relative average barcode frequency of all barcode frequencies larger than n. If the barcode frequency distribution $p(n)$ has the negative binomial form Eq. (7), the first incomplete moment acquires an exponential dependence on the barcode frequency,

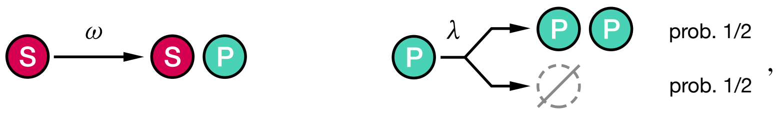
$$\mu(n) = \frac{1}{\mathcal{N}_1} \mathrm{e}^{-n/n_0} , \tag{9}$$

where $\mathcal{N}_1$ is another normalization constant. Since the first incomplete moment, together with the average barcode frequency $\langle n \rangle$, carries the same information as the original barcode frequency distribution[1], an exponential dependence of μ is completely equivalent to a negative binomial barcode frequency distribution.

Fig. 2b in the main text and Extended Data Figs. 5-6 show all first incomplete moments of the experimental barcode frequency distributions. They clearly exhibit an exponential behavior (linear on a logarithmic scale) over many decades of barcode frequencies, indicating negative binomial distributions across different patients, xenografts, passages, and replicate experiments. By definition of the first incomplete moment, data points with large barcode frequencies outside the negative binomial distribution show up as a strong deviation from the exponential behavior (see, e.g., red arrowhead in Fig. 2b). This is caused by the barcode frequency entering as a multiplicative term in the definition Eq. (8). Importantly, this does not affect its ability to detect negative binomial distributions for small barcode frequencies.

3 Emergence of negative binomial distributions

What can the barcode frequency distribution tell us about the proliferative dynamics underlying tumor growth? A generic mechanism giving rise to a negative binomial distribution is a process long-known in population dynamics, termed 'critical birth-death process with immigration' (Bailey, 1990; Simons, 2016). Translated into the language of cell population dynamics, such a process can be realized by a population of cells that stochastically divide ('birth') and differentiate ('death') with equal probability ('critical'), with a slow influx of cells from another cell compartment ('immigration') through differentiation. In the tumor context, such a process could naturally arise if there is (i) a slowly cycling glioblastoma stem cell (GSC) compartment at the apex of a proliferative hierarchy that sporadically gives rise to progenitor cells by asymmetric division and (ii) the resulting progenitor population undergoes division and differentiation that are balanced on the population level. Schematically, the dynamics of stem cells (S) and progenitors (P) can be expressed as



(10)

where λ is the loss-and-replacement rate of the progenitors and ω is the asymmetric division rate of the stem cells, also called 'immigration rate' since it describes the rate at which cells enter the progenitor compartment. If the immigration rate

[1] The barcode frequency distribution can be retrieved from the first incomplete moment via the relation $p(n) = \langle n \rangle [\mu(n) - \mu(n-1)]/n$.

ω is small compared to the loss-and-replacement rate λ, uniquely barcoded stem-like cells produce clones with a negative binomial barcode frequency distribution, Eq. (7). To show this, we describe the cell fate dynamics shown in scheme (10) as independent Poisson processes with rates ω and λ, respectively. Formally, the corresponding master equation that governs the dynamics of the probability $P = P(n,t)$ to find n progenitor cells in a clone derived from a single uniquely labelled S-type cell,

$$\frac{\partial P}{\partial t} = \left\{ \omega(\hat{E}^{-} - 1) + \frac{\lambda}{2}(n-1)\hat{E}^{-} + \frac{\lambda}{2}(n+1)\hat{E}^{+} - \lambda n \right\} P \, , \tag{11}$$

where we have introduced the ladder operators $\hat{E}^{\pm}$, defined by $\hat{E}^{\pm}P(n,t) = P(n \pm 1, t)$. The first term in brackets describes the asymmetric division of a single S-type cell whereas the remaining three terms describe symmetric division and death of P-type cells[2]. Note that asymmetric division of S-type cells leaves the number of S-type cells unchanged so that it is sufficient to only describe the number n of P-type cells.

The master equation (11) describes the dynamics of S-type and P-type cells shown in the scheme (10) as independent Poisson processes. An analytical solution can readily be obtained by standard methods (Walczak et al., 2012). For initially no progenitor cells being labelled, $P(n,0) = \delta_{n,0}$, the exact solution to the master equation (11) is given by the negative binomial distribution

$$P(n,t) = \frac{1}{n!}\frac{\Gamma(\zeta+n)}{\Gamma(\zeta)}\left(\frac{n_0(t)}{1+n_0(t)}\right)^{n}\left(1-\frac{n_0(t)}{1+n_0(t)}\right)^{\zeta} \, , \tag{12}$$

where $n_0(t) = \lambda t/2$, the dimensionless parameter $\zeta = 2\omega/\lambda$ is the ratio of immigration rate and progenitor loss-and-replacement rate, and $\Gamma(x) = \int_0^\infty u^{x-1}\mathrm{e}^{-u}\,\mathrm{d}u$ is the Gamma function. On average, barcode frequencies grow linearly in time,

$$\langle n(t) \rangle = 1 + \omega t \, , \tag{13}$$

[2] The structure of the master equation (11) can be understood by considering, for instance, a reduced dynamics that only describes asymmetric divisions of the S-type cell. This amounts to setting $\lambda = 0$ in Eq. (11) which yields the reduced equation $\partial P/\partial t = \omega P(n-1,t) - \omega P(n,t)$, where we have used the definition of the ladder operator $\hat{E}^{-}$. This equation describes the rate of change of the probability $P(n,t)$ to find n P-type cells. A state with n P-type cells can only be reached if there are already $n-1$ P-type cells and an asymmetric division of an S-type cell occurs, giving rise to another P-type cell. The corresponding contribution $\omega P(n-1,t)$ to the rate of change $\partial P/\partial t$ is given by the probability $P(n-1,t)$ to find the system in the state $n-1$ multiplied by the rate ω of asymmetric divisions. Conversely, the state with n P-type cells is left if another asymmetric division of the S-type cell occurs, raising the number of P-type cells to $n+1$. The analogous contribution $-\omega P(n,t)$ enters with a negative sign as it describes the process of leaving the state n. The other terms in the full master equation (11) follow the same logic. For more detailed reviews on general master equations and birth-death processes, we refer the reader to standard textbooks (Gardiner, 2009; Bailey, 1990)

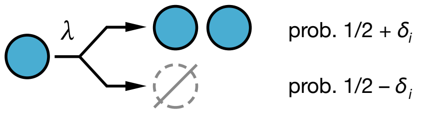
where $\langle n(t)\rangle = 1+\sum_n nP(n,t)$ is the average barcode frequency with the first term accounting for the stem cell. For small immigration rates ω, the barcode frequency distribution of cell populations with at least one progenitor cell, given by $P_*(n,t) = P(n,t)/(1-P(0,t))$, acquires the form Eq. (7),

$$\begin{aligned} P_*(n,t) &= \frac{1}{\ln(1+n_0)}\frac{1}{n}\left(\frac{n_0(t)}{1+n_0(t)}\right)^n + \mathcal{O}(\omega) \\ &\approx \frac{1}{\mathcal{N}_0(t)}\frac{\mathrm{e}^{-n/n_0(t)}}{n}\,, \end{aligned} \tag{14}$$

where $\mathcal{N}_0(t) = \ln n_0(t)$. For non-vanishing but small immigration rates ω, the resulting barcode frequency distribution is still well-approximated by Eq. (14). Thus, the dynamics (10) generically give rise to negative binomial barcode frequency distributions and hence are the starting point for our quantitative analysis.

Comparison with clone size distributions emerging from engrained proliferative heterogeneity

Could a negative binomial barcode frequency distribution also be caused by engrained proliferative heterogeneity instead of equipotency? To address this question, let us consider a large population of clones in which the cell of each clone i undergoes loss and replacement with clone-specific probabilities. For concreteness, we consider the following cell fate dynamics in which each cell undergoes loss and replacement with different probabilities,



. (15)

The parameter δ_i determines whether cell i is primed for proliferation ($\delta_i > 0$) or loss ($\delta_i < 0$). The average size of a clone derived from cell i evolves according to $\langle n_i(t)\rangle = \mathrm{e}^{2\delta_i\lambda t}$ and on average, clones will thus either grow exponentially ($\delta_i > 0$) or die out ($\delta_i < 0$). In this picture, engrained proliferative heterogeneity is captured by a broad distribution of the δ_i, so that some clones expand faster than others while some clones die. For a clone with a given δ_i, the resulting surviving clone size distribution at large times is exponential (Bailey, 1990),

$$p(n|\alpha_i) \simeq \alpha_i \mathrm{e}^{-\alpha_i n}\,. \tag{16}$$

with an exponent α_i that depends on the proliferative potential and on time. Hence, a distribution in engrained proliferative advantages δ_i entails a distribution in the shape parameter α_i of the clone size probabilities for the different clones. As an example, let us consider the distribution of α_i at a fixed time $t = t_0$. For simplicity, we here consider a Gamma distribution[3] for α, which ensures that $\alpha > 0$,

$$\bar{p}(\alpha) = \frac{\alpha^{m-1}\mathrm{e}^{-m\alpha/\alpha_0}}{(\alpha_0/m)^m\Gamma(m)} \,. \tag{17}$$

The clone size distribution resulting from this distribution of clone size scales is given by

$$p(n) = \int_0^\infty p(n|\alpha)\bar{p}(\alpha)\,\mathrm{d}\alpha = \frac{\alpha_0}{(\alpha_0 n/m + 1)^{m+1}} \,, \tag{18}$$

which asymptotically has the power law behavior $n^{-(m+1)}$ and is therefore distinctly different from the negative binomial form $\mathrm{e}^{-n/n_0}/n$. Which distribution of proliferative potentials would be needed to generate a negative binomial clone size distribution under these circumstances? In fact, a negative binomial form can only be obtained under very artificial conditions: the distribution for α would have to take the non-normalizable discontinuous form $\bar{p}(\alpha) \propto \alpha^{-1}\Theta(\alpha - \alpha_0)$ where Θ is the Heaviside step function; in this case, the clone size distribution would sensitively depend on the position α_0 of the step as it determines the characteristic scale of clone sizes, $p(n) = \mathrm{e}^{-\alpha_0 n}/n$. While being simplistic, this minimal model of engrained proliferative heterogeneity illustrates that negative binomial clone size distributions do not generically arise from a mere loss and replacement of clones—rather, the cell fate dynamics has to display certain distinctive features, such as the minimal hierarchy of the type (10), which robustly leads to such clone size distributions.

4 Theoretical model of tumor growth

In Section 3, we have shown how a negative binomical barcode frequency distribution can arise from a single uniquely labelled stem cell at the apex of a critical birth-death process with immigration. However, there are several reasons why growth of glioblastoma as observed in serial transplantation experiments warrant a more comprehensive model: First, the model (10) only considers strictly asymetrically dividing stem cells, leading to linear growth of barcode frequencies on average. However, there is no reason to a priori rule out *symmetric* stem cell divisions, which potentially provide a considerable contribution to tumor growth. Second, in the model (10), loss of the stem cell leads to a remaining progenitor cell population that will not grow on average and will eventually die out (Clayton et al., 2007). In the serial transplantation experiments, only small fractions of a harvested tumor

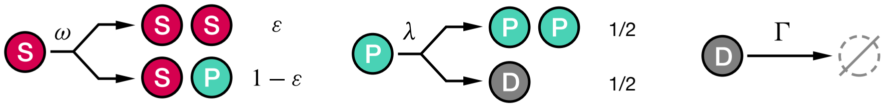
[3] The Gamma distribution as defined in Eq. (17) has mean α_0 and variance α_0^2/m; in the limit of large m, it is approximately equal to a normal distribution with the same mean and variance. For $m = 1$, the Gamma distribution reduces to an exponential distribution.

($\sim 5\%$) are chosen for reinjection. If clones were indeed maintained by a single stem cell, it would thus be likely that the stem cell is lost upon reinjection, giving rise to a massive loss of barcodes across passages which is not observed in experiments. Third, the model (10) neglects the potential presence of a non-proliferating compartment undergoing apoptosis that may affect the tumor size and composition. This non-proliferating compartment may be the differentiating progeny of the progenitor population or a quiescent progenitor population.

Therefore, in this section, we now formulate a more comprehensive model of glioblastoma growth and study its predictions on tumor growth and composition. Our model makes falsifiable predictions and to compare it with experiments, we introduce a simulation procedure that combines the clonal dynamics with harvesting and reinjection scheme to mimic the experimental procedure (Section 5). Subsequently, we compare our model to experimental data and show that it captures the key features of clonal dynamics (Section 6).

4.1 Stochastic dynamics of cell division and differentiation

Our model of tumor growth describes the dynamics of three cell compartments: a stem-like cell compartment (S), a progenitor compartment (P), and a non-proliferating compartment that may account for differentiating progeny (D). In our model, stem-like cells divide symmetrically with a probability ε and asymmetrically with probability $1-\varepsilon$. Progenitor cells either divide symmetrically or differentiate into their progeny, both with probability $1/2$, so that division and differentiation are balanced on the population level. The differentiating compartment has a finite lifetime and constitutes the lowest level of the differentiation hierarchy in our model. Schematically, the model can be expressed as



, (19)

where ω and λ are the division rates of stem cells and progenitors, respectively, and Γ is the apoptosis rate of the differentiating progeny. Defining $P(n^{\mathrm{S}}, n^{\mathrm{P}}, n^{\mathrm{D}}, t)$ as the probability to find n^{S} stem cells, n^{P} progenitor cells, and n^{D} differentiated cells at time t within a clone, we write down a master equation governing the stochastic dynamics in the same spirit as in the previous section,

$$\begin{aligned}\frac{\partial P}{\partial t} = \Big\{ & \varepsilon\omega\left(n^{\mathrm{S}}-1\right)\hat{E}_{\mathrm{S}}^{-} + (1-\varepsilon)\omega n^{\mathrm{S}}\hat{E}_{\mathrm{P}}^{-} - \omega n^{\mathrm{S}} + \frac{\lambda}{2}\left(n^{\mathrm{P}}-1\right)\hat{E}_{\mathrm{P}}^{-} \\ & + \frac{\lambda}{2}n^{\mathrm{P}}\hat{E}_{\mathrm{P}}^{+}\hat{E}_{\mathrm{D}}^{-} - \lambda n^{\mathrm{P}} + \Gamma\left(n^{\mathrm{D}}+1\right)\hat{E}_{\mathrm{D}}^{+} - \Gamma n^{\mathrm{D}}\Big\} P \ ,\end{aligned} \tag{20}$$

where we have again used ladder operators defined by $\hat{E}_{\mathrm{S}}^{\pm} P(n^{\mathrm{S}}, n^{\mathrm{P}}, n^{\mathrm{D}}, t) = P(n^{\mathrm{S}} \pm 1, n^{\mathrm{P}}, n^{\mathrm{D}}, t)$ and analogously for the other cell compartments P and D. Together with an initial condition $P(n^{\mathrm{S}}, n^{\mathrm{P}}, n^{\mathrm{D}}, 0)$ that characterizes the initially barcoded population, Eq. (20) permits to compute the clone composition and barcode frequency distribution of our model at any later time. The distribution $p(n,t)$ of total barcode frequencies $n = n^{\mathrm{S}} + n^{\mathrm{P}} + n^{\mathrm{D}}$ is obtained from the joint distribution P by summing over all barcode frequency configurations that lead to a total size n,

$$p(n,t) = \sum_{n'=0}^{n} \sum_{n''=0}^{n-n'} P(n', n'', n - n' - n'', t) \; . \tag{21}$$

While the full clonal dynamics of our model can only be explored by means of numerical simulations, several important insights about growth and composition of the tumor can be drawn from analytical arguments.

4.2 Composition of the tumor

Using the master equation (20), we can obtain insights into the composition of the tumor in our model, i.e., its relative content of stem cells, progenitors and differentiating progeny. The time evolution of the mean cell numbers is given by

$$\begin{aligned}
\langle \dot{n}^{\mathrm{S}} \rangle &= \varepsilon\omega \langle n^{\mathrm{S}} \rangle \; , \\
\langle \dot{n}^{\mathrm{P}} \rangle &= (1-\varepsilon)\omega \langle n^{\mathrm{S}} \rangle \; , \\
\langle \dot{n}^{\mathrm{D}} \rangle &= \tfrac{1}{2}\lambda \langle n^{\mathrm{P}} \rangle - \Gamma \langle n^{\mathrm{D}} \rangle \; ,
\end{aligned} \tag{22}$$

where the dot denotes the time derivative. In particular, the evolving clone, while steadily growing, acquires a steady-state composition characterized by a constant relative amount of stem-like cells, progenitor cells, and differentiated cells: Defining the relative cell contents $\phi^{\mathrm{S}} = \langle n^{\mathrm{S}} \rangle / \langle n \rangle$, $\phi^{\mathrm{P}} = \langle n^{\mathrm{P}} \rangle / \langle n \rangle$, and $\phi^{\mathrm{D}} = \langle n^{\mathrm{D}} \rangle / \langle n \rangle$ where $\langle n \rangle = \langle n^{\mathrm{S}} \rangle + \langle n^{\mathrm{P}} \rangle + \langle n^{\mathrm{D}} \rangle$ is the total barcode frequency, this stationary composition satisfies $\dot{\phi}^{\mathrm{S}} = \dot{\phi}^{\mathrm{P}} = \dot{\phi}^{\mathrm{D}} = 0$ and is given by

$$\phi^{\mathrm{S}} = \varepsilon \Omega^{-1} \; , \qquad \phi^{\mathrm{P}} = (1-\varepsilon)\Omega^{-1} \; , \qquad \phi^{\mathrm{D}} = 1 - \Omega^{-1} \; , \tag{23a,b,c}$$

with Ω being a dimensionless parameter given by

$$\Omega = 1 + \frac{\lambda}{2} \frac{1-\varepsilon}{\Gamma + \varepsilon\omega} \; . \tag{24}$$

Eqs. (23a–c) show that the probability ε for symmetric stem cell division determines the relative fraction of stem-like and progenitor cells while the composite parameter Ω determines the relative fraction of the differentiating progeny and the remaining two compartments. Note that in general the ratio of averages does not correspond to the average of the ratio, $\langle n^X \rangle / \langle n \rangle \neq \langle n^X / n \rangle$ for $X = \mathrm{S}, \mathrm{P}, \mathrm{D}$. However, simulations show that Eqs. (23a–c) are excellent approximations for the averages $\langle n^X / n \rangle$ in the considered parameter ranges.

4.3 Tumor expansion

The average growth of a clone (and thus the tumor) can be determined from Eqs. (22) as well. Defining the fold-change in cell number compared to the initial barcode frequency, $\gamma(t) = \langle n(t)\rangle / \langle n(0)\rangle$, we obtain

$$\gamma(t) = \mathrm{e}^{\varepsilon\omega t} \ , \tag{25}$$

given that, from the outset, the tumor has the stationary composition given by Eqs. (23). Hence, the tumor expands exponentially with the growth speed given by the rate $\varepsilon\omega$ of symmetric stem cell divisions.

5 Simulation of transplantation experiments

To capture the dynamics of the serial transplantation experiments, we develop a simulation of the clonal dynamics involving the repeated procedure of injection, unperturbed growth, and harvesting of the tumor. To this end, we use a stochastic simulation algorithm to compute many realizations of the clonal dynamics (Gillespie, 1977). The simulation consists of (i) the injection of a single uniquely labelled cell, (ii) unperturbed clonal dynamics according to the process (19), and (iii) subsequent harvesting of cells for sequencing and reinjection. Key observables such as barcode frequency distributions, numbers of surviving barcodes, and clonal growth are then obtained by performing statistics over the computed realizations.

5.1 Primary injection

To mimic the experimental procedure in our simulation, we start the primary passage by injecting a single labelled S or P cell, each with a probability that reflects the steady-state fractions given in Eqs. (23a,b). Differentiating progeny (represented by the D compartment in our model) are unlikely to survive the process of serial transplantation. The corresponding initial condition for the probability P is thus given by

$$P(n^{\mathrm{S}}, n^{\mathrm{P}}, n^{\mathrm{D}}, 0) = \frac{\phi^{\mathrm{S}}\delta_{n^{\mathrm{S}},1}\delta_{n^{\mathrm{P}},0}\delta_{n^{\mathrm{D}},0} + \phi^{\mathrm{P}}\delta_{n^{\mathrm{S}},0}\delta_{n^{\mathrm{P}},1}\delta_{n^{\mathrm{D}},0}}{\phi^{\mathrm{S}} + \phi^{\mathrm{P}}} \ . \tag{26}$$

5.2 Tumor growth

After the injection, the clone is subject to unperturbed growth according to Eq. (20) for the duration τ_i of the corresponding passage i.

5.3 Harvesting and reinjection

After each passage, the next passage i is initiated by reinjecting cells harvested from the previous passage $i - 1$. This amounts to setting a new initial condition for the probability P at the injection time t_i^{inj}, which coincides with the harvesting time $t_{i-1}^{\mathrm{harv}} = \sum_{j=1}^{i-1} \tau_j$ of the previous passage, where τ_i is the passage duration of passage i. Again, assuming that it is unlikely for differentiating cells (D) to survive the process of serial transplantation, only stem-like cells (S) and progenitors (P) are reinjected, each such a cell with a probability p_i^{inj}. The probability p_i^{inj} is determined by requiring that on average, the number n_i^{inj} of injected cells matches the number in the corresponding experiment. The probability p_i^{inj} can be calculated as follows. From Eqs. (22), the average growth of a clone can be calculated for any initial composition of the clone. If only S and P cells are injected, with cell numbers that reflect the stationary composition given by Eqs. (23a,b), the fold change $\gamma(t) = \langle n(t)\rangle / \langle n(0)\rangle$ in cell number is given by

$$\gamma(t) = \Omega \mathrm{e}^{\varepsilon\omega t} - (\Omega - 1)\mathrm{e}^{-\Gamma t} \ . \tag{27}$$

with Ω defined in Eq. (24). Hence, the total tumor size after passage i is given by $n^{\mathrm{inj}}\gamma(\tau)$ where n^{inj} is the number of injected cells and τ is the passage duration. Since the composition of the tumor quickly acquires the stationary composition given by Eqs. (23a,b,c) during the passage, the total number of S and P cells upon harvesting is given by $(\phi^{\mathrm{S}} + \phi^{\mathrm{P}})n^{\mathrm{inj}}\gamma(\tau)$. Therefore, to inject an average of n_i^{inj} cells at the beginning of passage i, the probability p_i^{inj} must be chosen as

$$p_i^{\mathrm{inj}} = \frac{n_i^{\mathrm{inj}}}{(\phi^{\mathrm{S}} + \phi^{\mathrm{P}})n_{i-1}^{\mathrm{inj}}\gamma(\tau_{i-1})} \ . \tag{28}$$

Then, the system again evolves according to Eq. (20) till t_{i+1}^{harv} and the same procedure is repeated for the next passage.

5.4 Example

Fig. S1 and Fig. 2d in the main text show numerical examples of the simulation. The upper panel displays different trajectories of barcode frequencies across three passages. Because of stochastic cell fate decisions, clones stochastically grow or shrink during a passage. Therefore, individual trajectories may emerge above and drop below a detection threshold (shaded area in Fig. S1) several times over the course of time (see yellow trajectory in Fig. S1 for an example). While the majority of clones is lost, a few clones grow very large by chance, acquiring several hundreds of cells. After each passage, all barcode frequencies abruptly drop due to harvesting and reinjection of a small sample of the tumor ($\sim$ 5%). From many realizations of the system, statistical properties of the clones such as barcode frequency distributions and correlations can be obtained: the lower panel of Fig. S1 shows, e.g., the average barcode frequency. Note that the average barcode frequency is strongly affected by the majority of clones becoming extinct very quickly while only a few

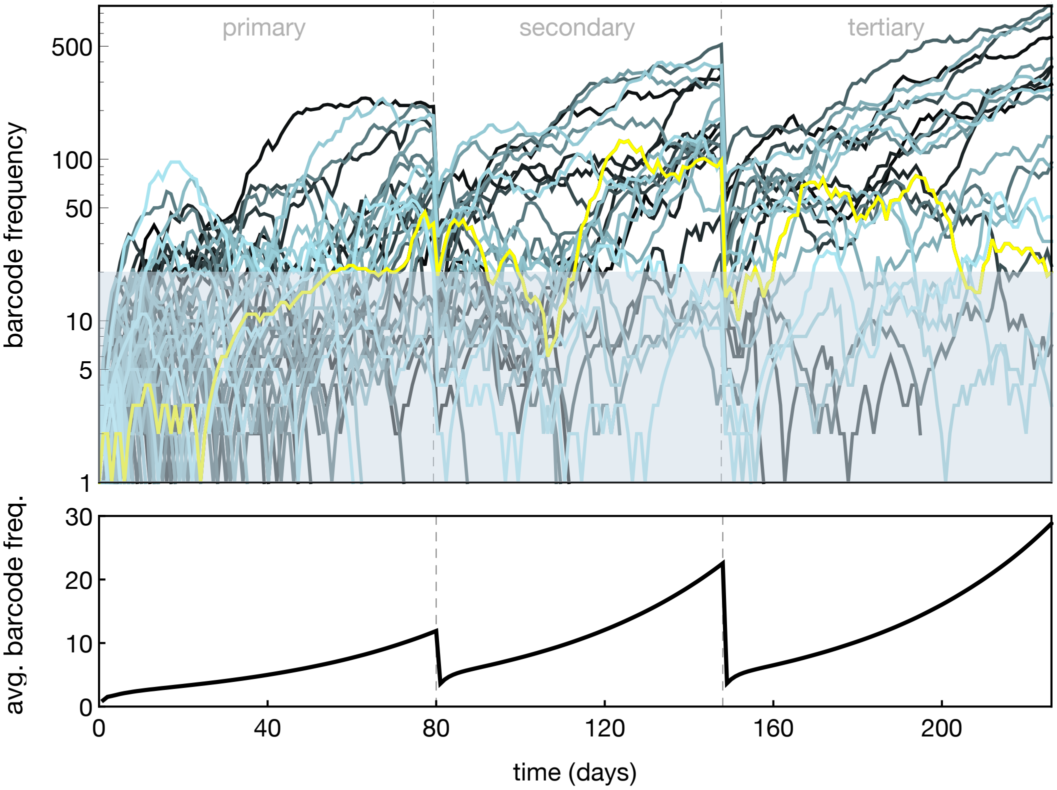


Figure S1 Numerical examples of barcode frequency trajectories across three serial passages on a logarithmic scale. The shaded area indicates an example detection threshold. The yellow curve shows a clone that emerges above and drops below the detection threshold several times. The lower plot shows the average over all trajectories. Parameters are given in Table S3.

clones become large. Fig. 2e in the main text shows the first incomplete moment of the barcode frequency distribution, revealing a negative binomial distribution over many decades as discussed in Section 2. We now use these simulations to systematically compare experimental data with our theory.

6 Comparison of theory and experiments

We now compare our theory with experiments. First, we discuss biologically sensible parameter ranges for our model. We then compare barcode frequency distributions and number of barcodes that survive the serial transplantations with experiments, highlighting that many qualitative key features of our theory are actually independent of the specific choice of parameters.

6.1 Parameter estimates

Can all experiments be characterized by the same set of parameters? Experimental data show a considerable degree of variation in the growth of different xenografts: for instance, referring to Table S4, the tertiary xenograft of the transplantation series labelled $(1, 2, 1)_{719}$ grows by 42-fold over a duration of 55 days, while the tertiary xenograft of $(1, 2, 3)_{719}$ grows by only 26-fold over the longer duration of 78 days, with both xenografts having been derived from the same secondary xenograft $(1, 2)_{719}$. While there are many potential sources for these variations among replicate experiments, this example already indicates that it is not possible to characterize all experiments with a single set of parameters. Rather, it suggests a corresponding degree of variation for the proliferation and differentiation rates of stem-like cells and progenitors as well as the apoptosis rate of the differentiating progeny. Here we aim at constraining plausible parameter ranges using experimental data.

An estimate for the apoptosis rate Γ of the differentiating progeny can be inferred from the steady-state composition of the tumor: we used Ki67 staining of xenograft samples to determine the relative amount of proliferating cells as 50% on average, see Extended Data Fig. 3d. Based on this estimate we fixed the relative amount of progenitor cells among the progenitor population, $\psi = n^{\mathrm{P}}/(n^{\mathrm{D}} + n^{\mathrm{P}})$, as $\psi \approx 0.5$. Using Eqs. (23) and (24), the apoptosis rate Γ can be expressed in terms of ψ and the other parameters as

$$\Gamma = \frac{\lambda}{2} \frac{1}{\psi^{-1} - 1} - \varepsilon\omega \; . \tag{29}$$

Hence, given numerical values for the other parameters ω, ε, and λ, this fixes the value of Γ. For the loss-and-replacement rate λ of the progenitors, we choose an upper bound of $\lambda = 1.5/\text{day}$, motivated by the fact that in mammalian cells, the typical S phase duration is already 5 to 6 hours which constrains the cell cycle speed. In Section 3, we have seen that progenitors have to divide much faster than stem-like cells ($\lambda \gg \omega$) in order to generate the characteristic negative binomial form of the barcode frequency distribution. Therefore, we restrict the stem cell division rate ω to values of at least an order of magnitude less, $\omega \lesssim 0.3/\text{day}$. In our model, overall growth of the tumor crucially depends on the rate $\varepsilon\omega$ of symmetric stem cell divisions (see Sections 4.3 and 5.3). Considering fast death of the differentiating progeny ($\Gamma \gg \varepsilon\omega$) and a small ratio of symmetric divisions ($\varepsilon \ll 1$), Eq. (27) enables to estimate the symmetric division rate of the stem cells as $\varepsilon\omega \approx \tau^{-1} \ln \gamma\psi$, where τ is the passage duration, γ is the fold-change in cell number from injection to harvesting and ψ is the amount of proliferating cells introduced above. Given the range of values for γ and τ given in Table S4, we obtain an estimate for the range of $\varepsilon\omega$ of $0.02 \ldots 0.06/\text{day}$. Since $\varepsilon < 1$ this automatically yields the lower bound $\omega \gtrsim 0.02/\text{day}$ for the stem cell division rate. In our model, the ratio ε of symmetric

Param.	Range	Description
ω	$0.02\ldots0.3\,\mathrm{d}^{-1}$	stem cell division rate
ε	$<1\ldots20\%$	probability of symmetric stem cell division
λ	$0.5\ldots1.5\,\mathrm{d}^{-1}$	progenitor loss-and-replacement rate
Γ	$0.2\ldots1.5\,\mathrm{d}^{-1}$	death rate of the differentiating progeny

Table S2 Parameter ranges for the model of tumor growth, Eq. (19).

stem cell divisions sets the relative size of the stem cell pool and the progenitor pool, see Eq. (23). Assuming that the stem-like cells form a minority population, we here restrict $\varepsilon \lesssim 20\%$.

A summary of the thus inferred parameter ranges is given in Table S2. To show that these estimates for the parameter ranges are consistent with the clonal behavior observed in experiments, we now compare numerical solutions of the model with experimental data.

6.2 Barcode frequency distributions

A direct quantitative comparison of barcode frequency distributions is currently not possible because of limitations in experimentally determining absolute barcode frequencies. However, the characteristic functional shape of the barcode frequency distributions is independent of absolute barcode frequencies and can be compared with experiments. To assess the barcode frequency distributions generated by our model, we obtain their first incomplete moment μ as defined in Eq. (8) from Eq. (21). Fig. 2e in the main text shows examples for μ for each passage, obtained from a numerical simulation of 2×10^6 realizations of the system. The linear behavior over many decades of barcode frequencies indicates a negative binomial size distribution as discussed in Section 2. In fact, we find these negative binomial distributions within a large range of parameters. This linear behavior is preceded by a short non-linear transient behavior for very small barcode frequencies that are likely below the experimental detection threshold.

6.3 Barcode survival

The survival of barcodes is reflected by the number of detected barcodes across passages. In experiments, the number of detected barcodes depends on the detection threshold and the fraction of sequenced cells. To obtain a measure for clone survival that is independent of these experimental constraints, we make use of the fact that barcode frequency distributions have the negative binomial form Eq. (7), which entails a characteristic barcode frequency n_0. This enables to define the number of clones that exceed a specified fraction θ of the characteristic barcode frequency n_0 as $\sum_{n>\theta n_0} h(n)$, where $h(n)$ is the number of clones with size n. The ratio of clones derived from initially injected barcoded cells that exceed the size θn_0 at a given passage therefore serves as a measure for barcode survival,

$$\beta_\theta = \frac{1}{N_\mathrm{B}} \sum_{n>\theta n_0} h(n) \, , \tag{30}$$

where N_B is the number of uniquely barcoded cells injected before the first passage, given by $N_\mathrm{B} = \eta n_1^\mathrm{inj}$ with η being the labelling efficiency and n_1^inj being the number of cells injected.

Fig. S2 shows the results from simulations[4] covering the parameter ranges indicated in Table S2, along with the corresponding experimental data[5]. Density bars show the distribution of values for $\beta_{1/2}$, dots show experimental data points[6]. Clearly, most of the values obtained in the biologically plausible parameter range also capture the experimentally obtained values. Moreover, simulations show a systematic decline of the growth probability with increasing passage number.

6.4 Correlations of barcode frequencies across passages

We now make use of the fact that unique barcoding enables to identify clones throughout different passages and replicate experiments. A characteristic feature of the clonal dynamics that includes this longitudinal data is the correlation of the size of a uniquely labelled clone across passages, see Extended Data Fig. 4f and Fig. 2h in the main text. To quantify these correlations, we define the normalized cross correlation

[4] A total of 108 parameter sets equally distributed in the parameter ranges for ω, ε, and λ indicated in Table S2 have been used to sample the parameter space. The parameter Γ was fixed according to Eq. (29). Each simulation consists of 100 000 realizations of clones using the passage times and number of injected cells reported in Table S4.

[5] We obtain β_θ from experimental data as follows. Barcode frequency distributions $h(x)$ with x being the relative barcode frequency are generated by binning the experimentally obtained barcode frequencies with a bin size of $(x_\mathrm{max} - x_\mathrm{min})/100$ where x_max and x_min are the largest and smallest relative barcode frequencyies, respectively. We then fitted the resulting barcode frequency distributions using the negative binomial form $p(x) = \mathcal{N}_0^{-1} \mathrm{e}^{-x/x_0}/x$ with n_0 and the normalisation constant $\mathcal{N}_0$ as fit parameters. Since the detection threshold from sequencing may distort the distributions for small barcode frequencies, we truncate the barcode frequency distributions from below (within the first 20 data points) such that the coefficient of determination R^2 of the fit is maximized. This yields the characteristic barcode frequency x_0 and β_θ is readily obtained as $\beta_\theta = N_\mathrm{B}^{-1} \sum_{x>\theta x_0} h(x)$. The standard error σ_{x_0} on x_0 obtained from the fit is used to calculate positive and negative errors for β_θ as $\sigma_\beta^\pm = N_\mathrm{B}^{-1} \sum_{x>\theta(x_0 \pm \sigma_{x_0})} h(x)$.

[6] The value $\theta = 1/2$ was chosen because the corresponding threshold $n_0/2$ lies well above the detection threshold from sequencing and at the same time takes into account most of the acquired data.

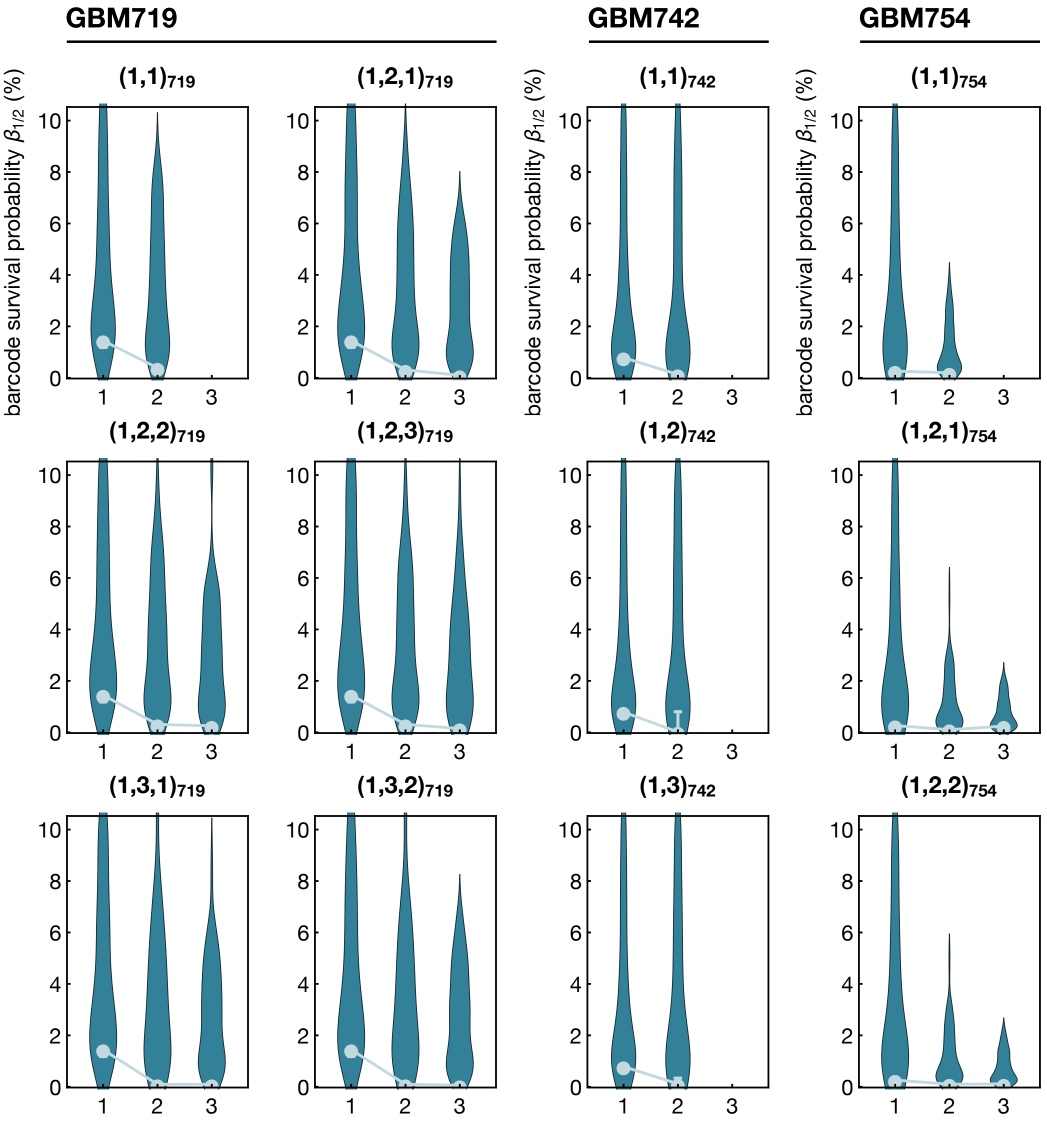


Figure S2 Fraction $\beta_{1/2}$ of initially injected barcodes growing above half of the characteristic barcode frequency $n_0/2$ as defined in Eq. (30) for all experimental trajectories given in Table S4. Density bars show the distribution of simulation results pooled over the parameter ranges indicated in Table S2. Dots show experimental data. The plot titles indicate the experimental trajectory as given in the first column of Table S4.

of the barcode frequency for passages i and j as

$$C_{ij} = \frac{\langle n_i n_j \rangle - \langle n_i \rangle \langle n_j \rangle}{\sqrt{\langle n_i^2 \rangle - \langle n_i \rangle^2}\sqrt{\langle n_j^2 \rangle - \langle n_j \rangle^2}} \,, \tag{31}$$

	ω	ε	λ	Γ	φ
Figs. 2d,e,h and S1, S4	$0.15\,\mathrm{d}^{-1}$	15%	$1\,\mathrm{d}^{-1}$	$0.48\,\mathrm{d}^{-1}$	0 %
Figs. 3e,f	$0.1\,\mathrm{d}^{-1}$	10%	$1.5\,\mathrm{d}^{-1}$	$0.74\,\mathrm{d}^{-1}$	0.5 %

Table S3 Parameter values used for the numerical examples in Figs. S1 and S4 and Figs. 2 and 3 in the main text. The parameters ω, ε, λ, and Γ are introduced in Sec. 4.1; the parameter φ is introduced in Sec. 6.5. These parameter sets are used to illustrate the model behavior and have therefore been chosen to be located in the center of the biologically plausible parameter ranges indicated in Table S2.

where $n_i = n(t_i^{\mathrm{harv}})$ is the barcode frequency after passage i. The normalized cross correlation C_{ij} takes values between -1 and 1, where $C_{ij} = 1$ indicates perfect correlation of barcode frequencies (i.e., small/large clones in passage i correspond to small/large clones in passage j), $C_{ij} = 0$ indicates that barcode frequencies are completely uncorrelated, and $C_{ij} = -1$ indicates perfect anticorrelation (i.e., large clones in passage i correspond to small clones in passage j and vice versa).

Fig. S3 shows a comparison of the correlations for the same simulations and experimental data sets as in Fig. S2. Density bars show the distribution of values for the cross correlations C_{ij}, dots show experimental data points. Without a fine tuning of the parameters, the theoretically computed cross correlations not only cover the experimentally obtained values in most cases but also clearly capture the correct trend between different pairwise comparisons within a particular injection series. In the case of the GBM754 experiment, deviations from experimental results is likely due to the comparably small number of detected clones which makes the cross correlation a less reliable measure; nevertheless, that the trend of correlations is largest between the secondary and tertiary passage is correctly captured.

6.5 Effects of chemotherapy

In the main text, we observed that the clonal behavior after chemotherapeutical treatment of xenografts with temozolomide (TMZ) can be characterized by two distinctive groups of small and large clones (termed Group A and Group B, respectively), see Fig. 3a–d. There we hypothesized that such a behavior is consistent with a subset of clones exhibiting a resistance to apoptosis. To assess whether our theory supports this scenario, we modified the simulation such that with a certain probability φ, a clone's differentiating progeny does not die off during the second passage ($\Gamma = 0$ for the respective clones). Fig. 3e,f in the main text shows the resulting correlations of barcode frequencies for the clones resisting apoptosis (blue dots, $\Gamma = 0$) and clones following the unperturbed dynamics (green dots, $\Gamma \neq 0$) for an example simulation with parameters given in Table S3. Indeed, the resulting behavior recapitulates

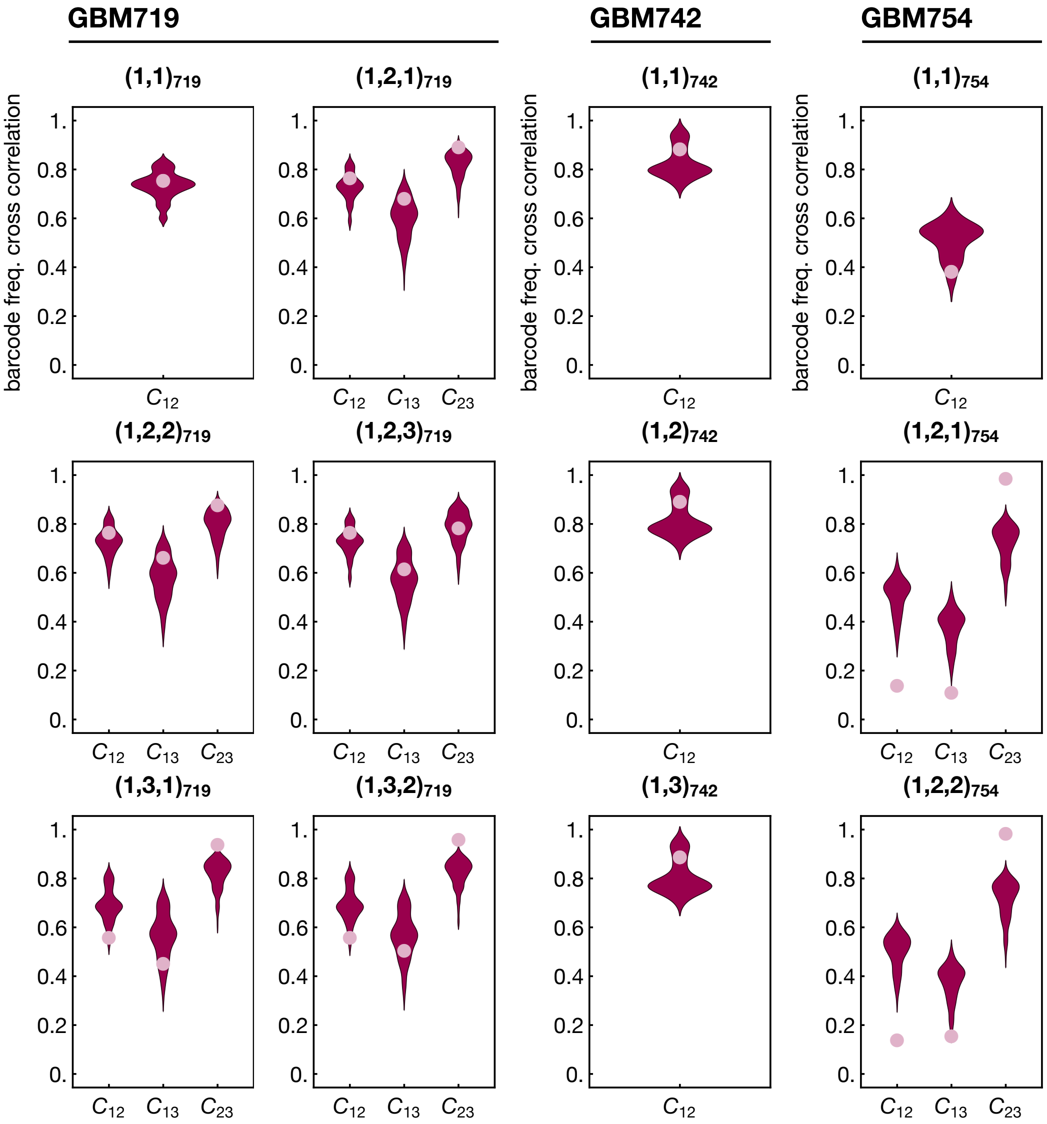


Figure S3 Barcode frequency cross correlations C_{ij} as defined in Eq. (31) for all experimental trajectories given in Table S4. Density bars show the distribution of simulation results pooled over the parameter ranges indicated in Table S2. Dots show experimental data. The plot titles indicate the experimental trajectory as given in the first column of Table S4.

the experimental findings: two clusters of small and large clones, respectively, with the size of large clones being positively correlated between subsequent passages, see Figs. 3a,b in the main text. These qualitative features of these correlations robustly appear without fine-tuning and within a large range of parameters, supporting that resistance to apoptosis of a subset of clones generically leads to the observed behavior.

7 Exome deep sequencing as a window on the mutational heterogeneity and clonal dynamics of GBM cells

To probe the mutational heterogeneity of the parent tumour and its evolution over time, we applied exome deep sequencing to xenografts from GBM719, focusing first on passage (p)2 and 3 of the untreated system. This analysis identified 546 mutations at p2 with variable allele frequencies (VAFs) that were above the threshold of detection, and 112 at p3. Analysis of the distribution of VAFs revealed a wide variation, with the majority clustered around the threshold value while some appeared to be clonally fixed within the population with VAFs of 0.5 or more. (Note that copy number variation can amplify VAFs above the value of 0.5, the value expected for a heterozygous point mutation that has become clonally fixed across of the population). Comparison of the mutational signature between p2 and p3 identified 68 mutant clones that were shared by both groups and therefore likely to be present in the parental tumour, emphasizing the mutational heterogeneity of both the parent tumour sample, and its conservation in the xenograft model.

As well as indicating the mutational heterogeneity of the tumour sample, the VAF also carries quantitative information on the relative abundance of point mutations within a sample and therefore carries information about the relative size of host mutant clones. Indeed, such data sets can often be used to identify cancer drivers and, in some cases, the phylogeny of mutations that drive non-neutral transformation (Williams et al., 2016; Eirew et al., 2015). However, in the present context, the current barcoding study indicates "neutral" competition between growing mutant clones suggesting that the vast majority of heterozygous point mutations, even when they occur in cancer genes, may leave the fate behaviour of tumour cells largely unperturbed. In this case, we can instead use point mutations as a surrogate clonal mark from which information on clonal dynamics of tumour cells can be inferred from the statistics of the ensemble of mutations. However, in contrast to cellular barcoding, where the clonal mark is created at a given instant in time, mutations occur sporadically leading to modified "clonal" distributions. As a result, the VAFs obtained from exome sequencing represent a product of both the underlying fate dynamics of the mutant cells within the sample and the mutational dynamics (Simons, 2016), involving the ongoing acquisition of new point mutations and copy number variations. Nevertheless, when copy number variation is low, such approaches can be used to quantify cell fate behaviour, as exemplified by a recent study of stem cell dynamics in physiological normal human epidermis obtained from punch biopsies of eyelid epidermis (Martincorena et al., 2015; Simons, 2016).

To develop a similar approach here, we reasoned that biopsies from primary tumours are likely to contain geographically restricted mutations (Johnson et al., 2014), further compounding the potential complexity of the VAF distribution. However, since normal cells are unlikely to survive passaging through the xenograft, we reasoned that VAFs obtained at p2 and p3 were likely to be rooted in the tumour-maintaining population. To address this data, we first considered the qualitative behaviour of the raw VAF distributions in both control (untreated) samples from p2 and p3. If, for a given locus, mutations of both alleles occur at a negligible rate, a VAF of 0.5 indicates a mutation present in the entire cell population and is therefore fixed across the population. Indeed, the VAF distributions in both samples (Extended Data Fig. 6d) exhibit an abundance of small clones as well as a smaller peak at VAFs of around 0.5, which likely corresponds to mutations that have already become fixed in the population after the respective passage.

Examining the correlations of VAFs between passages in xenografts (Extended Data Fig. 6e), we found a population of larger clones that are present after both passages, as expected for mutations that have become fixed (or almost fixed) at the end of p2. Alongside these clones, we also found both (i) clones that became extinct (or, more accurately, fell below the threshold of the deep sequencing) during repopulation and expansion in p3 as well as (ii) new clones that emerge during p3. If we assume that these new mutant clones arise from new mutations acquired during p3 (rather than from pre-existing clones that grew above the detection threshold), we can use the dynamics inferred from the barcoding to derive expected features of the VAF distribution of these newly-generated clones.

To predict the large-scale dependence of the VAF distribution, we adapted our simulation to take into account random "induction" of clones through mutations during the tertiary passage[7]. Model simulations suggested that the resulting VAF distribution again approximates a negative binomial form, or, equivalently, acquires an approximately exponential first incomplete moment consistent with experiment (Extended Data Fig. 6f, Fig. S4). Remarkably, focusing on the first incomplete moment of the 44 clones that emerge during the tertiary passage, we find that the first incomplete moment of the VAF distribution again reveals an exponential distribution (Extended Data Fig. 6g), in accordance with expectations from the barcoding study. By comparison, the TMZ-treated samples show (i) a much larger number of newly acquired mutations during p3 (Extended Data Fig. 6d) and (ii) a broad distribution

[7] Considering a constant mutation rate for each locus in each cell (Simons, 2016), the probability for a mutation to occur is proportional to the instantaneous number of cells in the tumour. Therefore, knowing that in our model specified by Eq. (20), the time-dependent fold-change in cell number is given by $\gamma(t) = \mathrm{e}^{\varepsilon\omega t}$, see Eq. (25), we reasoned that the time-dependent probability distribution for a mutation to have occured during the tertiary passage is given by $p_{\mathrm{ind}}(t) = \gamma(t)/\int_0^{\tau_3} \gamma(t)\,\mathrm{d}t$ with $0 \leq t \leq \tau_3$ where $t = 0$ refers to the start of the passage and where τ_3 is the passage duration of the tertiary passage. Hence, for each clone, we drew a time t_{ind} from the distribution p_{ind} and simulated the respective clone for the time $\tau_3 - t_{\mathrm{ind}}$, i.e., the remaining time from induction during the tertiary passage to the end of the passage. We then obtained the clone size distribution and first incomplete moment from the resulting clone population.

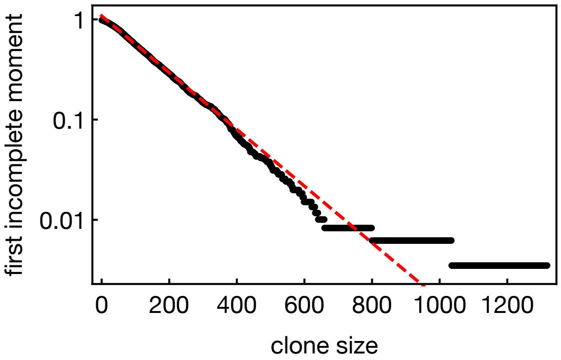


Figure S4 First incomplete moment of the clone size distribution obtained from a simulation of 10^5 clones with random induction times during the tertiary passage[7]. Parameters are given in the first row of Table S3. The red dashed line shows an exponential fit of the first incomplete moment.

of VAFs after p3 with a considerable subset of clones displaying VAFs larger than 0.5 (Extended Data Fig. 6d), both pointing at a treatment-induced higher genomic variability.

Although the agreement between the theoretical prediction based on the barcoding data and experiment is encouraging, we must also exercise some caution. While correction of VAFs to account for copy number variation (CNV) is already challenging in the parent tumour, with new mutations, the challenge is even greater. When CNV occurs before the mutation, the VAF provides a faithful read-out of clone size; where it occurs afterwards, the VAF is corrupted by the amplification. The correlation between VAFs associated with shared mutations between p2 and p3 of the control xenograft suggests that CNV may be rather infrequent as compared to the clonal dynamics, consistent with the systematic behaviour of the measured clone size distribution as predicted by a conserved proliferative hierarchy. In addition, we repeated the same analysis only taking genomic regions that are predicted to be diploid within each sample based on exome sequencing. After filtering, the VAF distributions continue to conform to the negative binomial (Extended Data Fig. 6h-i). However, a more detailed quantitative analysis would require a comprehensive investigation and understanding of the interplay between tumour growth, mutational dynamics and, indeed, chemotherapy-induced mutation (Johnson et al., 2014), which are beyond the scope of the current study.

8 Remarks

Here we have introduced a theoretical model of human glioblastoma (GBM) growth based on a critical birth-death process with immigration, describing the stochastic cell fate dynamics of a proliferative hierarchy with glioblastoma stem cells (GSCs) at the apex. Our model is able to robustly capture key features of the clonal dynamics assessed experimentally: importantly, it explains the characteristic negative binomial barcode frequency distributions across all serial passages observed in experiments. Moreover, comparison of (i) the number of surviving barcodes across serial passages and (ii) correlations of barcode frequencies between serial passages show that the inferred parameter range covers the observed behavior in the overwhelming majority of cases.

Note that the model presented here is still a minimal model in the sense that more complex alterations and refinements are conceivable. These may include a slight imbalance between loss and replacement of progenitors as well as multiple progeny compartments. Also, small amounts of cell death may occur in the stem cell and progenitor compartments. However, if the death rate was of comparable size (or larger) than the rate of symmetric proliferation, we would expect a massive loss of clones. If, on the contrary, cell death only represents a small contribution relative to the symmetric proliferation rate, it could be accounted for by an effective adjustment of the other model parameters that, e.g., determine the net growth of tumor and would only be visible in subtle changes of the barcode frequency distributions that are impossible to detect in the experimentally given distributions. However, these alterations do not change the basic characteristics of our model. Moreover, we have neglected the spatial aspect of tumor growth and potential ongoing driver gene mutations (Michor et al., 2006; Waclaw et al., 2015), assuming that cell division and loss-and-replacement occur at constant rates as the tumor expands within the brain. The fact that, despite its simplicity, our model is able to capture the main features of the clonal dynamics indicates a remarkably simple proliferative behavior of human GBM despite the genomic variability of GBM cells.

References

Bailey, N. T. J. (1990). *The Elements of Stochastic Processes with Applications to the Natural Sciences*. A Wiley publication in applied statistics. Wiley.

Blanpain, C. and Simons, B. D. (2013). Unravelling stem cell dynamics by lineage tracing. *Nat. Rev. Mol. Cell Biol.*, 14:489–502.

Clayton, E., Doupé, D. P., Klein, A. M., Winton, D. J., Simons, B. D., and Jones, P. H. (2007). A single type of progenitor cell maintains normal epidermis. *Nature*, 446:185–189.

Eirew, P., Steif, A., Khattra, J., Ha, G., Yap, D., Farahani, H., Gelmon, K., Chia, S., Mar, C., Wan, A., Laks, E., Biele, J., Shumansky, K., Rosner, J., McPherson, A., Nielsen, C., Roth, A. J. L., Lefebvre, C., Bashashati, A., de Souza, C., Siu, C., Aniba, R., Brimhall, J., Oloumi, A., Osako, T., Bruna, A., Sandoval, J. L., Algara, T., Greenwood, W., Leung, K., Cheng, H., Xue, H., Wang, Y., Lin, D., Mungall, A. J., Moore, R., Zhao, Y., Lorette, J., Nguyen, L., Huntsman, D., Eaves, C. J., Hansen, C., Marra, M. A., Caldas, C., Shah, S. P., and Aparicio, S. (2015). Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*, 518(7539):422–426.

Gardiner, C. (2009). *Stochastic Methods: A Handbook for the Natural and Social Sciences*. Springer Series in Synergetics. Springer.

Gillespie, D. T. (1977). Exact Stochastic Simulation of Coupled Chemical Reactions. *J. Phys. Chem.*, 93555(1):2340–2361.

Johnson, B. E., Mazor, T., Hong, C., Barnes, M., Aihara, K., McLean, C. Y., Fouse, S. D., Yamamoto, S., Ueda, H., Tatsuno, K., Asthana, S., Jalbert, L. E., Nelson, S. J., Bollen, A. W., Gustafson, W. C., Charron, E., Weiss, W. A., Smirnov, I. V., Song, J. S., Olshen, A. B., Cha, S., Zhao, Y., Moore, R. A., Mungall, A. J., Jones, S. J. M., Hirst, M., Marra, M. A., Saito, N., Aburatani, H., Mukasa, A., Berger, M. S., Chang, S. M., Taylor, B. S., and Costello, F. F. (2014). Mutational Analysis Reveals the Origin and Therapy-Driven Evolution of Recurrent Glioma. *Science*, 343(6167):189–193.

L. V. Nguyen, M. Makarem, *et al.* (2014). Clonal Analysis via Barcoding Reveals Diverse Growth and Differentiation of Transplanted Mouse and Human Mammary Stem Cells. *Cell Stem Cell*, 14:253–263.

Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D. C., Fullam, A., Alexandrov, L. B., Tubio, J. M., Stebbings, L., Menzies, A., Widaa, S., Stratton, M. R., Jones, P. H., and Campbell, P. J. (2015). High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, 348(6237):880–886.

Michor, F., Iwasa, Y., and Nowak, M. A. (2006). The age incidence of chronic myeloid leukemia can be explained by a one-mutation model. *Proc. Natl. Acad. Sci. USA*, 103(40):14931–14934.

Simons, B. D. (2016). Deep sequencing as a probe of normal stem cell fate and preneoplasia in human epidermis. *Proc. Natl. Acad. Sci. USA*, 113:128–133.

Waclaw, B., Bozic, I., Pittman, M. E., Hruban, R. H., Vogelstein, B., and Nowak, M. A. (2015). A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. *Nature*, 525(7568):261–264.

Walczak, A. M., Mugler, A., and Wiggins, C. H. (2012). Analytic methods for modeling stochastic regulatory networks. *Meth. Mol. Biol.*, 880(Chapter 13):273–322.

Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A., and Sottoriva, A. (2016). Identification of neutral tumor evolution across cancer types. *Nat. Genet.*, 48(3):238–244.

ID	Passage	inj. cells n^{inj}	τ	surv. prob. $\beta_{1/2}$	growth γ
		GBM719			
$(1)_{719}$	Primary	1.25×10^5	79 d	$1.44^{+0.}_{-0.26}\%$	40
$(1,1)_{719}$	— Secondary	3×10^5	64 d	$0.32^{+0.06}_{-0.07}\%$	7.7
$(1,2)_{719}$	— Secondary	3×10^5	68 d	$0.31^{+0.08}_{-0.07}\%$	18
$(1,2,1)_{719}$	—– Tertiary	3×10^5	55 d	$0.13^{+0.07}_{-0.03}\%$	42
$(1,2,2)_{719}$	—– Tertiary	3×10^5	70 d	$0.25^{+0.04}_{-0.03}\%$	43.3
$(1,2,3)_{719}$	—– Tertiary	3×10^5	78 d	$0.18^{+0.04}_{-0.03}\%$	26
$(1,3)_{719}$	— Secondary	3×10^5	89 d	$0.03^{+0.05}_{-0.01}\%$	16.3
$(1,3,1)_{719}$	—– Tertiary	3×10^5	66 d	$0.13^{+0.03}_{-0.}\%$	15.1
$(1,3,2)_{719}$	—– Tertiary	3×10^5	62 d	$0.07^{+0.03}_{-0.01}\%$	47.7
		GBM754			
$(1)_{754}$	Primary	1.25×10^5	99 d	$0.26^{+0.}_{-0.}\%$	26
$(1,1)_{754}$	— Secondary	6×10^4	79 d	$0.19^{+0.}_{-0.}\%$	42.7
$(1,2)_{754}$	— Secondary	6×10^4	86 d	$0.1^{+0.}_{-0.}\%$	98.3
$(1,2,1)_{754}$	—– Tertiary	6×10^4	72 d	$0.24^{+0.}_{-0.}\%$	56
$(1,2,2)_{754}$	—– Tertiary	6×10^4	73 d	$0.11^{+0.}_{-0.}\%$	31.7
		GBM742			
$(1)_{742}$	Primary	2.4×10^4	78 d	$0.78^{+0.}_{-0.}\%$	530
$(1,1)_{742}$	— Secondary	3×10^5	43 d	$0.13^{+0.06}_{-0.}\%$	5.7
$(1,2)_{742}$	— Secondary	3×10^5	47 d	$0.01^{+0.78}_{-0.01}\%$	1.8
$(1,3)_{742}$	— Secondary	3×10^5	50 d	$0.12^{+0.22}_{-0.03}\%$	8.5

Table S4 Experimental data sets used to compare with theory. Here, n^{inj} is the number of injected cells, τ is the passage duration, s is the fraction of cells sequenced, $\beta_{1/2}$ is the fraction of initially injected barcodes growing above half of the characteristic barcode frequency $n_0/2$, as defined in Eq. (30), and γ is the estimated fold-change in cell number between injection and harvesting, which quantifies tumor growth. In all cases, cells were harvested and injected from the ipsilateral side.