A cellular hierarchy in melanoma uncouples growth and metastatic phenotypes

Table of contents

Supplementary Table Legends

Supplementary Table 1: Differentially expressed genes of malignant Seurat clusters (mouse) related to Fig. 1 and Extended Data Fig. 1. P value was calculated using non-parametric Wilcoxon rank sum test and adjusted p value with Bonferroni correction for multiple testing.

Supplementary Table 2: Functionally enriched terms for top 120-150 sign. overexpressed in each Seurat cluster, related to Fig.1c. Significance was assessed using Fisher exact test, two-tailed p value**.**

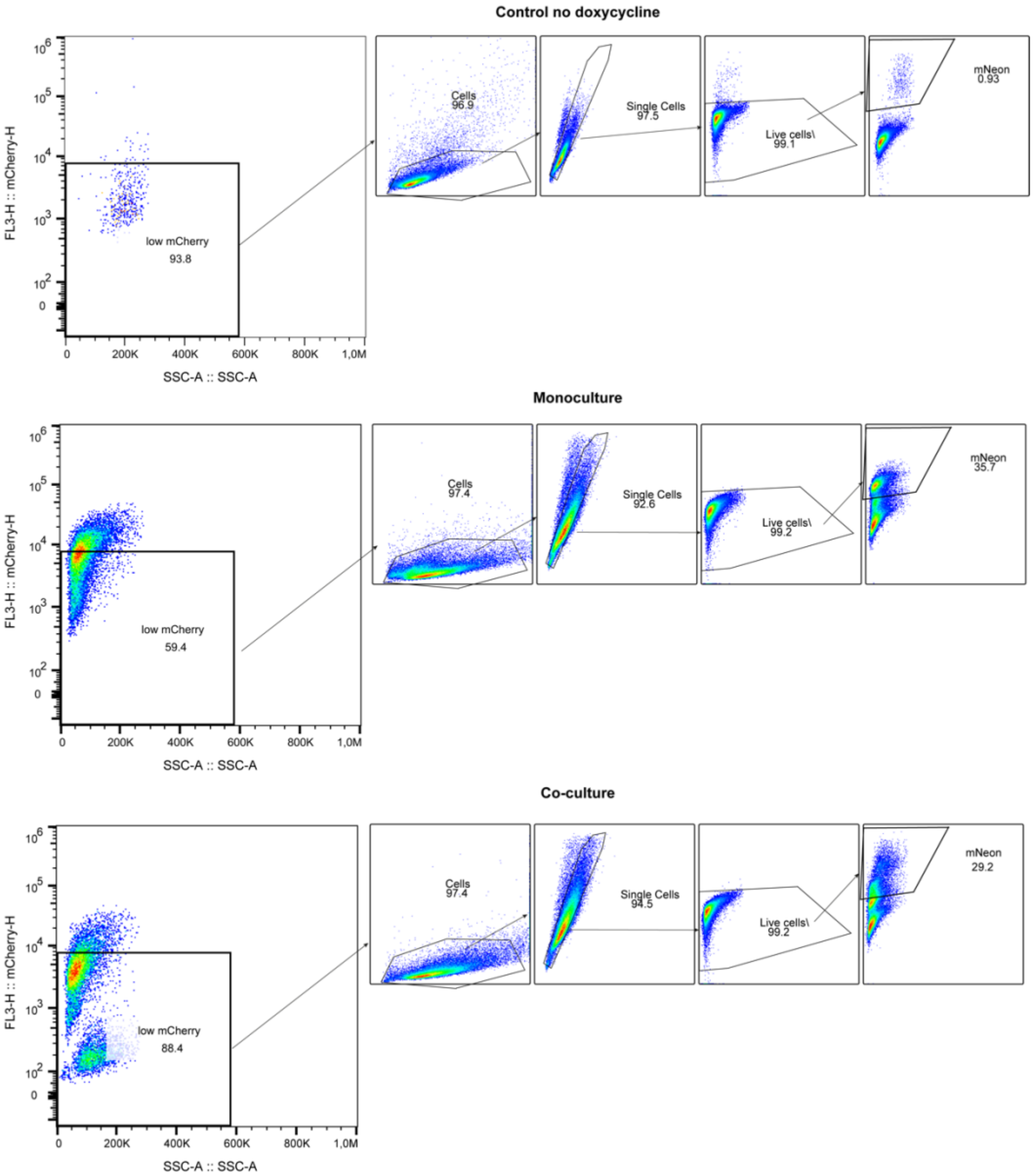
Supplementary Table 3: Functionally enriched gene expression signatures related to Fig.1d, e, g and Extended Data 1,2.

Supplementary Table 4: Parameters estimated from the fit of the bi-exponential decay to the CDF of clone sizes for each sample (see Model fits section of the Supplementary Note, Fig. 2f and Extended data Fig. 4b-e). The chase time (t_0), total number of clones and number of singlets (single cell clones) are shown for each sample, together with the average clone size (considering all clones) and measured tumour expansion relative to its volume at the start of the chase period. The clone size threshold for finding the long term decay, the fitted values for the large and small clone size decays, $\bar{n}$ and $\bar{n}_p$, and composite parameter p_0, are shown. These fits were made by first removing single cell clones from the data, to account only for clones rooted in proliferative cells, using a least-square method (see Supplementary Note).

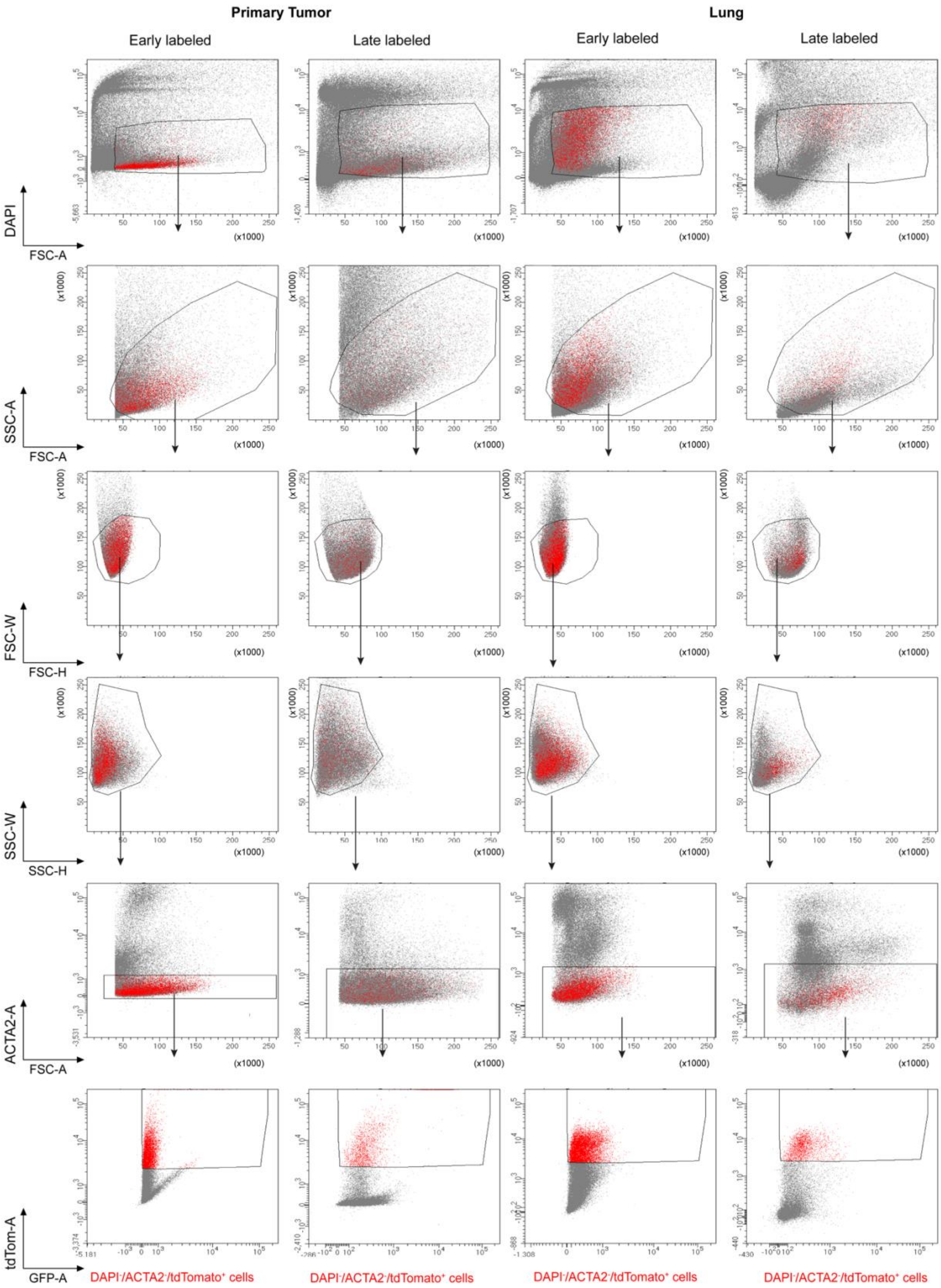
Supplementary Table 5: Parameters of the two-compartment stem-progenitor cell model, estimated using the fitted values for the large and small clone size decays, $\bar{n}$ and $\bar{n}_p$, and composite parameter p_0 (see Supplementary Table 4), together with the predicter tumour volume. The parameters shown here were used in the numerical simulations of the full stochastic model (see Supplementary Note and Extended Data Fig. 4g-k).

Supplementary Table 6: Gene list inferred from scRNA seq of *Tyr::NRAS*$^{Q61K/\circ}$*;Ink4a*$^{-/-}$ allograft model and projected to Stereo-Seq technology to spatially resolve pre-EMT and hypoxic cell states in respect to Endothelial Cells. Related to Fig. 3c-d and Extended Data Fig. 6d-g.

Supplementary Table 7: Molecular Cartography probes used in this study. The discriminative markers were selected based on scRNA seq of *Tyr::NRAS*$^{Q61K/\circ}$*;Ink4a*$^{-/-}$ allograft model to spatially resolve cell type and melanoma cell state diversity. Related to Fig. 3e,f and Extended Data Fig. 6h,i,k,m.



Supplementary Figure 1. Analysis of mCherry dilution in Watermelon melanoma cells upon co-culturing with Bend3 cells *in vitro*. Gating strategy for analyzing low mcherry population. mCherry fluorescence was evaluated within mNeon-positive population, allowing to exclude Bend3 cells. Watermelon expressing *NRAS*$^{Q61K/°}$;*Ink4a*$^{-/-}$ cells non treated with doxycycline were used as a negative control (see Reporting Summary). Final gates are presented in Extended Data Fig. 8d. The outcome of the analyses is shown in Fig. 3j. Similar gating strategy was applied for the experiments with Watermelon-expressing cells upon Notch3 genetic inactivation shown in Fig. 3n.


Primary Tumor
Lung
Early labeled
Late labeled
Early labeled
Late labeled
DAPI
FSC-A
SSC-A
FSC-A
FSC-W
FSC-H
SSC-W
SSC-H
ACTA2-A
FSC-A
tdTom-A
GFP-A
(x1000)
DAPI-/ACTA2-/tdTomato+ cells
DAPI-/ACTA2-/tdTomato+ cells
DAPI-/ACTA2-/tdTomato+ cells
DAPI-/ACTA2-/tdTomato+ cells

Supplementary Figure 2. Isolation of tdTom$^+$ cells from *Prrx1::CreER-GFP;Tyr-NRAS$^{Q61K/°}$;Ink4a$^{-/-}$;ROSA26R$^{LSL\text{-}tdTomato/LSL\text{-}tdTomato}$*. Gating strategy for the isolation of tdTom$^+$ cells from early vs late labeled mice (upon tamoxifen administration). Isolated fractions were subsequently subjected to scRNA sequencing using 10x platform. The outcome of the analyses is presented in Fig. 4e-g, i and Extended Data Fig. 10a-c, g.

Supplementary Note

Here, we present in further detail the theoretical framework used to investigate the clonal dynamics of melanoma growth, and in particular how the experimental observations impose constraints on the cellular dynamic. In section 1, we describe how clonal information on tumour cell dynamics is extracted from thick tumour sections. In section 2, we show how the measured distribution of clone sizes provides evidence for a proliferative hierarchy, which we formalise mathematically as a two-compartment model. In section 3, we describe the strategy used to fit the two-component model to the experimental clone size data and extract the model parameters. Finally, in section 4, we show how short-term clonal labelling and staining of vasculature provides further support for the hierarchical model. Throughout, we emphasize the value and strengths of the modelling scheme, as well as discussing its limitations. Readers who would prefer to focus only on the definition of the two-component model and its fit to the clonal data may turn to the Model fits section 3 below, where a summary of the model, its fit parameters and the fitting procedure are presented in full.

The scripts and data used in the clonal analysis are available at `https://github.com/ibordeu/scripts_Karras_et_al_2022_git`.

1 Clone reconstruction

As detailed in the main text, we used a multicolour genetic lineage tracing approach based on the Confetti reporter system to study the dynamics of tumour growth, with approximately 0.5-1% of all cells marked by one of four Confetti colours when samples were collected. Tumours were collected and imaged once they had expanded to around 10 times their initial size of around 150 mm^3 (see column marked “Tumour expansion” in Supplementary Table 4 for precise values). To avoid challenges in identifying cells marked by membrane-bound

CFP, and the scarcity of nuclear GFP labelled cells, we focused our quantitative analysis on cells positive for either YFP or RFP expression.

From a qualitative analysis of tumour sections, we found that cells marked by a common colour tended to associate spatially into clusters. In some cases, clusters were large and densely labelled by marked cells (Fig. 2c). In other cases, clusters showed evidence of dispersion, with fragments of marked cell clusters separated from other clusters by unlabelled cells. However, many clusters were spatially isolated and contained just a few cells or less. Together, these findings were suggestive of a bimodal type of behaviour, with some tumour cells founding large clones that grow and progressively disperse, while other tumour cells give rise to a more limited output.

To gain quantitative insight into the tumour cell dynamics, we sought to reconstruct total clone sizes based on 3D reconstructions from the confocal images of thick sections. Specifically, we first used a fluorescence intensity-based pipeline to identify the position and volume of clusters of cells marked by a common Confetti colour and in close physical contact. Then, by computing the spatial correlation function of clusters marked by the same colour and different colours, we determined that clusters separated by a distance of 100μm or less (translating to around 6-7 cell diameters) were likely to belong the same clone (Fig. 2d, Extended Data Fig. 4a and Methods). Based on this assignment, we then performed a visual inspection of reconstructed clones to check on the consistency of the clustering and to remove potential artefacts, such a false positive labelling. In doing so, we noticed that there were a few clones that, based on spatial morphology, were likely to be associated with the chance merger of large proximate clones. In one dataset (specifically the RFP channel of sample 3, Extended Data Fig. 4c) we found that an elevated induction efficiency led to a preponderance of over-sized clones, consistent with a multiplicity of chance of merger events, leading us to conclude that this dataset was below the threshold at which clonal reconstruction could be performed reliably (see Extended data Fig. 4c). However, for the rest of the samples and colour channels, the clonal assignments led to a size distribution that was broadly consistent across samples and colours (Fig. 2e), both in scale and shape. The total number of reconstructed clones for each sample and colour channel are shown in Supplementary Table 4.

Once clones were identified, we then inferred their corresponding constituent cell number by scaling their volume against the estimated average volume of individual cells based on fluorescence intensity, the latter estimated at around 1,800 μm^3, for an average diameter of 15 μm (see Methods). Notably, based on this calibration, we found that the average clone size was broadly consistent with the net expansion of the tumour volume, providing further support for the integrity of clonal assignments (Supplementary Table 4). Consistent with the qualitative analysis of the sectional data, the cumulative clone size distributions based on the 3D reconstructions showed evidence of a bimodal behaviour, with small clone sizes characterised by a rapid exponential-like decay with a sharp crossover to a slower exponential-like decay at larger clone sizes (Fig. 2f and Extended Data Fig. 4b-e).

This observation was revealing: Alongside potential intrinsic heterogeneities associated with the complex organization of the tumour cell hierarchy, divergences in the mutational landscape together with spatial fluctuations associated with local tumour microenvironments might be expected to lead to an unstructured clone size distribution. Yet, these findings suggest that the heterogeneity of clone sizes, from the smallest clones containing just a handful of cells, to the largest clones with hundreds or thousands of cells, can be captured quantitatively with just three parameters: The two exponential decay rates and the crossover clone size. We therefore sought to question what is the minimal model of cell fate that could capture the observed dynamics, turning later to consider the potential limitations of this scheme. In the following, we will show that the observed clone size behaviour is characteristic of a hierarchical proliferative organization with an expanding stem cell-like population giving rise to a second progenitor cell-like population.

2 Theory

To assess the potential validity of a hierarchical organization, we considered the dynamics of a two-compartment model comprised of a stem and a progenitor cell population. To simplify the analysis, we considered a minimal model in which stem cells select stochastically (i.e., probabilistically) between cell duplication and loss through differentiation, while progenitor cells expand through stochastic cell duplication (Fig. 2g),

$$\begin{aligned} &S \xrightarrow{\sigma} \begin{cases} S+S & \text{Pr. } r \\ P & \text{Pr. } 1-r \end{cases} \\ &P \xrightarrow{\Delta_P} P+P. \end{aligned} \tag{1}$$

Here $r\sigma$ defines the stem cell duplication rate, $(1-r)\sigma$ denotes the effective differentiation rate into the progenitor cell compartment, and Δ_P denotes the expansion rate of the progenitor cell population. Later, we will consider how generalizations of the model give rise to the same long-term clone size dependencies.

Since the two-compartment model does not admit a formal exact analytic solution, we first questioned its behaviour based on the results of stochastic simulation. Using a Gillespie algorithm to simulate the model dynamics (see implementation details below), numerical analysis showed that, over a wide range of parameters, the chance of finding a clone of size n cells, $T_n(t)$, after a time t showed a bi-exponential type dependence, with

$$T_n(t) \simeq \frac{1-c_0}{\bar{n}_1(t)} e^{-n/\bar{n}_1(t)} + \frac{c_0}{\bar{n}_2(t)} e^{-n/\bar{n}_2(t)}, \tag{2}$$

where, with $\bar{n}_2(t) \gg \bar{n}_1(t)$, the first term represents the small clone size dependence and the latter the large. We therefore used this empirical observation to determine approximate analytical expressions for the coefficients c_0, $\bar{n}_1(t)$ and $\bar{n}_2(t)$.

As a starting point, we noted that, within the framework of the two-compartment model, the average size of the stem and progenitor cell populations, S and P, are defined by the rate equations

$$\begin{aligned} \dot{S} &= \Delta_S S \\ \dot{P} &= \Delta_P P + (1-r)\sigma S, \end{aligned}$$

where $\Delta_S = \sigma(2r-1)$ denotes the net expansion rate of the stem cell population. With the initial condition $S(0) = 1$ and $P(0) = 0$, these equations have the solution

$$\begin{aligned} S(t) &= e^{\Delta_S t} \\ P(t) &= \frac{(1-r)}{2r-1} \frac{\Delta_S}{\Delta_S - \Delta_P} \left(e^{\Delta_S t} - e^{\Delta_P t}\right). \end{aligned} \tag{3}$$

With the stem cell dynamics defined by a general birth-death type process, the chance of finding a clone of size n stem cells after a chase time t is given by [60]

$$S_n(t) = \begin{cases} \alpha & n = 0 \\ (1-\alpha)(1-\beta)\beta^{n-1} & n > 0 \end{cases}, \tag{4}$$

defining $\omega \equiv r/(1-r)$, $\alpha(t) = (S(t)-1)/(\omega S(t)-1)$ and $\beta(t) = \omega(S(t)-1)/(\omega S(t)-1)$.

To obtain an approximation for the total clone size distribution, T_n, we separated the clone size dependence as

$$T_n(t) = T_n^{<}(t) + T_n^{>}(t),$$

where $T_n^<$ represents the small clone size contribution derived predominantly from "extinct" clones, defined as those containing no stem cells, and $T_n^>$ represents the contribution from "surviving" (i.e., stem cell-containing) clones. From the results of the stochastic simulation of the hierarchical model, we expect that $T_n^>(t) = \frac{c_0}{\bar{n}(t)} e^{-n/\bar{n}(t)}$. In the long-time limit, we reasoned that extinct clones will make a negligible contribution to the average total clone size since, with $r > 1/2$, their contribution will become rapidly overwhelmed by the expanding stem cell-containing clones. Therefore, to leading order, we can approximate $T_n^<(t) \simeq T_0^< \delta_{n,0}$. With these definitions, we noted the further constraints from the normalization and the average total clone size, which require that

$$\sum_{n=1}^{\infty} T_n(t) = 1 \tag{5}$$

$$\sum_{n=1}^{\infty} n T_n(t) = T(t). \tag{6}$$

From the normalisation condition (5), it follows that $c_0 = 1 - T_0^<$. From the second constraint (6), it follows that the average total clone size satisfies the relation $T(t) = (1 - T_0^<)\bar{n}(t)$, which translates to the identity $\bar{n}(t) = T(t)/(1 - T_0^<)$. However, from Eq. (4), it follows that the number of extinct clones is given by $T_0^< \equiv \lim_{t\to\infty} S_0(t) = 1/\omega$. With $1 - T_0^< = 1 - 1/\omega = 2 - 1/r = (2r-1)/r$, we thus have that

$$\bar{n}(t) = \frac{r}{2r-1} T(t) = \frac{r}{2r-1}(S(t) + P(t)).$$

In summary, to a first approximation, we find that the total clone size distribution takes the form

$$T_n(t) \simeq \frac{1-r}{r}\delta_{n,0} + \frac{2r-1}{r}\frac{1}{\bar{n}(t)} e^{-n/\bar{n}(t)}.$$

Although this equation fits well with the empirical long-term large clone size dependence, it was instructive to develop a more refined approximation that captures the shorter-time behaviour. If the first stem cell fate decision leads to a transition $S \mapsto P$, the resulting progenitor will give rise to an exponential contribution to the clone size distribution, $P_n(t) \simeq e^{-n/\bar{n}_P(t)}/\bar{n}_P(t)$ with $\bar{n}_P = e^{\Delta_P t}$. Alternatively, a stem cell might multiply and expand before finally becoming extinct, generating multiple progenitor cells over its lifetime. However, if $r \sim 1$, the majority of extinction events will derive from the first decision. (Note that, for a super-critical birth-death process, $r > 1/2$, once a clone has expanded beyond a certain size, its chance of extinction becomes negligible). Therefore, taking the leading order contribution from extinction events, we can approximate the small clone size distribution by the exponential distribution of a progenitor cell-derived clone created at close to the induction time,

$$T_n^<(t) = \frac{1}{\omega}\frac{1}{\bar{n}_P(t)}e^{-n/\bar{n}_P(t)}.$$

Although the revised distribution,

$$T_n(t) = \frac{1-r}{r}\frac{1}{\bar{n}_P(t)}e^{-n/\bar{n}_P(t)} + \frac{2r-1}{r}\frac{1}{\bar{n}(t)}e^{-n/\bar{n}(t)}$$

is still normalised, since the first term now contributes to the average clone size, we have to revise $\bar{n}(t)$ accordingly. In this case, going back to the constraint on the average clone size, we now have the condition

$$T(t) = \frac{1-r}{r}\bar{n}_P(t) + \frac{2r-1}{r}\bar{n}(t) \tag{7}$$

from which follows the relation

$$\bar{n}(t) = \frac{1}{2r-1}\left(rT(t) - (1-r)\bar{n}_P\right). \tag{8}$$

In summary, to leading order, the total clone size distribution takes the form

$$T_n(t) \simeq \frac{1-r}{r}\frac{1}{n_P(t)}e^{-n/n_P(t)} + \frac{2r-1}{r}\frac{1}{\bar{n}(t)}e^{-n/\bar{n}(t)}. \tag{9}$$

Note that this expression becomes accurate in the limit $r \to 1$, but becomes increasingly unreliable as $r \to 1/2$. In this limit, further contributions to the extinct clone population will affect the small clone size dependence of the distribution, which in turn effects the net normalization. Indeed, extrapolating to the next order of approximation, one may show that

$$T_n(t) \simeq \frac{1-r}{r(1+r(1-r))}\left(1 + r(1-r)\frac{n}{\bar{n}_P}\right)\frac{e^{-n/\bar{n}_P}}{\bar{n}_P} + \frac{2r-1}{r}\frac{e^{-n/\bar{n}}}{\bar{n}}.$$

However, in the following, since the major contribution to the small clone sizes experimentally arises from the direct induction of progenitors, we kept only the leading order of approximation (9) for simplicity.

Finally, to complete the analysis, we then considered how the contribution of progenitor cell-derived clones revises the expression for the clone size distribution. If, on induction, a fraction f_S of induced cells belong to the stem cell compartment, with the remaining $1 - f_S$ belonging to the progenitor compartment, from the results above, it follows that the total clone size distribution is given by

$$T_n(t) \simeq \left(1 - f_S \frac{2r-1}{r}\right) \frac{1}{\bar{n}_P(t)} e^{-n/\bar{n}_P(t)} + f_S \frac{2r-1}{r} \frac{1}{\bar{n}(t)} e^{-n/\bar{n}(t)}. \qquad (10)$$

Note that, as mentioned above, if the stem cell fraction f_S is small, then the progenitor cell contribution dominates the small clone size dependence, mitigating to some extent the need to include the higher order corrections for r close to $1/2$. Lastly, within this framework, the total tumour size grows as

$$V(t) = f_S T(t) + (1 - f_S)\bar{n}_P(t),$$

the first term arising from the stem cell-derived clones and the second from the progenitor population. Then, using (7), we obtain the leading approximation

$$V(t) \simeq f_S \frac{2r-1}{r} \bar{n}(t) + \left(1 - \frac{2r-1}{r} f_S\right) \bar{n}_P(t). \qquad (11)$$

Model generalisations

Before turning to the fit of the model to the clonal data, it is helpful to first revisit how potential generalisations of the model can be captured within the framework of the same minimal scheme (1). Starting with the stem cell compartment, previously we considered a dynamics based on a simple birth-death type process. Such a model would capture, for example, a situation in which stem cell competence was linked to proximity to a niche site. As cells move away from the niche, they enter irreversibly into a differentiation programme that leads to their transition into a progenitor P state. However, if fate were assigned during division, we might consider a generalised model in which all three fate outcomes contribute,

$$S \xrightarrow{\sigma} \begin{cases} S + S & \text{Pr. } u \\ S + P & \text{Pr. } v \\ P + P & \text{Pr. } 1 - u - v \end{cases} . \qquad (12)$$

In this case, the kinetic equations for the average cell numbers are given by

$$\dot{S} = \Delta_S S$$

$$\dot{P} = (2 - 2u - v)\sigma P + \Delta_P P,$$

where $\Delta_S = (2u + v - 1)\sigma$. With the initial condition $S(0) = 1$ and $P(0) = 0$, these equations have the solution

$$\begin{aligned} S(t) &= e^{\Delta_S t} \\ P(t) &= \frac{(2 - 2u - v)}{2u + v - 1} \frac{\Delta_S}{\Delta_S - \Delta_P} \left(e^{\Delta_S t} - e^{\Delta_P t} \right). \end{aligned}$$

At the same time, the distribution of stem cell clone sizes will take the same exponential-like form as above (4), but with $r = u/(1 - v)$. Therefore, taken together, these results lead to the same bi-exponential clone size dependence of the distribution, where the effective rate constants and probabilities are appropriately renormalised.

Similarly, in the model, Eq. (1), we considered a progenitor cell compartment that was purely duplicative. In practice, it is likely that the progenitor cells can both proliferate and transition into a more differentiated non-cycling state,

$$\begin{aligned} & P \xrightarrow{\lambda} \begin{cases} P + P & \text{Pr. } q \\ D & \text{Pr. } 1 - q \end{cases} \\ & D \xrightarrow{\gamma} 0. \end{aligned}$$

However, once again, the dynamics of the progenitor cell population will still be defined by a birth-death process, giving rise to an exponential distribution of clone sizes, at least for the progenitor cell compartment. Therefore, provided that the dynamics remains supercritical $q > 1/2$, and/or the differentiated cells D are short-lived (i.e., have a high loss rate γ), apart from a revision of the parameters, such a generalisation would not change the bi-exponential form of the clone size distribution.

In our minimal modelling scheme, we have considered just one progenitor cell type, P. By contrast, the scRNA-seq data suggests that tumour cells may be heterogeneous, comprising multiple proliferative cell types, mirroring the developmental trajectories of neural crest sublineages. In this case, one might wonder whether the stem cell-like population could transition into a diversity of progenitor sublineages, P_1, P_2, etc. However, provided that the dynamics of the progenitor cell population is dominated either by a single sublineage, as would be expected for a net bias towards exponential growth, or a common progenitor, the dynamics would still be captured by the minimal model.

Finally, in formulating the model, we have not considered how spatial influences could impact on the clonal dynamics. A niche-like organization, such as that imposed by the vasculature, could limit the clonal expansion of the stem cell compartment along preferred directions. However, previous studies have shown that such constraints impact only in low spatial dimension [61]. In the context of the three-dimensional tissue, and the extended arrangement of the vascular niche, these constraints would not be expected to impact, allowing us to apply the minimal zero-dimensional description above.

While these arguments provide a rationale to explain the integrity of the observed bi-exponential clone size dependence against potential complexities of tumour cell fate and heterogeneities of the tumour microenvironment, they also impose some limitations on the predictive power of the modelling scheme. In particular, while we may hope to retrieve the parameters of the model - the expansion rates and renewal bias of the stem cell compartment - individual cell fate behaviours (viz. symmetric vs. asymmetric division probabilities) and division rates, as well as potential substructures of the progenitor and differentiated cell compartments are not accessible. For this reason, in fitting the data, we have stayed within the framework of the minimal model (1).

3 Model fits

In section 2, we defined a stochastic two-compartment model of tumour growth, Eq. (1), based on a stem-progenitor cell hierarchy. The model has three parameters: the stem cell cycling rate σ, the duplication probability r, and the progenitor cell expansion rate Δ_p. However, these parameters are not accessible directly from fits to the observed bi-exponential dependence of the clone size distribution, but can be inferred *a posteriori*. Here, we detail the strategy used in fitting the predicted bi-exponential dependence to the experimental data and the resulting inference of the model parameters. Equation (10) has three effective fitting parameters: the average size of stem cell-containing clones $\bar{n}(t)$ and progenitor-only clones $\bar{n}_P(t)$, which translate to the exponential decay rate of the large and small clones size distribution, respectively. The third parameter

$$p_0 \equiv f_S \frac{(2r-1)}{r}$$

is a composite function of the fraction f_S of induced cells belonging to the stem cell compartment and the probability r that a stem cell divides symmetrically.

Before proceeding to the model fits to the clonal data, we noted that there was a preponderance of clones containing just a single cell (see Supplementary Table 4). Moreover, the fraction of single cell clones was much larger than what would be expected from the labelling of a purely proliferative population, or by our two-compartment model (1), where such events would originate from the small fraction of labelled progenitor cells that have yet to divide during the chase period. The over-abundance of single cell clones could indicate the existence of a sub-population of long-lived quiescent or non-dividing tumour cells or, alternatively, may correspond to fragments of larger clones that could not be classified correctly by the clone segmentation strategy. Therefore, to circumvent this uncertainty, we first focused attention on the statistical ensemble of clones containing 2 cells or more, which must have been rooted in a proliferative cell at the time of induction. In this case, the resulting renormalised clone size distribution takes the form

$$T'_{n>1}(t) = \frac{T_n(t)}{1 - s(t)}, \tag{13}$$

where $T_n(t)$ is given by Eq. (10) and $s(t) = T_0(t) + T_1(t)$.

To fit the predicted clone size distribution to the experimental data, we followed a two-step procedure. We first used the analytical approximation, Eq. (13), to identify the parameter values, and then used stochastic simulation of the two-compartment model (1) to validate the predictions. Specifically, we first used the predicted long-term size dependence of the complementary cumulative distribution function (CDF)

$$C_{n>1}(t) = 1 - \sum_{m=2}^{n} T'_{m>1}(t) \overset{n \gg 1}{\simeq} \frac{f_S}{1 - s(t)} \frac{2r - 1}{r} e^{-n/\bar{n}(t)} = \frac{p_0}{1 - s(t)} e^{-n/\bar{n}(t)} \tag{14}$$

to determine the decay parameter $\bar{n}(t)$ and then, from the $n = 0$ intercept, obtained the ratio $p_0/(1 - s(t))$ (see Supplementary Table 4). Then, to obtain the exponential fit to the smaller clone size dependence, we subtracted the exponential fit of the large clone size dependence,

$$C'_n(t) = C_n(t) - \frac{p_0}{1 - s(t)} e^{-n/\bar{n}(t)} \propto \exp(-n/\bar{n}_P(t)), \tag{15}$$

from which we could fit the second decay parameter $\bar{n}_P$ (see Supplementary Table 4). Note that, to perform the fit of Eq. (14), we had to identify the clone size threshold at which the clone sizes were best described by the long-term exponential decay. This threshold was found for each sample and colour independently through numerical optimisation, by finding the threshold that minimised the standard error of the fit between the empirical and theoretical CDFs (see "clone size threshold" column in Supplementary Table 4). The values of the three fit parameters $\bar{n}$, $\bar{n}_P$ and $p_0/(1-s(t))$ are shown in Supplementary Table 4.

To obtain the model parameters, we needed to extract the parameter p_0, which required an estimate of $s(t) = T_0(t) + T_1(t)$. In principle, the value of T_0, the probability that a clone was altogether lost during the chase period, is not accessible. However, in the context of the two-compartment model, Eq. (1), there is no clone loss, from which it follows that $T_0(t) = 0$. Moreover, as the clone size distribution at small clone sizes is dominated by contributions from the progenitor cell compartment, we could infer an approximate value of $T_1(t)$ as the $n = 1$ intersect of the small clone size decay, i.e., $C'_1(t)$, such that $1 - s(t) \simeq 1 - T_1(t) \approx C'_1(t)$. From this result, we could estimate the corresponding values of the parameter p_0.

Based on the three parameters, $\bar{n}$, $\bar{n}_P$ and p_0, we then tried to infer the original parameters of the two-component model (1). From $\bar{n}_P$, we could estimate the expansion rate of the progenitor cell population through the relation $\Delta_P = (\ln \bar{n}_P)/t_0$, where t_0 is the given chase time (Supplementary Table 4). However, with the remaining model parameters, Δ_S, r and f_S, with only two fit parameters, $\bar{n}$ and p_0, there is a degeneracy of fits. This degeneracy becomes clear in the limit $\Delta_S \gg \Delta_P$ when, using Eq. (8), it follows that

$$\bar{n}(t) \simeq \left(\frac{r}{2r-1}\right)^2 e^{\Delta_S t} = \left(\frac{f_S}{p_0}\right)^2 e^{\Delta_S t}.$$

We thus have just two relations,

$$\Delta_S = \frac{1}{t_0} \ln\left[\left(\frac{p_0}{f_S}\right)^2 \bar{n}(t_0)\right] \quad \text{and} \quad r = \frac{f_S}{2f_S - p_0}, \tag{16}$$

constraining three adjustable parameters, Δ_S, r and $f_S \geq p_0$. This translates to a line of degeneracy, where we can continuously adjust values of Δ_S, r and f_S along a trajectory in the parameter space without changing the distribution of clone sizes, e.g., a reduction in r can be compensated by an increase in Δ_S, etc. Even when relaxing the condition $\Delta_S \gg \Delta_P$, a similar degeneracy of fit parameters exists both within the approximation scheme and in the exact solution of the model.

Therefore, operationally, to estimate roughly the three remaining model parameters, we made use of the analytical approximation above, setting $r = 0.75$ to break the degeneracy. A different choice of r would translate to a different estimate of the parameters Δ_S and f_S (see Extended Data Fig. 4f); but the overall scale would be similar. Specifically, from

Eqs. (16), we could estimate the stem cell expansion rate Δ_S. Additionally, from the definition $p_0 = f_S \frac{(2r-1)}{r}$, we could estimate labelled stem cell fraction f_S (Supplementary Table 5). In Supplementary Table 5, we summarise the associated fit parameters for each sample and colour, which show that the inferred stem cell expansion rate Δ_S is some 2-3 times that of progenitors, Δ_P. These parameter values result in estimates for the tumour expansion given by Eq. (11) (using the estimated value of f_S as defined above) that are consistent with the empirical measurement (see Supplementary Tables 4 and 5). Estimates of the stem cell fraction, f_S were variable, ranging from 10-32% (with a mean $\pm$ SD of $21 \pm 7\%$), and larger than the reported size of the pre-EMT NC-like cell population from the scRNA-seq data of $6 \pm 3\%$. However, we should note that, since progenitor cell-derived clones could, in principle, be lost if there is cell death, the inferred value of f_S may not equate to the true stem cell fraction at induction. Moreover, the induction of tumour cell types may not be representative of the respective tissue fractions. Finally, our analysis is insensitive the existence and abundance of non-dividing tumour cells, which would be included in the scRNA-seq analysis.

Sensitivity to different choices of the duplication probability r

To break the fit degeneracy implied by Eqs. (16), we made an arbitrary choice of the stem cell duplication, setting $r = 0.75$. Different choices of r alter the respective estimates of the effective stem cell cycling rate and expansion rates, σ and Δ_S, as well as the stem cell fraction, f_S. To assess the sensitivity of these parameters on r, the estimates were recalculated for a wide range of r values (see Extended Data Fig. 4f). As expected, increasing the value of r led to higher estimate of Δ_S and a lower estimate of both σ and f_S, with the variations becoming more pronounced as r approached the singular value $r = 1/2$ (at which the stem cell compartment becomes perfectly renewing). For example, setting $r = 0.9$ led to an average variation of Δ_S, σ and f_S of 16%, 27% and 25%, respectively, relative to the reference values at $r = 0.75$. Consistent with the degeneracy of the model, numerical simulations of the full stochastic model led to equally good fits to the data when considering different values of r.

Stochastic simulation

To explore the precise dynamics of the two-component model (1), and validate the analytical approximation defined above, we turned to stochastic simulations. Specifically, we made use of a stochastic, Gillespie-type [62], simulation that considers the proliferative dynamic of individual cells. Given an initial number of stem cells $S(t=0)$ and progenitors $P(t=0)$, at time $t=0$, the algorithm operates as follows:

1. Compute the propensity functions $w_{s1} = \sigma r S(t)$, $w_{s2} = \sigma(1-r)S(t)$ and $w_p = \Delta_P P(t)$, and define the overall effective rate $w = w_{s1} + w_{s2} + w_p$.

2. Generate a uniformly distributed random number $r_1 \in (0,1]$, and calculate the time of the next event $t+\tau$, where $\tau = -\frac{1}{w}\ln r_1$

3. Decide which of the sub-processes takes place in the time interval τ, for which a new uniformly distributed random number $r_2 \in (0,1]$ is generated. Based on the propensity functions, the number of stem cells S and progenitors P at time $t+\tau$ is obtained according to

$$[S(t+\tau), P(t+\tau)] = \begin{cases} [S(t)+1, P(t)] & r_2 \leq w_{s1}/w \\ [S(t)-1, P(t)+1] & w_{s1}/w < r_2 \leq (w_{s1}+w_{s2})/w \\ [S(t), P(t)+1] & (w_{s1}+w_{s2})/w < r_2 \leq 1 \end{cases}$$

4. The previous steps are repeated while $t < t_{\max}$, where $t_{\max}$ is the chase time of the clonal assay.

Each individual simulation describes the stochastic evolution of a single clone, where the total clone size corresponds to $P(t_{\max}) + S(t_{\max})$ after the chase time $t_{\max}$. To compare the predictions of the model with the data and analytical theory, for each dataset, we considered the statistical distribution from a clonal ensemble of size equal to that analysed in the experimental condition without considering single cell clones, as discussed in the Model fits section 3 (see Supplementary Table 4). Then, to obtain estimates for the size of statistical fluctuations for a given set of parameters, we repeated this procedure over $N = 10^4$ such trials, computing the predicted CDF for each of them. To account for different fractions of the induced proliferative cell population, we initialise a fraction f_S of clones with initial condition $S(0) = 1$ and $P(0) = 0$, and the remaining fraction $(1 - f_S)$ with initial condition

$S(0) = 0$ and $P(0) = 1$. Finally, to compare the simulation with the experimental data, we considered the distribution of clones comprised of 2 cells or more, removing clones of size unity. The results are shown in Extended data Fig. 4g-k, showing the average CDF and standard deviation from $N = 10^4$ trials. A two-sample Kolmogorov-Smirnov (KS) test was performed to compare the empirical and simulated distributions of clone sizes (see Extended data Fig. 4g-k). This shows that, with the exception of the YFP$^+$ clone of sample 1 (right panel, Extended data Fig. 4g), the distributions of clone sizes obtained from our two-compartment model are statistically indistinguishable from the empirical distributions, validating the analytical approach. The KS-test rejected the null hypothesis that the YFP$^+$ clones of sample 1 belong to the same distribution of sizes as the simulated clones (p-value < 0.05). However, we note that this dataset had the smallest sample size, with only 35 clones (ignoring singlets, see Supplementary Table 4), so that random fluctuations may affect the result of the statistical test.

4 Clonal organisation around the vasculature

To provide further support to the stem-progenitor cell model proposed, we performed short-term clonal labelling (4- and 10-days chase) and stained thick sections with the vasculature marker CD31 (Methods). This allowed us to segment both the clones and the vasculature, and to classify clones according to whether they were attached or detached from the vasculature, using a minimum distance threshold of 20 μm. This revealed that indeed those clones that are attached to the vasculature (nearest distance < 20 microns) were significantly larger in size compared to those detached (nearest distance ≥ 20 microns), with the difference becoming more pronounced at the 10 day time point (see Extended data Fig. 7c-d). This behaviour was consistent with a faster expansion of the stem cell-like population localized near the vasculature and corroborates our interpretation of the bi-exponential distribution of clone sizes observed in the long-term clone data, where positional information on the proximity to vasculature was not available.

References

[60] Norman TJ Bailey. *The elements of Stochastic Processes with applications to the natural sciences*. John Wiley & Sons, 1964.

[61] Allon M Klein and Benjamin D Simons. Universal patterns of stem cell fate in cycling adult tissues. *Development*, 138(15):3103–3111, 2011.

[62] Radek Erban, Jonathan Chapman, and Philip Maini. A practical guide to stochastic simulations of reaction-diffusion processes. *arXiv preprint arXiv:0704.1908*, 2007.