Neutral competition within a long-lived population of symmetrically dividing cells shapes the clonal composition of cerebral organoids

Florian G. Pflug[1*], Simon Haendeler[1], Christopher Esk[2], Dominik Lindenhofer[2], Jürgen A. Knoblich[2], Arndt von Haeseler[1,3]

[1] Center for Integrative Bioinformatics Vienna (CIBIV), Max Perutz Laboratories, University of Vienna and Medical University of Vienna, Vienna Bio Center (VBC), Vienna, Austria
[2] Institute of Molecular Biotechnology of the Austrian Academy of Science (IMBA), Vienna Bio Center (VBC), Vienna, Austria
[3] Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Vienna, Austria
*Correspondence: florian.pflug@univie.ac.at

Summary

Cerebral organoids model the development of the human brain and have become an indispensable tool for studying neural development and neuro-developmental diseases. Comprehensive whole-organoid lineage tracing has revealed the number of progeny arising from each initial stem cell to be highly diverse, with lineage sizes ranging from one to more than 20,000 cells. This variability exceeds what can be explained by existing stochastic models of corticogenesis, which indicates that an additional source of stochasticity must exist. We propose the quantitative SAN model in which this additional source of stochasticity is neutral competition within a long-lived population of symmetrically dividing cells. In this model, the eventual size of a lineage is determined by its survival time within this population of symmetrically dividing cells, which due to neutral competition varies widely between individual lineages. We demonstrate the SAN model to explain the experimentally observed variability of lineage sizes and use it to derive a formula that captures the quantitative relationship between survival time and lineage size. Finally, we show that our model implies the existence of a mechanism which keeps the size of the population of symmetrically diving cells approximately constants, and that it enables this mechanism to be probed experimentally.

Introduction

The development and maintenance of the tissues and organs comprising complex organisms rely on sophisticated genetic programs to coordinate the differentiation of cells in both space and time. In many cases, this “program” does not consist of fully deterministic decision chains but instead contains stochastic components; examples include the development of the cortex (Klingler and Jabaudon, 2020; Llorca et al., 2019) and stem cell homeostasis in intestinal crypts (Snippert et al., 2010).

During cortical development, neurons are produced (directly or indirectly) by progenitor cells in the ventricular zone called radial glial cells (RGCs). In mice, the neuronal output of individual RGCs was observed to vary by about one to two orders of magnitude between seemingly identical progenitors, which suggests a stochastic model of cortical neurogenesis (Llorca et al., 2019). Cerebral organoids grown from human stem cells (Lancaster et al., 2017) show even stronger variability of offspring numbers; comprehensive whole-organoid lineage tracing data shows the sizes of individual lineages arising from each ancestral stem cells to vary over up to four to five orders of magnitude (Esk et al., 2020). For RGCs, an alternative explanation of the apparent stochasticity of their number of offspring are hidden variables (Zechner et al., 2020) like transcriptional state within the seemingly homogenous population of progenitors. But in organoids, we expect the pool of ancestral stem cells to be homogenous, and thus conclude that lineage sizes vary predominantly due to stochastic effects. While RGCs output may vary more widely in humans than in mice, varying RGC output alone still cannot account for lineage sizes varying over 4 to 5 orders of magnitude in human cerebral organoids. There is thus likely an additional source of stochasticity in organoid development beyond the stochastic model of neurogenesis proposed by Llorca et al (2019).

In this study, we propose the source of this additional stochasticity to be neutral competition within a long-lived population of roughly 10,000 symmetrically dividing stem cells (S-cells). Neutral competition between stem cells has previously been shown to shape the clonal composition of tissues in homeostasis (Snippert et al., 2010; Corominas-Murtra et al., 2020), and to accurately predict the time until monoclonality (the time until all but a single lineage has died out). We show that in growing tissue like cerebral organoids, neutral competition does not lead to eventual monoclonality. Instead, the tissue records the changing clonal composition of its stem cell population, which causes the sizes of individual lineages to grow increasingly diverse over time. To quantify this effect and its dependence on the size of the S-cell population, we introduce the stochastic SAN model and show that

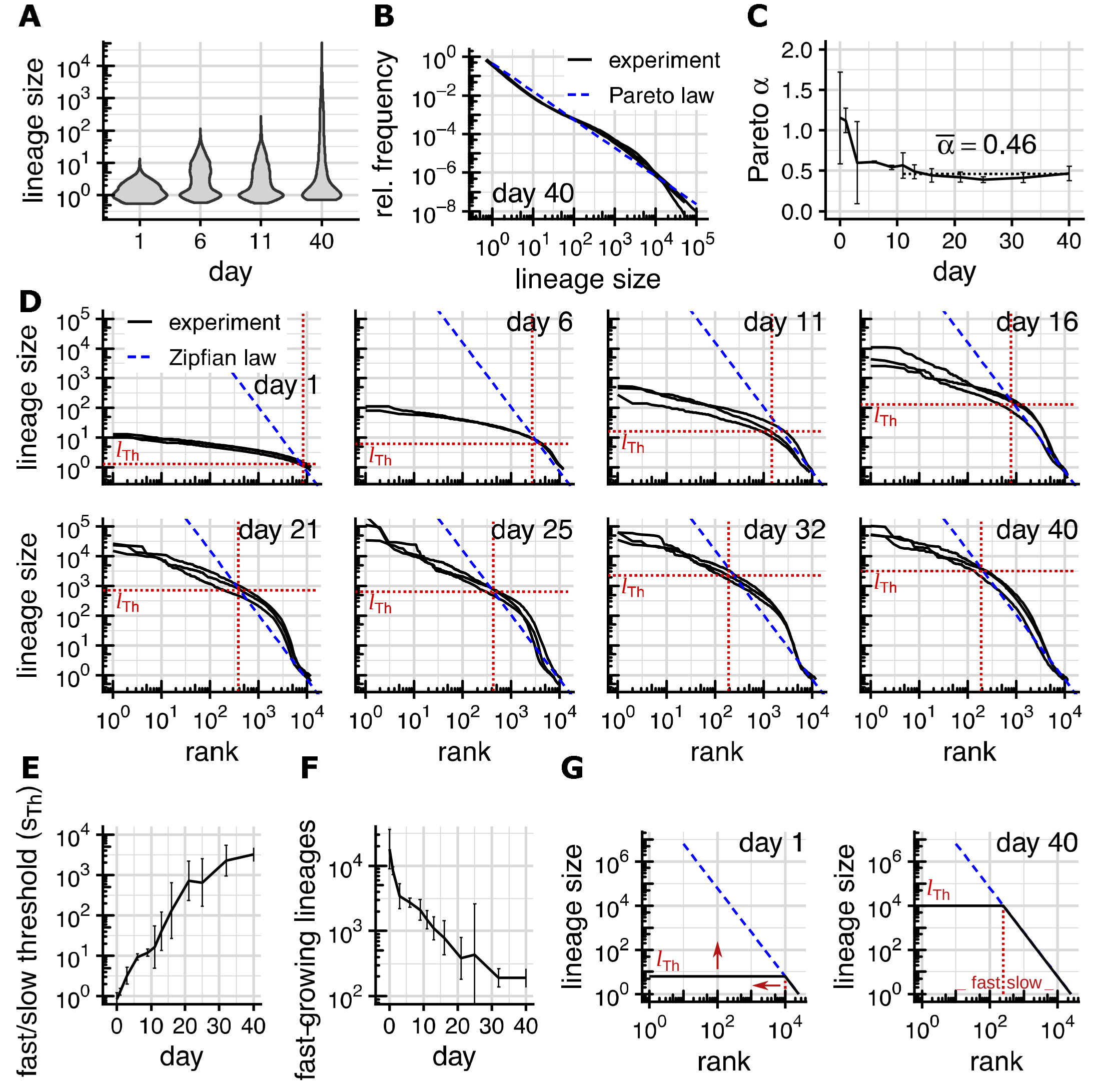

Figure 1. Lineages switch from fast to slow growth**. (experimental data from Esk et al., 2020, error bars and shaded areas show two standard deviations across the three replicates)* ***(A) *Observed lineage size distributions at different time points (only replicate 1 shown, others are similar).* ***(B)*** *Relative frequencies of different lineage sizes on day 40 vs. Pareto power law with* equality index . ***(C)*** *Convergence of Pareto equality index to α = 0.46 (dotted line, represents the average from day 11 onward).* ***(D)*** *Observed rank-size distributions, threshold size s_{Th} (horizontal dotted line) and threshold rank (vertical dotted line). Threshold size and rank mark the truncation point of the Zipfian power law $r^{-1/\alpha}$ separating fast- and slow-growing lineages.* ***(E)*** *Size threshold s_{Th} and* ***(F)*** *number of fast-growing lineages (threshold rank l_{Th}) over time.* ***(G)*** *Over time, the size of fast-growing lineages grows but their number drops, causing the lineage size distribution to approach a power law.*

neutral competition within its S-cell population suffices to explain the observed variation of lineage sizes over four to five orders of magnitude.

Results

Empirical lineage size distribution

In the experiment conducted by Esk *et al.* (Esk et al., 2020), cerebral organoids were grown from roughly 24,000 stem cells, genetically identical except for a distinct genetic barcode in each cell serving as a *lineage identifier* (LID). To determine the contribution of each initial stem cell to organoids of different ages, organoids were subjected to amplicon high-throughput sequencing. The sequencing reads (after filtering and error-correction) corresponding to each LID were counted, and the per-LID read counts normalized to an approximate number of cells comprising each lineage (see *experimental procedures* for details).

The resulting *lineage size distribution* (figure 1A) shows, as expected, small and equally sized lineages for organoids harvested at day 1 (lineages sizes around 1 cell). The distribution grows more uneven until day 11 (up to 30 cells/lineage) and extends over 4 to 5 orders of magnitude (up to 100,000 cells/lineage) after 40 days.

A common mathematical model for distributions extending over multiple orders of magnitude are so-called (Pareto) power laws where the frequency of objects of size l or larger is proportional to $l^{-\alpha}$. Parameter α is called the (Pareto) equality index because it determines how even (large α) or uneven (small α) object sizes are distributed. In double-logarithmic frequency vs. size plots, power laws appear as straight lines with slope α, which we find matches the lineage size distribution on day 40 well for $\alpha \approx 0.46$ (figure 1B). We remark that $\alpha \approx 0.46$ represents a small equality index (i.e. diverse lineage sizes); in applications of Pareto distributions values of α often lie between 1 and 2.

While the unevenness of the lineage size distribution grows considerably between days 11 and 40 (figure 1A), the equality index stays close to $\alpha \approx 0.46$ from day 11 onwards (figure 1C). The equality index thus fails to capture the large increase in non-uniformity of lineage sizes between days 11 and 40.

Truncated Zipfian rank-size distribution

To describe the evolution of lineage sizes over time, we thus instead rank lineages by size (largest lineage first) and plot the resulting *rank-size distributions* (figure 1D). For lineage sizes governed by a Pareto law with index α, the rank-size distribution would be expected to be governed by a Zipfian power law, meaning lineage sizes should decrease proportional to $r^{-1/\alpha}$ with increasing rank r (Adamic and Huberman, 2002).

Instead, we observe *truncated* Zipfian laws in which lineages sizes obey a Zipfian law ($\alpha \approx 0.46$) only up to a certain threshold size l_{Th} above which lineages are multiple orders of magnitudes smaller and more uniform than the Zipfian law would predict (figure 1D). The threshold size l_{Th} grows more than 1,000-fold (figure 1E) over 40 days, while the ratio between threshold size and largest lineage size grows only by a factor of 2 (from 8.5 to 21; figure 1D); lineages above the threshold therefore grow roughly uniformly. Lineages below the threshold, in contrast, show no overall shift towards larger lineages sizes over time, indicating that growth has mostly ceased for these lineages.

Lineages switch from fast to slow growth

The size threshold l_{Th} thus partitions lineages according to their growth regime into *fast-growing* and *slow/non-growing*. Of the (on average) 10,851 lineages that contribute to the final organoid 8,389 lineages fall into the fast-growing category on day 1; but on day 11 their number has dropped to 1.496, and on day 40 only 191 (about 2%) fast-growing lineages remain (figure 1F). Lineages thus start out fast-growing, and one by one switch to a regime of slow/no growth as time progresses. The later that switch occurs for a particular lineage, the bigger it has become before its growth ceases, leading to larger and larger lineages in the slow-growing regime and consequently to l_{Th} increasing as time progresses. In this coarse approximation, lineages are assumed to have the same size as long as they are fast-growing (figure 1G); experimentally we observe a spread of 1.5 orders of magnitude within the sizes of fast-growing lineages versus a spread of 3.5 orders of magnitude within the slow-growing regime (figure 1D).

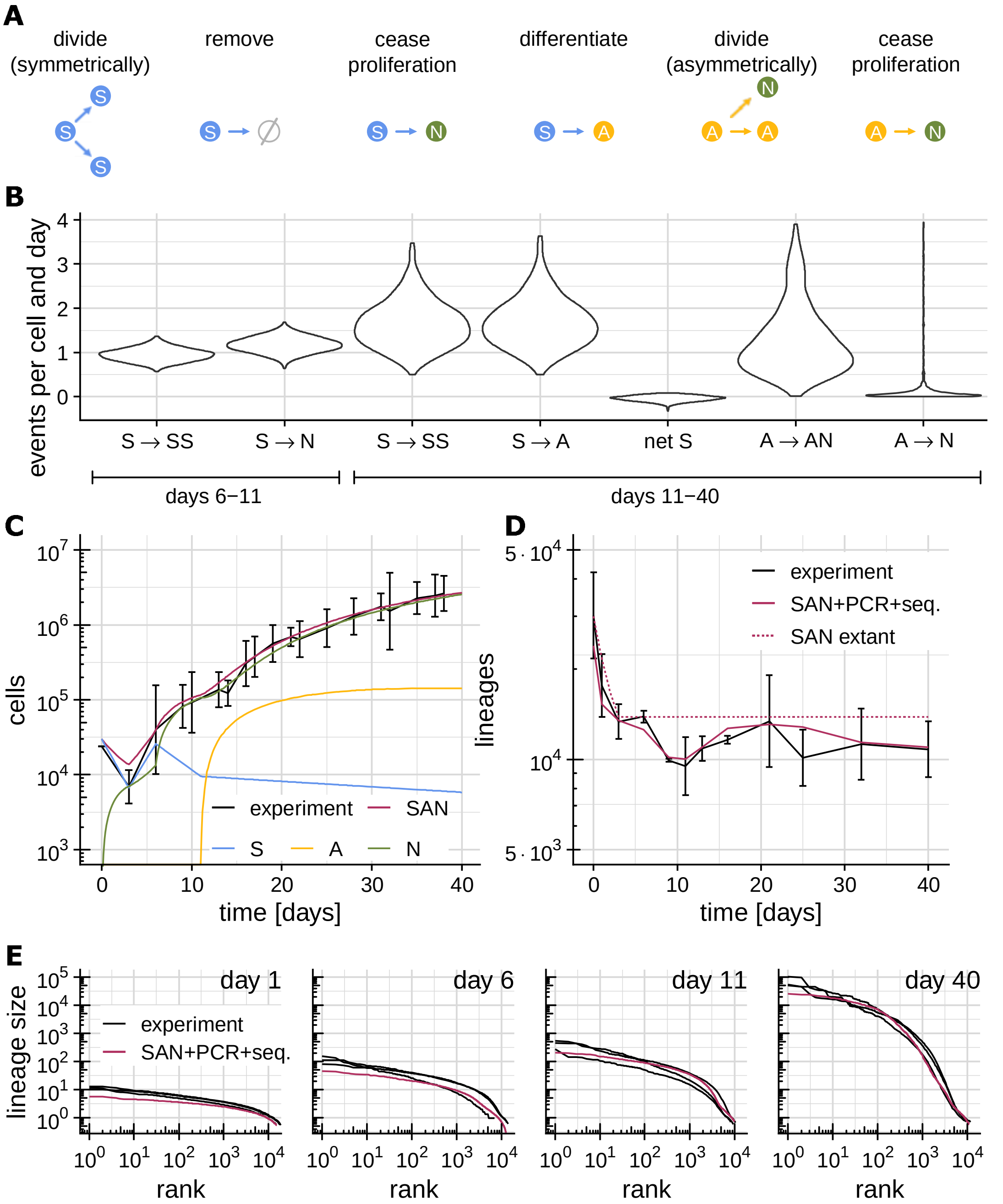

Figure 2. The SAN model. *(experimental data from Esk et al., 2020)* ***(A)*** *Division and differentiation events that S-, A- and N-cells can undergo (see table 1 for the corresponding rates).* ***(B)*** *Posterior distribution for the estimated rates of division and differentiation events for days 6-11 and 11-40 (events not shown are assumed to have rate 0).* ***(C)*** *Total organoid size (number of cells) observed experimentally (black) and predicted by the SAN model (red), and the predicted number of S- (blue), A- (yellow) and N-cells (green).* ***(D)*** *Experimentally observed number of lineages (black), predicted number of extant lineages (dotted red) and predicted number of observed lineages (red; based on the SAN model and a model of NGS-based lineage tracing).* ***(E)*** *Rank-size distributions observed experimentally (black) and predicted by the SAN model plus a model of NGS-*

This proposed lineage-specific switching from fast to slow growth can also quantitatively reproduce the observed truncated Zipfian with $\alpha = 0.46$, with one mathematically simple example being fast-growing lineages growing exponentially with rate γ and the number of fast-growing lineages declining exponentially with rate $\sigma = \alpha\gamma$. But while this simple example assumes an unspecified biological mechanism behind the lineage-specific growth regime switches, we show in the following that no such mechanism is in fact necessary. Instead, we show that such growth regime switches emerge naturally from a cellular model of organoid growth.

SAN model

In the SAN model of organoid growth we distinguish between three types of cells based on the proliferation behavior they exhibit (figure 2A). Cells are either *symmetrically* dividing (S-cells), *asymmetrically* dividing (A-cells) or *non-dividing* (N-cells). In this model, S-cells have the ability to self-renew indefinitely through symmetric division and can thus be considered stem cells. They form the initial cell population of an organoid, and apart from dividing symmetrically they differentiate into either A- or N-cells or are removed permanently. A-cells are cells that have committed to a differentiation trajectory and produce N-cells through asymmetric division, while N-cells do not further divide. We emphasize that S, A and N refer solely to a cell's proliferation behavior, not its functional cell type.

All these division and differentiation events occur randomly and independently for each cell with specific time-dependent rates (table 1); from a single-lineage perspective, the SAN model is thus stochastic in nature. Any difference between the trajectories of the lineages arising from different ancestral cells is thus assumed to be purely the result of random chance, not of cell fate decisions or spatial configuration. From a whole-organoid perspective, on the other hand, the SAN model is deterministic, because random effects average out over the roughly 10,000 lineages comprising an organoid.

Division and differentiation rates

To find the rates of cell division and differentiation, we split the organoid development into four time intervals (days 0-3, 3-6, 6-11 and 11-40) according to the main phases of the protocol of Lancaster *et al.* (Lancaster et al., 2017). Until day 6, formation of embryoid bodies (EBs) is still ongoing, and organoid development thus does not reflect development *in vivo*. For these time intervals we manually chose rates of S-cell division (S → S S), removal (S → ∅) and death (S → N; dead cells present in the EB are still counted by NGS-based lineage tracing) for which predicted and observed numbers of cells, lineages, and lineage sizes match (table 1). After day 6, EB formation is complete, and no further cells are removed from the organoid. Until day 11 S-cells are then assumed to either divide symmetrically (S → S S) or cease proliferation (S → N), but to not produce A-cells yet. After embedding the organoids into Matrigel droplets on day 11, organoid growth enters the asymmetric division phase where S-cells are assumed to multiply (S → S S) and to differentiate into A-cells (S → A), which then produce N-cells through asymmetric division (A → A N) before they eventually cease to proliferate (A → N).

From day 6 onwards organoid development reflects development in vivo and we hence desired to identify the range of likely rates for each event in addition to a single most-likely value. We thus adopted a Bayesian model comprising log-normally distributed measurement inaccuracies on top of the SAN model, and used Markov chain Monte Carlo (MCMC) sampling to find 1,000 likely rate combinations and their (posterior) probabilities (figure 2B). To arrive at a single set of most-likely values for the rates to be estimated, we then computed MAP (maximum a-posteriori) estimates (table 1) from this posterior distribution.

Both the posterior distribution (figure 2B) and the MAP estimates (table 1) show the *net* rate of S-cell proliferation (the rate with which the S-cell population grows or shrinks, i.e. the difference between the rates of S → S S and S → A) to lie close to zero. From this, we conclude that the size of the S-cell population changes only slowly from day 11 onwards. The posterior distributions of the individual rates are, on the other hand, much broader. While the MAP estimates are thus arguably the single most likely set of rates, other combinations of rates are possible as well.

Model validation

For the MAP rate estimates (table 1), the predicted organoid sizes between day 0 and 40 agree well with the experimentally determined number of cells (figure 2C). Similarly, the lineage size distribution predicted by the SAN model matches the observed lineage size distribution both in the original data of Esk et al. (figure 2E) as well as in independent replicate experiments (figure S1). In particular, the predictions show the same truncated Zipfian distributions as the experimental data recapitulate the spread over 4 – 5 orders of magnitude. The SAN model predicts the number of *extant* lineages (lineages containing at least one S-, A or N-cell) to drop to about ≈13,700 on day 3 where it then remains. This drop in the number of extant lineages is caused by lineages that do not make it into the organoid during EB formation. While the predicted number of remaining lineages slightly exceeds the experimental observation (≈10,900 on day 40 on average), the numbers match closely once we account for non-observed lineages due to the stochastic nature of sequencing (figure 2D).

A-cell output

According to the SAN model (table 1) a single A-cell has two options, either to divide asymmetrically (probability $r_{A\to AN}/(r_{A\to AN}+r_{A\to N}) = 91\%$) or to cease proliferation (probability $r_{A\to N}/(r_{A\to AN}+r_{A\to N}) = 9\%$). The likely range of additional N-cells produced over the lifetime of an A-cell is thus 0 to 30 (95% quantile), with an average of $r_{A\to AN}/r_{A\to N} = 10$. This slightly exceeds (about 3x) the observed stochasticity of the neuronal output of RGCs in mice (Llorca et al., 2019) and is consistent with a previously observed increase of intermediate progenitor divisions in humans.

Predicted S-cell population size

The MAP rate estimates (table 1) predict that organoids contain ≈9,500 S-cells on day 11 and still ≈5,800 S-cells on day 40. To take the inherent ambiguity of the MAP estimate due to the broadness of the posterior distribution into account, we

day	S → S S	S → ∅	S → N	S → A	A → A N	A → N	phase
0-3	-	0.35	0.15	-	-	-	EB formation
3-6	0.6	-	0.15	-	-	-	EB formation
6-11	0.94	-	1.14	-	-	-	neural induction
11-40	1.68	-	-	1.69	0.71	0.07	asymmetric division

***Table 1. Division and conversion rates in the SAN model**. Rates specify the number of expected events per cell and day*

computed the posterior distribution of these population sizes (figure 3A). We find that the total S-cell population size on day 11 is well-defined up to a factor of at most 2 around 10,000 cells. On day 40, the estimates are more dispersed, owning to the large cumulative effect that rates have over 30 days; yet while the exact population size is difficult to estimate, the finding that organoids contain a significant number of S-cells on day 40 is robust.

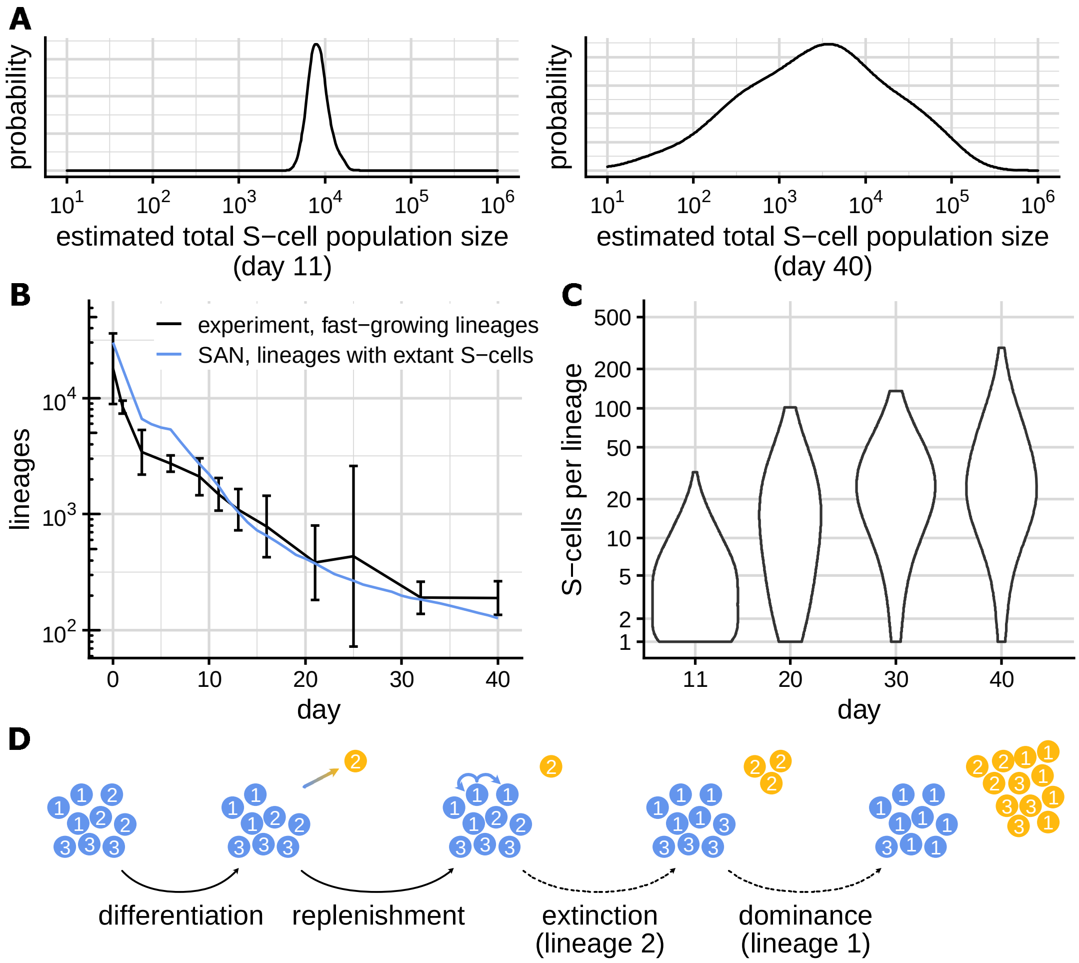

Figure 3. The S-cell population. (A) *Likely total S-cell population sizes on (left) day 11 (left) and (right) day 40 (posterior distribution obtained with MCMC, see text).* ***(B)*** *Number of fast-growing lineages (black; from figure 1E) number of lineages with extant S-cells as predicted by the SAN model (blue).* ***(C)*** *Distribution of S-cells per lineages (considering only lineage with extant S-cells) on days 11, 20, 30, 40 predicted by the SAN model.* ***(D)*** *Neutral competition among lineages (lineage identifiers 1,2,3) within the S-cell population. The clonal composition of the S-cell distribution changes through stochastic differentiation and replenishment events, causing some lineage to eventually lose all S-cells and others to become dominant.*

Fast-growing lineages contain S-cells

While the total size of the S-cell population changes only slowly, its clonal composition changes rapidly. From the ≈13,700 lineages comprising the organoid from day 3 forward, ≈1,700 lineages still contain S-cells on day 11, and until day 40 that number has dropped to ≈100 (figure 3B). This drop in the number of lineages with extant S-cells is offset by an increase in the number of S-cells each of these lineages contains (figure 3C; average grows from ≈5 cells/lineage on day 11 to ≈36 cells/lineage on day 40).

The number of lineages with extant S-cells matches the number of lineages classified as fast-growing by our fast-slow model well (figure 3B). This highlights S-

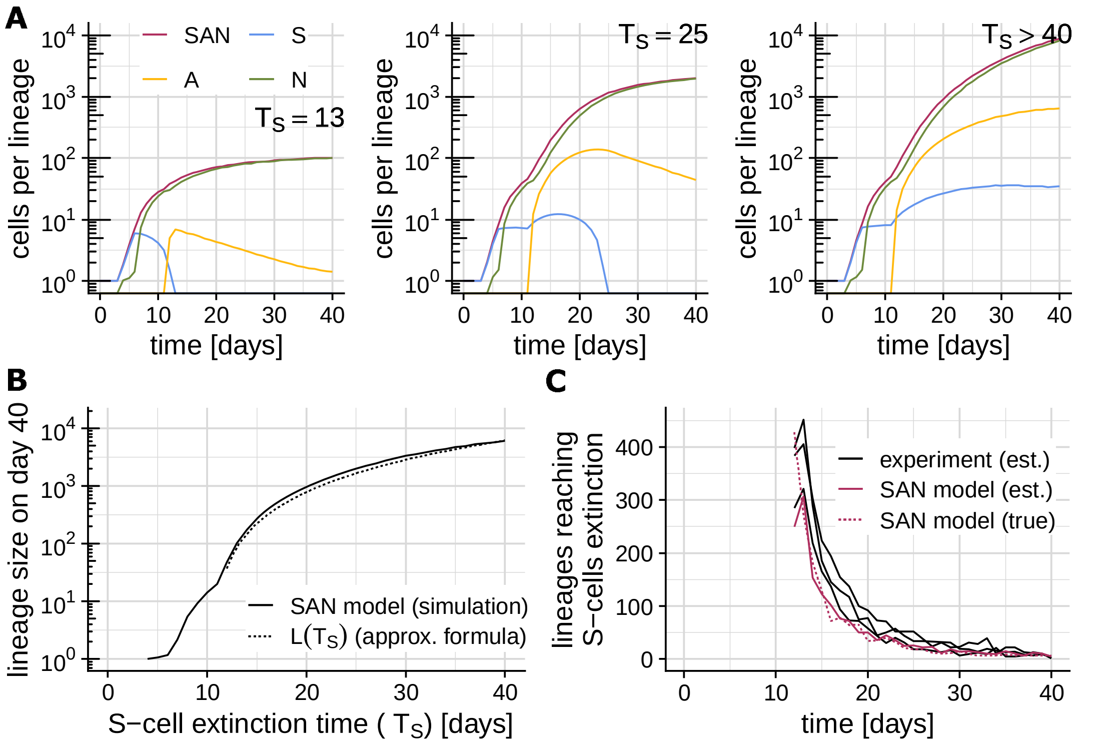

Figure 4. Lineage-specific S-cell extinction time determines linage size. *(Error bars show two standard deviations across the three replicates of Esk et al, shaded areas show the range between the 2.5% and 97.5% quantile across 5 billion simulations).* ***(A)*** *Lineage-specific growth trajectories under the SAN model stratified by the lineage's S-cell extinction time T_S. Plot show the most likely total lineage size (red), and number of S- (blue), A- (yellow) and N- (green) cells comprising the lineage.* ***(B)*** *S-cell extinction time T_S vs. final lineage size on day 40. Plot shows simulation results (solid) and the analytical approximation $L(T_S)$ (dotted).* ***(C)*** *Recovering S-cell extinction times from lineage sizes on day 40. Plot shows the number of lineages reaching S-cell extinction on each day estimated using $L(T_S)$ from experimental data (black) and simulated data (red), and the true number of such lineages according to the SAN model.*

cells as being the main driver of lineage growth; as stated above a single differentiating S-cells eventually on average produces 10 additional N-cells, and once a lineage contains no more S-cells its growth will thus slow down and eventually cease.

Neutral competition shapes S-cell clonal composition

Over time, not only does the average number of S-cells found within lineages with extant S-cells grow, but so does the spread between the lineages' S-cells counts (from about 1 - 30 S-cells per lineage on day 11 to about 1 – 300 S-cells per lineage on day 40). The clonal composition of an organoid's S-cell population thus grows more and more non-uniform over time. Under the SAN model, this change is the result of *neutral competition* (figure 3D) amongst S-cells, a term introduced to describe the population dynamics of stem cells within intestinal crypts (Snippert et al., 2010).

Qualitatively, the dynamics of an organoid's S-cell population under neutral competition mimic the population-genetic Moran model (Moran, 1958) in which individuals (cells in our case) carrying different neutral alleles (lineage identifiers in our case) are randomly removed (differentiate) and are replaced (through symmetric division) by offspring of another randomly selected individual. Once the last S-cell of a particular lineage has differentiated, the lineage cannot reappear within the organoid's S-cell population. The observed disappearance (figure 3B) of lineages from the organoids S-cell population is thus a result of more S-cells differentiating than dividing due to random chance. Similarly, the observed growth of the remaining lineages (figure 3C) results from more symmetric divisions than differentiations, again due to random chance.

Using the SAN model, we now study the effects of neutral competition between S-cells on the clonal composition quantitatively.

Lineage-specific S-cell extinction times determine final lineage sizes

Under the population-genetic Moran model, alleles eventually either disappear from a population or become fixed. Tissue homeostasis driven by a stem cell population under neutral competition likewise leads to eventual monoclonality, i.e. to all extant cells being eventually derived from a single ancestral stem cell. In growing neural tissue like cerebral organoids however, the lack of constant cell turn-over restricts eventual monoclonality to S-cells. The clonal composition of the N-cell population instead records the evolution of the S-cell's clonal composition over time; lineages whose last S-cell was lost later and/or which contained more S-cells will contribute more N-cells than lineages which die out quickly from the S-cell population.

To study the effects of S-cell extinction on lineage sizes quantitatively, we stratified simulated lineage growth trajectories according to their *S-cell extinction time* (T_S; the time at which a particular lineage loses the last S-cell). Lineages whose S-cell population goes extinct at day $T_S = 13$ (figure 4A left) respectively day $T_S = 25$ (figure 4A middle) show diminished growth and a declining number of A-cells after losing their S-cells at time T_S. In contrast, lineages whose S-cell population survives past day 40 (figure 4A right) grow considerably faster and reach a considerably larger size. Comparing the variations in lineage sizes on day 40 between S-cell extinction time strata shows the variation due to T_S to dominate the variations within each stratum (figure 4B). Thus, while other random factors have some influence, their influence on a lineage's sizes on day 40 is negligible compared to the time the lineage loses its last S-cell.

Mathematical analysis of the SAN model yields the approximate expression

(Eq. 1)

$$L(\Delta T_S) = \left(\frac{1}{3} s_0 \Delta T_S + \frac{r_{S\to SS}}{2\sqrt{6}} \Delta T_S^2\right) r_{S\to A} \left(1 + \frac{r_{A\to AN}}{r_{A\to N}}\right)$$

for the final lineage size of a lineage comprising s_0 S-cells on day 11 ($s_0 \approx 5$ for rates in table 1; and we assume $r_{S\to SS} \approx r_{S\to A}$) and whose S-cell population goes extinct ΔT_S days later. We note that *final lineage size* here does not refer to the size on day 40 (or any other particular point in time), but rather to the eventual size a lineage will have reached when its growth ceases. While this does not exactly match our simulation setup (we only simulate up to day 40) the approximate final linage sizes $L(\Delta T_S)$ still matches the simulation results well (figure 4B).

Recovering S-cell extinction times from final lineage sizes

By solving the equation $L(\Delta T_S) = L_i$, the time at which a lineage lost its last S-cell can be estimated from the final size (L_i) of that lineage. To gauge the reliability of this approach, we applied it to a simulate lineage size distribution for day 40, and found that it recovers the number of lineages that reached S-cell extinction on a particular day well (figure 4C). When applied to the experimentally observed lineage sizes on day 40, the estimated number of lineages reaching S-cell extinction lies close to the SAN model prediction, but slightly exceeds it up to about day 30.

Emergence of a Zipfian law

If A- and N-cells are disregarded, the SAN model is equivalent to the well-studied birth-death process, and in particular, the distribution of the S-cell extinction time ΔT_S is known (Feller, 1939). By translating this distribution via $L(\Delta T_S)$ into the corresponding distribution of lineage sizes, the (approximate) distribution of final lineage sizes (i.e. of lineages which have ceased growth) can be found. If we consider only sufficiently large S-cell extinction times ΔT_S, the probability of ΔT_S is (approximately) proportional to $1/\Delta {T_S}^2$ and translation into lineage sizes via $L(\Delta T_S)$ yields a Zipfian law with $\alpha = 0.5$. This theoretical prediction matches the empirical observation that lineage sizes approach a Zipfian law with $\alpha \approx 0.46$.

Discussion

We have empirically observed lineages in cerebral organoids to initially grow fast and roughly uniformly until some lineage-specific stopping time at which growth slows down significantly or ceases altogether; and have found the size of slow or non-growing lineages to follow a Zipfian power law with exponent $-1/\alpha$, $\alpha = 0.46$. While the destructive nature of NGS-based lineage tracing prevents us from directly observing lineages as they switch their growth regime, alternative hypotheses would necessarily involve either very early fate decisions, or lineage-specific proliferation rates to explain the large diversity of observed lineage sizes. Both alternative models seem unlikely given that organoids are grown from a homogenous population of stem cells.

To study the cause of the apparently random and lineage-wide switch of growth regime we introduced the cellular *SAN* model. This model accurately recapitulates all experimental data and shows that observed lineage growth dynamics to emerge from neutral competition within a proposal long-lived population of roughly 10,000 symmetrically dividing stem cells (S-cells). Under the SAN model the apparently lineage-wide switch of growth regime occurs despite the lack of either direct or indirect (e.g., through spatial colocation) lineage-wide events. Instead, growth of a lineage slows down and eventually ceases as the result of the lineage vanishing from the S-cell population through neutral competition; lineage survival time within the organoid's S-cell population is thus the major determinant of lineage size.

The relationship between a lineage's survival times within the organoids S-cell population and the size it eventually attains can be expressed by a formula. Inverting this formula allows the history of the organoids S-cell population that was recorded within its clonal composition to be read; doing so we found for days 11-30 a slight excess of lineages reaching S-cell extinction in the experimental data compared to the SAN model. We hypothesize that this might point to gradual reduction of division and differentiation rates in organoids; Since the SAN model assumes constant rates between days 11 and 40, a gradual reduction of rates would cause the model to appear to fall behind at first, and then to catch up once the true rates have fallen below the model's rates.

While we found that we cannot estimate the rates of most division and differentiation events in the SAN model precisely, we could robustly determine the rates of S-cell division and differentiation to be almost identical. This implies that the population of symmetrically dividing cells in cerebral organoids is long-lived, and in particular that organoids still contain a population of symmetrically dividing cells after 40 days.

Furthermore, the similarity of the S-cell division and differentiation rates implies the existence of a mechanism that controls the S-cell population size by linking S-cell differentiation and subsequent replenishment through symmetric divisions. Yet that link must be stochastic in nature; if S-cells simply divided asymmetrically to produce A-cells, or if after symmetric division exactly one offspring always differentiated, no neutral competition between S-cells would occur and the observed large variability of lineage size would remain unexplained. In the terminology of Simons and Clevers (2011), the mechanism must thus be of the *population asymmetric* type.

Given the similarity between the population-level link of S-cell division and differentiation and the dynamics within stem cell niches in intestinal crypts (Snippert et al., 2010), we conjecture that similar structures located within proliferation centers called *neural rosettes* (Esk et al., 2020). might be responsible for balancing division and differentiation of S-cells.

To study the mechanism controlling S-cell population size in more detail, it needs to be probed experimentally by perturbing organoids at specific points in time and observing their response. If a fraction of cells is killed, different mechanism would respond differently: A mechanism that relies on spatial constraints (i.e. stem cells being pushed out of a niche) would be expected to show a reduced rate of differentiations until the population has recovered. A regulatory mechanism which more directly links S-cell division to differentiation would respond differently; there we might expect the S-cell population to never reach its original size, but to instead increase its overall cell turn-over to make up for lost S-cells.

Since the SAN model accurately predicts the lineage sizes observed for cerebral organoids grown from wildtype cells, it is also useful both when planning organoid-based perturbation screens, and when analyzing the resulting data. During the planning phase the model makes it possible to judge the effect of proliferation phenotypes on final lineage size, and thus to estimate the statistical power of different screen designs. During statistical analysis of screening data, the model provides a baseline (null model), against which the sizes of (genetically) perturbed lineages can be compared.

To facilitate the adoption of the SAN model, we offer an implementation of the model both as an interactive online service (URL to be determined) as well as a R package (http://github.com/Cibiv/SANjar).

Funding

This work was supported by the Austrian Science Fund (FWF) project number F78.

Methods

Total organoid sizes

For days 0 through 21, organoid sizes were measured using fluorescence-activated cell sorting (FACS). For days 11 through 40, organoid volumes were estimated from microscopy images, and translated into cell counts using the average number of cells per volume for days 11 through 21 where both FACS and volume measurements were available.

NGS data processing

The lineage tracing data of Esk et al. (2020) was obtained from GEO (accession GSE151383, supplementary file GSE151383_LT47.tsv.gz), and organoids “H9-day06-03” and “H9-day09-01” removed as outliers. Based on the assumption that in all samples the most common lineage size is 1 cell, we located the mode of the log-transformed read count distribution for every sample and used it to normalize relative lineage sizes (reads) to absolute cell counts. The validity of the underlaying assumption is confirmed by the good agreement the sum of absolute lineage sizes and the FACS and area-derived estimates of total organoid size.

Pareto index and fast-slow threshold estimation

For each organoid, we used the observed lineage sizes $l_1, \dots, l_n$ to estimate the Pareto equality index α and minimal lineage size m with the maximum-likelihood estimator

$$\widehat{m} = \min_i l_i, \qquad \widehat{\alpha} = n\left(\sum_i \log\frac{l_i}{\widehat{m}}\right)^{-1},$$

and computed the steady-state average $\bar{\alpha}$ from the alpha estimates of all organoids sequenced on day 11 or later. To find the fast-slow threshold l_{Th} for a particular organoid, we first found intersect d^{Pareto} such that the Pareto-induced rank-size powerlaw $\log_{10} L^{\mathrm{Pareto}}(r) = -\bar{\alpha}^{-1}\log_{10} r + d^{\mathrm{Pareto}}$ fits the size of the smallest observed lineage, and determined the smallest rank R for which the actual lineage size $l_{(r)}$ matches or exceeds the power law $L^{\mathrm{Pareto}}(r)$. We then fit a separate log-log-linear model $\log_{10} L^{\mathrm{Large}}(r) = k\log_{10} r + d^{\mathrm{Large}}$ to lineages with ranks $1, \dots, \sqrt{R}$ (which we assume are surely not governed by the Pareto law), and set l_{Th} to the size at which the two laws intersect (meaning $l_{\mathrm{Th}} = L^{\mathrm{Pareto}}(r) = L^{\mathrm{Large}}(r)$.

SAN model simulation

The total number of S-, A- and N-cells that an organoid is predicted to comprise at time t is computed based on the deterministic SAN model (with rates $r_{S\to SS}$, $r_{S\to\emptyset}$, $r_{S\to\mathrm{A}}$, $r_{S\to\mathrm{N}}$, $r_{A\to\mathrm{AN}}$, $r_{A\to\mathrm{N}}$ of these events occurring per cell and per day). The deterministic SAN model is described by the ordinary differential equations (ODE),

$$s(0) = 30{,}000, \qquad a(0) = 0, \qquad n(0) = 0,$$
$$\dot{s} = (r_{S\to SS} - r_{S\to\emptyset} - r_{S\to A} - r_{S\to N})s,$$
$$\dot{a} = r_{S\to A}s - r_{A\to N}a,$$
$$\dot{n} = r_{S\to N}s + (r_{A\to AN} + r_{A\to N})a,$$

which can be solved analytically (for the time-homogenous case) and is then evaluated separately for each time interval within which rates are constant (The initial number $s(0)$ of S-cells is set to 30,000 instead of 24,000 to account for a slight excess in the number of observed lineages on day 0, likely due to a combination of multiple labelling and sequencing artefacts).

To find the predicted lineage size distribution at time t, the stochastic SAN model is simulated independently for each of the 30,000 lineages in an organoid. The simulation proceeds in discrete time steps Δt, which are chosen small enough to make the probability of a single cell undergoing two events negligible ($< 10^{-3}$). Given the numbers $S_i(t)$, $A_i(t)$, $N_i(t)$ of S-, A-, N-cells comprising lineage i at time t, the number of cells Δ_e undergoing event e is chosen from a Poisson distribution. Specifically,

$$\Delta_{S\to SS} \sim \mathrm{Poisson}(r_{S\to SS}S_i(t)\Delta t), \quad \Delta_{S\to\emptyset} \sim \mathrm{Poisson}(r_{S\to\emptyset}S_i(t)\Delta t),$$
$$\Delta_{S\to A} \sim \mathrm{Poisson}(r_{S\to N}S_i(t)\Delta t), \quad \Delta_{S\to N} \sim \mathrm{Poisson}(r_{S\to N}S_i(t)\Delta t),$$
$$\Delta_{A\to AN} \sim \mathrm{Poisson}(r_{A\to AN}A_i(t)\Delta t), \quad \Delta_{A\to N} \sim \mathrm{Poisson}(r_{A\to N}A_i(t)\Delta t),$$

and the number of S-, A-, N-cells at time $t + \Delta t$ is then set to be

$$s_i(t + \Delta t) = s_i(t) + \Delta_{S\to SS} - \Delta_{S\to\emptyset} - \Delta_{S\to A} - \Delta_{S\to N,}$$
$$a_i(t + \Delta t) = a_i(t) + \Delta_{S\to A} - \Delta_{A\to N,}$$
$$n_i(t + \Delta t) = n_i(t) + \Delta_{S\to N} + \Delta_{A\to \mathrm{AN}} + \Delta_{A\to N.}$$

Finally, the lineage size distribution $l_1, \ldots, l_{30{,}000}$ at time t is found by summing up the number of S-, A- and N-cells, $l_i(t) = s_i(t) + a_i(t) + n_i(t)$.

Technical noise simulation

The effect of PCR amplification and sequencing on the observed lineage sizes was simulated using a stochastic model of PCR amplification and sequencing (Pflug and von Haeseler, 2018) with parameters *PCR efficiency* and average *reads per molecule* (in our case per *lineage*). For every sampling time t, we simulated one read count (normalized to one read per cell on average) per lineage; parameters were *PCR efficiency* 35% (estimated from the day 0 data) and average *reads per lineage* $Wl_i/\sum_i l_i$ for a linage comprising l_i cells (W is the median experimental library size for time t). The simulated read counts where then normalized to cells by division by the average number of reads per cell ($W/\sum_i l_i$).

SAN rate estimation

For days 0-3 and 3-6, rates which replicate the experimental data well were found by trial and error. For the remaining 6 biologically relevant rates (of S → S S and S → N between 6 and 11, and S → S S, S → A, A → A N and A → N between days 11 and 40) we computed the posterior distribution given experimentally observed total organoid sizes $\hat{c}^{(dayt)}$ (on days $t \in \mathcal{D}$ = {0, 3, 6, 9, 10, 13, 14, 16, 17, 19, 21, 22, 25, 28, 31, 32, 35, 37, 38}) and ranked lineage sizes $\hat{l}_{(r)}^{(\text{day } 11)}$, $\hat{l}_{(r)}^{(\text{day } 40)}$ (on days 11 and 40, for ranks $r \in \mathcal{R}$ = {1, 2, 5, 10, 15, 25, 40, 60, 100, 150, 250, 400, 600, 1000, 1500, 2500, 4000, 6000, 10000, 15000, 25000}). To account for biological differences between replicates we assumed that experimental observations are log-normally distributed around the SAN model predictions $c^{(dayt)}$ and $l_{(r)}^{(\text{day } 11)}$, $l_{(r)}^{(\text{day } 40)}$; the likelihood of the rate vector θ (comprising the 6 rates mentioned above) given the experimental data is thus

$$l(\theta) = -\frac{1}{2}\sum_{t\in\mathcal{D}}\left(\frac{\mu\left[\hat{c}^{(\text{day } t)}\right] - c^{(\text{day } t)}}{\sigma\left[\hat{c}^{(\text{day } t)}\right]}\right)^2 - \frac{1}{2}\sum_{t\in\{11,40\}}\sum_{r\in\mathcal{R}}\left(\frac{\mu\left[\hat{l}_{(r)}^{(\text{day } t)}\right] - l_{(r)}^{(\text{day } t)}}{\sigma\left[\hat{l}_{(r)}^{(\text{day } t)}\right]}\right)^2$$

where $\mu[\ldots]$ and $\sigma[\ldots]$ denote the mean respectively standard deviation across biological replicates. Rates were restricted to lie between 0 and 4 and *a priori* assumed to be equally probable; the posterior probability of θ is thus proportional to $l(\theta)$. To find this posterior distribution, we sampled 1,000 random rate vectors according to their likelihoods by simulating 1,000 Markov chains using pseudo-marginal Metropolis-Hastings Markov chain Monte Carlo sampling (Beaumont 2003;

Andrieu & Roberts, 2009; Warne et al., 2020). We then computed the maximal mode of the (joint) posterior distribution with the mean-shift algorithm to obtain the MAP estimates (table 1).

Mathematical Analysis

If we consider only S-cells, the SAN model corresponds to the well-known birth-death process (Feller, 1939). We consider the diffusion approximation of this process and restrict our mathematical treatment to day 11 and later where the rates of symmetric division ($r_{S\to SS}$; birth) and of differentiation ($r_{S\to A}$; amounts to death since we consider only S-cells) are similar enough to be considered identical ($r_{S\to SS} = r_{S\to A} = \lambda/2$). The number of S-cells within a lineage at time t (where $t = 0$ represents day 11) is then governed by the stochastic differential equation (SDE)

$$ds(t) = \sqrt{\lambda s(t)}dW(t).$$

Using Onsager-Machlup theory (Onsager & Machlup, 1953; Dürr & Bach 1978) we find the most probably trajectory of a linage that contains s_0 cells at $t = 0$ and loses its last S-cell ΔT_S days later,

$$s_{\text{ext}}(t \mid s_0, \Delta T_S) = s_0\left(1 - \frac{t}{\Delta T_S}\right)\left(1 + \rho\frac{t}{\Delta T_S}\right) \quad \text{where} \quad \rho = \frac{\Delta T_S}{s_0}\lambda\sqrt{\frac{3}{8}} - 1.$$

On average, a lineage grows by $\lambda/2$ A-cells per S-cell and per day, and over its lifetime every A-cell will eventually produce $r_{A\to AN}/r_{A\to N}$ additional N-cells through asymmetric division. Eventually, a lineage that starts out with s_0 S-cells and loses its last S-cell ΔT_S days later will thus approximately grow to size

$$L(\Delta T_S) = \frac{\lambda}{2}\left(1 + \frac{r_{A\to AN}}{r_{A\to N}}\right)\int_0^{\Delta T_S} s_{\text{ext}}(t \mid s_0, \Delta T_S)dt.$$

Integration of this expression yields Eq. (1).

References

Adamic, L. A., and Huberman, B. (2002). Zipf's law and the Internet. Glottometrics *3*, 143–150.

Andrieu, C., and Roberts, G. (2009). The Pseudo-Marginal Approach for Efficient Monte Carlo Computations. The Annals of Statistics 37 (2).

Corominas-Murtra, B., Scheele, C.L.G.J., Kishi, K., Ellenbroek, S.I.J., Simons, B.D., van Rheenen, J., and Hannezo, E. (2020). Stem cell lineage survival as a noisy competition for niche access. Proceedings of the National Academy of Sciences *117*, 16969–16975.

Beaumont, M. (2003). Estimation of Population Growth or Decline in Genetically Monitored Populations. Genetics 164 (3), 1139-1160.

Dürr, D., and Bach, A. (1978). The Onsager-Machlup function as Lagrangian for the most probable path of a diffusion process. Communications in Mathematical Physics *60*, 153–170.

Esk, C., Lindenhofer, D., Haendeler, S., Wester, R.A., Pflug, F., Schroeder, B., Bagley, J.A., Elling, U., Zuber, J., von Haeseler, A., et al. (2020). A human tissue screen identifies a regulator of ER secretion as a brain-size determinant. Science *370*, 935–941.

Feller, W. (1939). Die Grundlagen der Volterraschen Theorie des Kampfes ums Dasein in wahrscheinlichkeitstheoretischer Behandlung. Acta Biotheoretica *5*, 11–40.

Klingler, E., and Jabaudon, D. (2020). Do progenitors play dice? ELife *9*, e54042.

Lancaster, M.A., Corsini, N.S., Wolfinger, S., Gustafson, E.H., Phillips, A.W., Burkard, T.R., Otani, T., Livesey, F.J., and Knoblich, J.A. (2017). Guided self-organization and cortical plate formation in human brain organoids. Nature Biotechnology *35*, 659–666.

Llorca, A., Ciceri, G., Beattie, R., Wong, F.K., Diana, G., Serafeimidou-Pouliou, E., Fernández-Otero, M., Streicher, C., Arnold, S.J., Meyer, M., et al. (2019). A stochastic framework of neurogenesis underlies the assembly of neocortical cytoarchitecture. ELife *8*, e51381.

Moran, P.A.P. (1958). Random processes in genetics. Mathematical Proceedings of the Cambridge Philosophical Society *54*, 60–71.

Onsager, L., and Machlup, S. (1953). Fluctuations and irreversible processes. Physical Review *91*, 1505–1512.

Pflug, F.G. and von Haeseler, A. (2018). TRUmiCount: Correctly counting molecules using unique molecular identifiers. *Bioinformatics* 34-18, 3137–3144.

Simons, B. D. and Clevers, H. (2011). Strategies for homeostatic stem cell self-renewal in adult tissues. Cell *145*, 851–862.

Snippert, H.J., van der Flier, L.G., Sato, T., van Es, J.H., van den Born, M., Kroon-Veenboer, C., Barker, N., Klein, A.M., van Rheenen, J., Simons, B.D., et al. (2010). Intestinal Crypt Homeostasis Results from Neutral Competition between Symmetrically Dividing Lgr5 Stem Cells. Cell *143*, 134–144.

Warne, D. J., Baker, R. E., and Simpson M. J. (2020). A Practical Guide to Pseudo-Marginal Methods for Computational Inference in Systems Biology. Journal of Theoretical Biology 496.

Zechner, C., Nerli, E., and Norden, C. (2020). Stochasticity and determinism in cell fate decisions. Development *147*, dev181495.

Supplemental Information

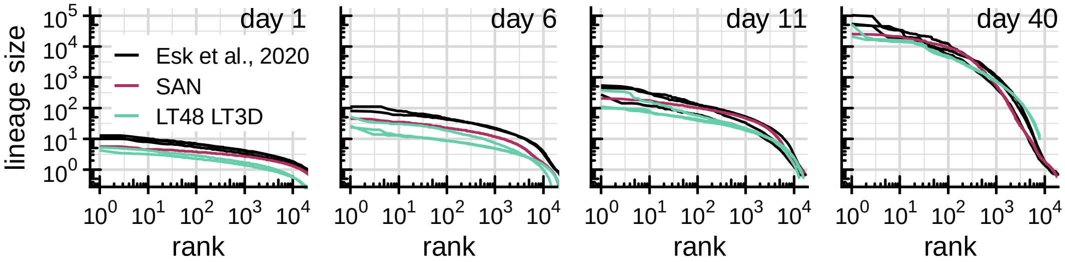

Figure S1. Replicate experiments. *Replicate experiments based on the same organoid protocol show similar lineage size distributions as the data from Esk et al. (2020). Ranks of the Esk et al. data and SAN model predictions were scaled to account for an 1.7-fold increase in the number of detected lineages in the replicate experiments.*