PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

# Neutral competition explains the clonal composition of neural organoids

**Florian G. Pflug**[1,2]*, **Simon Haendeler**[2,3], **Christopher Esk**[4,5], **Dominik Lindenhofer**[4,6], **Jürgen A. Knoblich**[4,7], **Arndt von Haeseler**[2,8]

**1** Biological Complexity Unit, Okinawa Institute of Science and Technology Graduate University (OIST), Onna, Okinawa, Japan, **2** Center for Integrative Bioinformatics Vienna (CIBIV), Max Perutz Labs, University of Vienna and Medical University of Vienna, Vienna BioCenter (VBC), Vienna, Austria, **3** Vienna Biocenter (VBC) PhD Program, a Doctoral School of the University of Vienna and the Medical University of Vienna, Vienna, Austria, **4** Institute of Molecular Biotechnology of the Austrian Academy of Science (IMBA), Vienna BioCenter (VBC), Vienna, Austria, **5** Institute of Molecular Biology, University of Innsbruck, Innsbruck, Austria, **6** Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany, **7** Department of Neurology, Medical University of Vienna, Vienna, Austria, **8** Faculty of Computer Science Bioinformatics and Computational Biology, University of Vienna, Vienna, Austria

* florian.pflug@univie.ac.at

OPEN ACCESS

**Citation:** Pflug FG, Haendeler S, Esk C, Lindenhofer D, Knoblich JA, von Haeseler A (2024) Neutral competition explains the clonal composition of neural organoids. PLoS Comput Biol 20(4): e1012054. https://doi.org/10.1371/journal.pcbi.1012054

**Editor:** Tunahan Cakir, Gebze Technical University, TURKEY

**Received:** October 9, 2023

**Accepted:** April 3, 2024

**Published:** April 22, 2024

**Data Availability Statement:** An R package that contains code and data is available at http://github.com/Cibiv/SANjar. Sequencing data for the additional replicates performed is available in GEO under accession GSE255232.

**Funding:** Work in the laboratory of A.v.H that contributed to this study was supported by the Austrian Science Fund (FWF) Special Research Program SFB-F7811-B. Work in the laboratory of J. A.K that contributed to this study was supported by the Austrian Academy of Sciences, the Austrian

## Abstract

Neural organoids model the development of the human brain and are an indispensable tool for studying neurodevelopment. Whole-organoid lineage tracing has revealed the number of progenies arising from each initial stem cell to be highly diverse, with lineage sizes ranging from one to more than 20,000 cells. This high variability exceeds what can be explained by existing stochastic models of corticogenesis and indicates the existence of an additional source of stochasticity. To explain this variability, we introduce the SAN model which distinguishes Symmetrically diving, Asymmetrically dividing, and Non-proliferating cells. In the SAN model, the additional source of stochasticity is the survival time of a lineage's pool of symmetrically dividing cells. These survival times result from neutral competition within the sub-population of all symmetrically dividing cells. We demonstrate that our model explains the experimentally observed variability of lineage sizes and derive the quantitative relationship between survival time and lineage size. We also show that our model implies the existence of a regulatory mechanism which keeps the size of the symmetrically dividing cell population constant. Our results provide quantitative insight into the clonal composition of neural organoids and how it arises. This is relevant for many applications of neural organoids, and similar processes may occur in other developing tissues both *in vitro* and *in vivo*.

## Author summary

Organoids are tissue that is grown *in vitro*, but which replicates many aspects of a specific organ. Neural organoids grown from around 20,000 human pluripotent stem cells in particular replicate many aspects of the developing human brain, and have become an important tool for studying human neurodevelopment. Lineage-tracing allows the contribution of each ancestral stem cells to the organoid to be measured quantitatively. Surprisingly,

Science Fund (FWF) Special Research Programme F7804-B and EU Horizon 2020 research and innovation programme 874769. S.H. is and D.L. was a member of the VBC doctoral program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** I have read the journal's policy and the authors of this manuscript have the following competing interests: J.A.K. is on the supervisory and scientific advisory board of a:head bio AG (https://aheadbio.com) and is an inventor on several patents relating to cerebral organoids.

these contributions vary between a single and more than 20,000 cells, which affects many applications of neural organoids. In this paper we study how these variations arise. We introduce the so-called SAN model, named after the symmetrically, asymmetrically, and non-dividing cells it distinguishes. This model explains the large variation in lineage sizes as being caused by *neutral competition* between symmetrically dividing cells, i.e. by random chance as cells divide and differentiate. Our model helps to understand the growth of neural organoids quantitatively, and provides way to account for the large difference in lineage sizes in applications of neural organoids. Neutral competition as the cause of large lineage size differences may also be relevant beyond neural organoids, and possibly in the development of tissue *in vivo*.

## 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 stem cell homeostasis in intestinal crypts [1] and recently the development of the cortex [2,3].

During cortical development, neurons are produced (directly or indirectly) by progenitor cells in the ventricular zone called radial glial cells (RGCs). In mice, Llorca *et al*. observed the neuronal output (i.e. number of neurons produced) of individual RGCs to vary by about one to two orders of magnitude between seemingly identical progenitors which lead them to suggest a stochastic model of cortical neurogenesis [3]. In human cerebral organoids [4], Esk *et al*. [5] used comprehensive whole-organoid lineage tracing to measure the contribution of each ancestral stem cell to an organoid, and after 40 days of growth found total number of offspring to vary over four to five orders of magnitude between individual stem cells (Fig 1A). This greatly exceeds what can be explained by the stochastic model of corticogenesis of Llorca *et al*. [3] alone, even if differences between the two model systems are considered. We are thus faced with the question: What causes lineage sizes in cerebral organoids to vary over multiple orders of magnitude, and how does the source of these variations fit into what is known about neurogenesis?

In the past, hidden variables like differing transcriptional state within a seemingly homogenous population of progenitors have been suggested as a possible non-stochastic source of varying offspring numbers [6]. For partially developed tissued with defined spatial structure and differing ancestries of cells, the existence of such hidden variables seems plausible. For cerebral organoids grown from a homogenous population of stem cells, however, differences between the internal states of cells extensive enough to explain multiple orders of variations between offspring numbers seem unlikely. Instead, we expect there to be a hitherto unobserved stochastic source of offspring variability.

Here, we show that neutral competition (also called neutral drift) within a long-lived population of roughly 10,000 symmetrically dividing stem cells, which we call S-cells, explains the observed variations in lineage size over multiple orders of magnitude. Neutral competition has previously been shown to shape the clonal composition of tissues in homeostasis [1,7], 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's clonal composition represents a record of the S-cell population's history, and size differences between individual lineages reflect

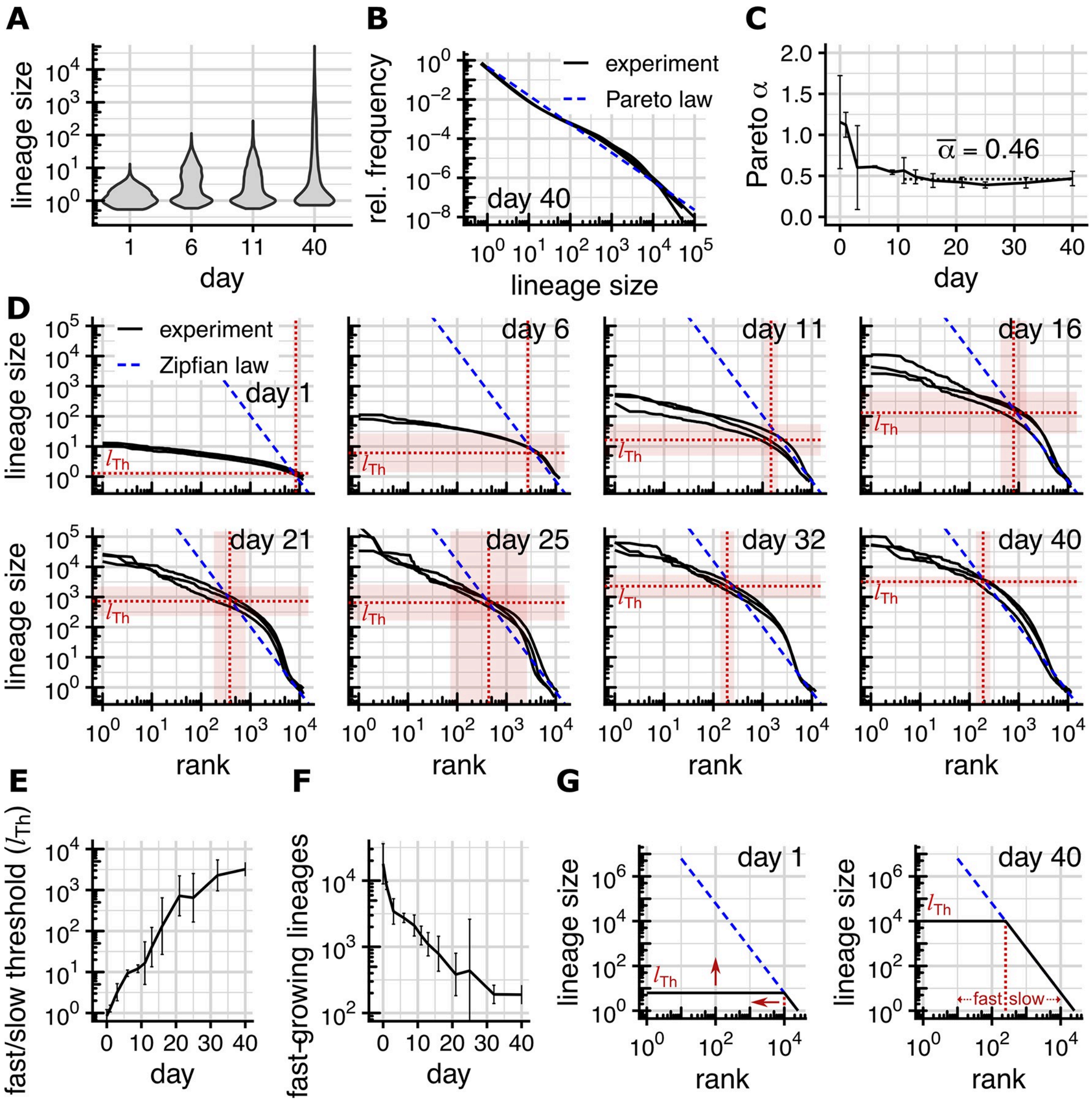

**Fig 1. Lineages switch from fast to slow growth.** (Experimental data from Esk et al., 2020. The error bars (panels C, E, F) and shaded areas (panel D) show ± two standard deviations across the three replicates) **(A)** Observed lineage size distributions at different time points (only replicate 1 shown, others are similar). **(B)** Relative frequencies of different lineage sizes on day 40 vs. Pareto power law with equality index $\bar{\alpha}$. **(C)** Convergence of Pareto equality index to α = 0.46 (dotted line, represents the average from day 11 onward). Per-replicate estimates are listed in S5 Table. **(D)** Observed rank-size distributions, threshold size $l_{\mathrm{Th}}$ (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 $l_{\mathrm{Th}}$ and **(F)** number of fast-growing lineages (threshold rank) over time. Per-replicate estimated are listed in S2 Table **(G)** Over time, the size of fast-growing lineages grows but their number drops, causing the lineage size distribution to approach a power law.

https://doi.org/10.1371/journal.pcbi.1012054.g001

differences between their fates. To quantify this effect and its dependence on the size of the S-cell population, we introduce the stochastic SAN model. In this model, organoids are assumed to comprise three types of cells, S-cells which either divide symmetrically or differentiate into A-cells, A-cells which divide asymmetrically or convert into N-cells, and N-cells which have ceased to divide. These division and differentiation events occur stochastically at specific rates. We estimate these rates, and show that 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.* [5], 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 were normalized to an approximate number of cells comprising each lineage (see *Methods* for details).

The resulting *lineage size distribution* (Fig 1A) shows, as expected, small and equally sized lineages for organoids harvested at day 1 (lineages sizes around 1 cell). The distribution grows more dispersed 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. In power-law distributions, the frequency of objects of size $l$ or larger is proportional to $l^{-\alpha}$, where the parameter $\alpha$ controls the amount of dispersion. The parameter $\alpha$ is called (Pareto) equality index because a small value of $\alpha$ indicates a large dispersion (i.e. diverse sizes), while a large values of $\alpha$ indicates clustered (i.e. similar) sizes. Compared to other commonly used distributions such as the normal distribution, power-law distributions typically represent very dispersed data. In double-logarithmic frequency vs. size plots, power laws appear as straight lines with slope $\alpha$. This fits the lineage size distribution on day 40 well for $\alpha \approx 0.46$ (Fig 1B). We remark that $\alpha \approx 0.46$ represents a small equality index (i.e. very dispersed lineage sizes); in applications of Pareto distributions values of $\alpha$ often lie between 1 and 2 [8].

At early time points, technical/experimental noise dominates over actual variations in lineage size, and we consequently find large sample-to-sample variations of the equality index (Fig 1C) between the three replicates for each time point. Starting with day 11, all samples show an equality index close to $\alpha \approx 0.46$ (Figs 1C and S1). During the same time, however, the dispersion of the linage size distribution grows considerably (Fig 1A).

### Truncated Zipfian rank-size distribution

To understand why the equality index converges while the inequality of lineage sizes grows, we rank lineages by size (largest lineage first) and plot the resulting *rank-size distribution* in a double-logarithmic plot (Fig 1D). In such a plot, lineage sizes governed by a Pareto law with index $\alpha$ would be expected to be governed by a Zipfian power law (i.e. lineage sizes decrease proportional to $r^{-1/\alpha}$ with increasing rank $r$). Because of the double-logarithmic nature of the plot, the rank-size distribution would thus be expected to form a straight line with slope $-1/\alpha$ [9].

We observe that even from days 11 onwards (where we found $\alpha \approx 0.46$), lineage sizes adhere to the Zipfian law only up to a certain lineage size threshold $l_{\mathrm{Th}}$. Lineages larger than the threshold $l_{\mathrm{Th}}$ are multiple orders of magnitudes smaller than the Zipfian law would predict, and more uniform in size (Figs 1D and S1). For most time points (very early ones excluded), the majority of lineages lie below the threshold and thus in the Zipfian regime. The threshold $l_{\mathrm{Th}}$ itself grows rapidly with time, increasing more than 1,000-fold (Fig 1E) over 40 days. At the same time the spread of lineage sizes above the threshold $l_{\mathrm{Th}}$ (i.e. ratio between $l_{\mathrm{Th}}$ and largest lineage size) grows only by a factor of about two (from 8.5 to 21; Fig 1D). We conclude that lineages outside the Zipfian regime grow quickly (since the threshold grows rapidly) and roughly similarly fast (since the spread of linage sizes in this regime increases only slowly). Lineages in the Zipfian regime, in contrast, show no overall shift towards larger lineages sizes over time, indicating that growth has mostly ceased for these lineages.

## Lineages switch growth regime

The threshold $l_{\text{Th}}$ 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 (Fig 1F). Lineages thus start in a fast-growing regime, 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 $l_{\text{Th}}$ increasing as time progresses. Within the fast-growing regime (i.e. lineages sizes above $l_{\text{Th}}$), the difference in size between the largest and smallest lineage is about 1–2 orders of magnitude (Fig 1D). Within the slow-growing regime, the spread between lineage sizes reaches 3.5 orders of magnitude on day 40 (Fig 1D). We thus neglect the differences between lineages sizes within fast-growing lineages to arrive at a simplified representation of this model. In this cartoon version of our data-derived model, at any particular point in time all fast-growing lineages have the same size. When they drop out of the fast-growing regime (at a random point in time) their size arrests (Fig 1G).

This proposed lineage-specific switching from fast to slow growth can also quantitatively reproduce the observed truncated Zipfian with $\alpha = 0.046$. One mathematically simple model are fast-growing lineages which grow exponentially with rate $\gamma$ and drop out of the fast-growing regime with rate $\sigma = \alpha\gamma$ (S1 Supplemental Methods). 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

We now consider a simplified, yet biologically realistic stochastic cellular model of organoid growth we call the SAN model that explains the observed linage size distributions. The model distinguishes between three types of cells based purely on the proliferation behavior they exhibit (Fig 2A). Cells are either *symmetrically* dividing (S-cells), *asymmetrically* dividing (A-cells) or *non-dividing* (N-cells). In our 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 lost, i.e. removed permanently). A-cells have committed to a differentiation trajectory and produce N-cells through asymmetric division, while N-cells do not further divide.

We categorize cells into S, A and N solely based on proliferation behavior, not by functional categories such as for example radial glial cells (RGCs). We furthermore only consider such divisions as *symmetric* that produce offspring cells which are functionally identical to the parent. This precludes divisions which may appear symmetric, but where one of the offspring in fact has an altered cellular state which puts it on differentiation trajectory that leads to eventually asymmetrical division or non-proliferation. While appearing symmetric, such divisions are in the context of the SAN model more appropriately modelled as S to A transitions, since the number of eventual N-cells resulting from such a division is relatively well-determined (yet not necessarily fully deterministic).

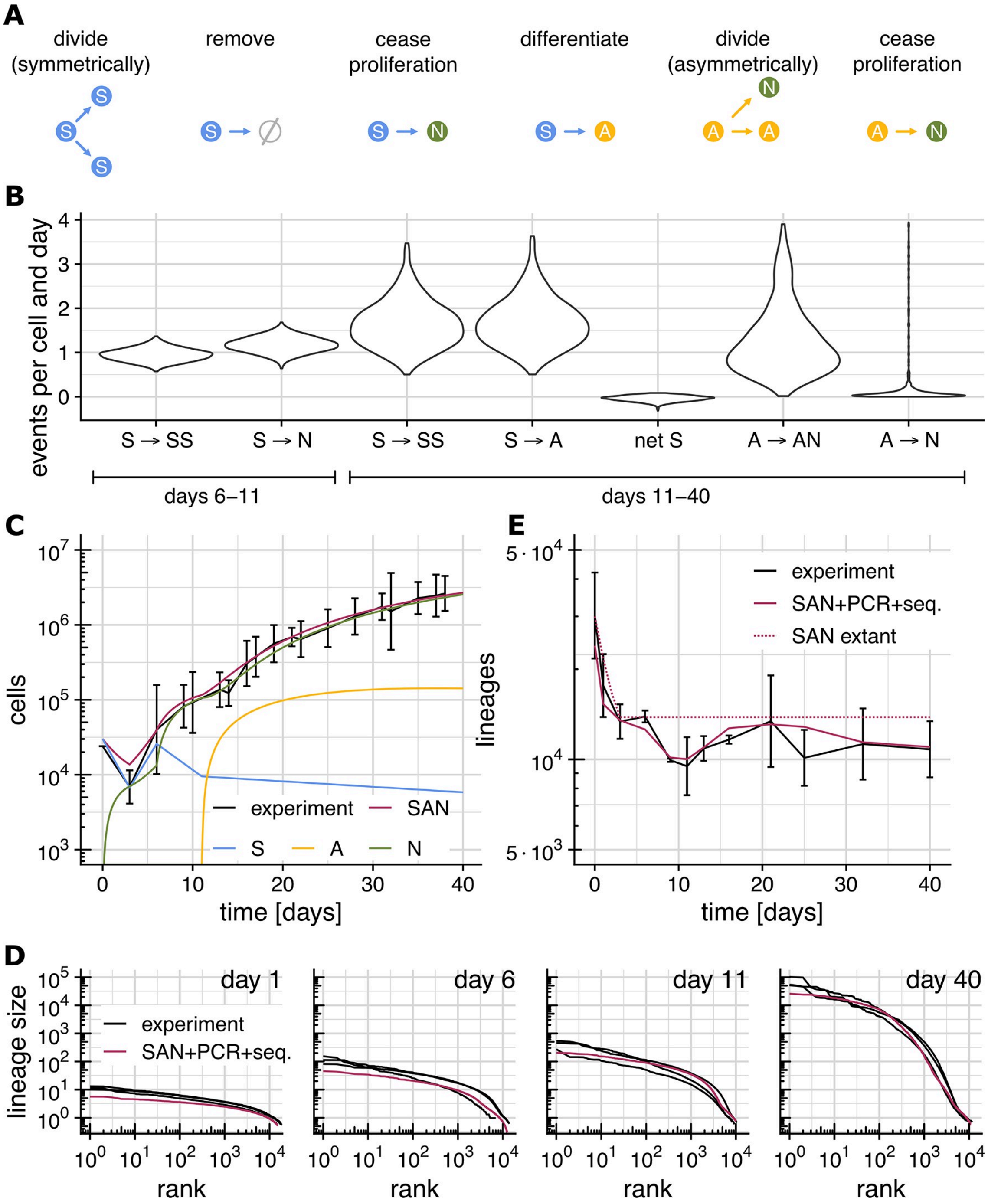

**Fig 2. The SAN model.** (Experimental data from Esk et al., [5]) **(A)** Division and differentiation events that S-, A- and N-cells can undergo (see Table *1* for the corresponding rates). **(B)** Posterior distribution for the estimated rates of division and differentiation events for days 6–11 and 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). Raw data available in S4 Table. **(D)** Rank-size distributions observed experimentally (black) and predicted by the SAN model plus a model of NGS-based lineage tracing (red). **(E)** 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, see methods "Simulation of NGS- based lineage tracing" for details). Number of obs. lineages in each replicate listed in S3 Table.

https://doi.org/10.1371/journal.pcbi.1012054.g002

**Table 1. Division and conversion rates in the SAN model.** Rates specify the number of expected events per cell and day. These rates are the maximum a-posteriori (MAP) estimates, the full posterior distribution is shown in Fig 2B.

| 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 |

https://doi.org/10.1371/journal.pcbi.1012054.t001

In the SAN model, all 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.* [4]. 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) for which predicted and observed numbers of cells, lineages, and lineage sizes match experimental data (Table 1). Here, distinguishing S → ∅ and S → N is necessary since dead cells within the embryoid body (EB) are still counted by sequencing-based lineage tracing, while cells which are not part of the EB are not. After day 6, EB formation is complete, and no further cells are removed from the organoid. Until day 11 S-cells are thus 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). We do not consider direct differentiations of S cells into N cells here since from a perspective of lineage sizes, the occurrence of such events is difficult to distinguish from slightly elevated rates of S → A and A → N. Note that A-cells decide randomly after each asymmetric division whether to continue dividing or to cease proliferation, and the N-cell output per A-cell is thus stochastic.

From day 6 onwards organoid development reflects development in vivo [5]. To obtain a range of likely rates for each event in addition to a single most-likely value, we adopted a Bayesian approach. In a Bayesian setting, one starts with an *a priori* assumption about likely parameter values and updates it based on experimental observations. This yields an *a posteriori* distribution that reflects how likely different parameter combinations are given the experimental evidence. We assumed *a priori* that all division and differentiation rates between 0 and 4 events/day (on average) are equally likely, and that measurement errors follow log-normal distributions. We then sampled 1,000 parameter combinations from the posterior distribution using a Markov Chain Monte Carlo (MCMC) approach (see Methods). To verify that the MCMC algorithm had converged to the true posterior distribution, we verified that the Gelman-Rubin [10] diagnostic lies below 1.2 (S1 Table). Finally, 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 (Fig 2B) and the MAP estimates (Table 1 and S6 Table) show a *net* rate of S-cell proliferation (the difference between the rates of S → S S and S → A) 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 other 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 maximum a-posteriori (MAP) rate estimates (Table 1), the organoid sizes predicted by the SAN model between day 0 and 40 agree well with the experimentally determined number of cells (Fig 2C). Similarly, the lineage size distribution predicted by the SAN model matches the observed lineage size distribution (Figs 2D and S2). In particular, the predictions show the same truncated Zipfian distributions as the experimental data and 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. This predicted number of remaining lineages (≈13,700) slightly exceeds the experimental observation (≈10,900 on day 40 on average). Since sequencing is an inherently stochastic sampling process, this excess of predicted over observed lineages caused by extant lineages which remain unobserved and is therefore to be expected. Once we take the stochastic nature of sequencing into account (see Methods for details) and compute the number of lineages we should expect to observe instead of the number of extant lineages, the numbers match closely (Fig 2E). To confirm the close fit of model and data, we replicated the experiments performed by Esk *et al.* [5] using the published protocol. These replicate experiments agree well with the original data of Esk et al. and show a similarly close fit between model and data (S3 Fig).

## Individual effect of each rate

To clarify the individual effect that each division and differentiation rate has on organoid size and lineage size distribution, we performed a sensitivity analysis where we modified the MAP estimate of each rate (Table 1) by ± two standard deviations of the posterior distribution (Fig 2B and S6 Table). The results show that the rates between days 6 and 11 strongly affect the shape of the lineage size distribution, but not the sizes of the largest lineage (S5 Fig). The large effect of individually modifying the rate of S → S S or S → A (S6 Fig) confirms the strong correlation of these rates, previously observed through the narrow posterior distribution of the net S-cell proliferation rate (Fig 2B).

## A-cell output

According to the MAP rate estimates for the SAN model (Table 1) a single A-cell has two options, either to continue to divide asymmetrically (probability $r_{A\to AN}/(r_{A\to AN}+r_{A\to N})$ = 91%) or to cease proliferation (probability $r_{A\to AN}/(r_{A\to AN}+r_{A\to N})$ = 9%). Since every additional asymmetric division produces an N-cell, the likely range of additional N-cells produced over the lifetime of an A-cell is thus 0 to 30 (95% quantile; $0.91^{30}\approx 0.05$), with an average of $r_{A\to AN}/r_{A\to N}$ = 10. This range of N-cells produced by a single A-cell is comparable to the range of neurons (1–35) produced by a single radial glial cell (RGC) in mice [3].

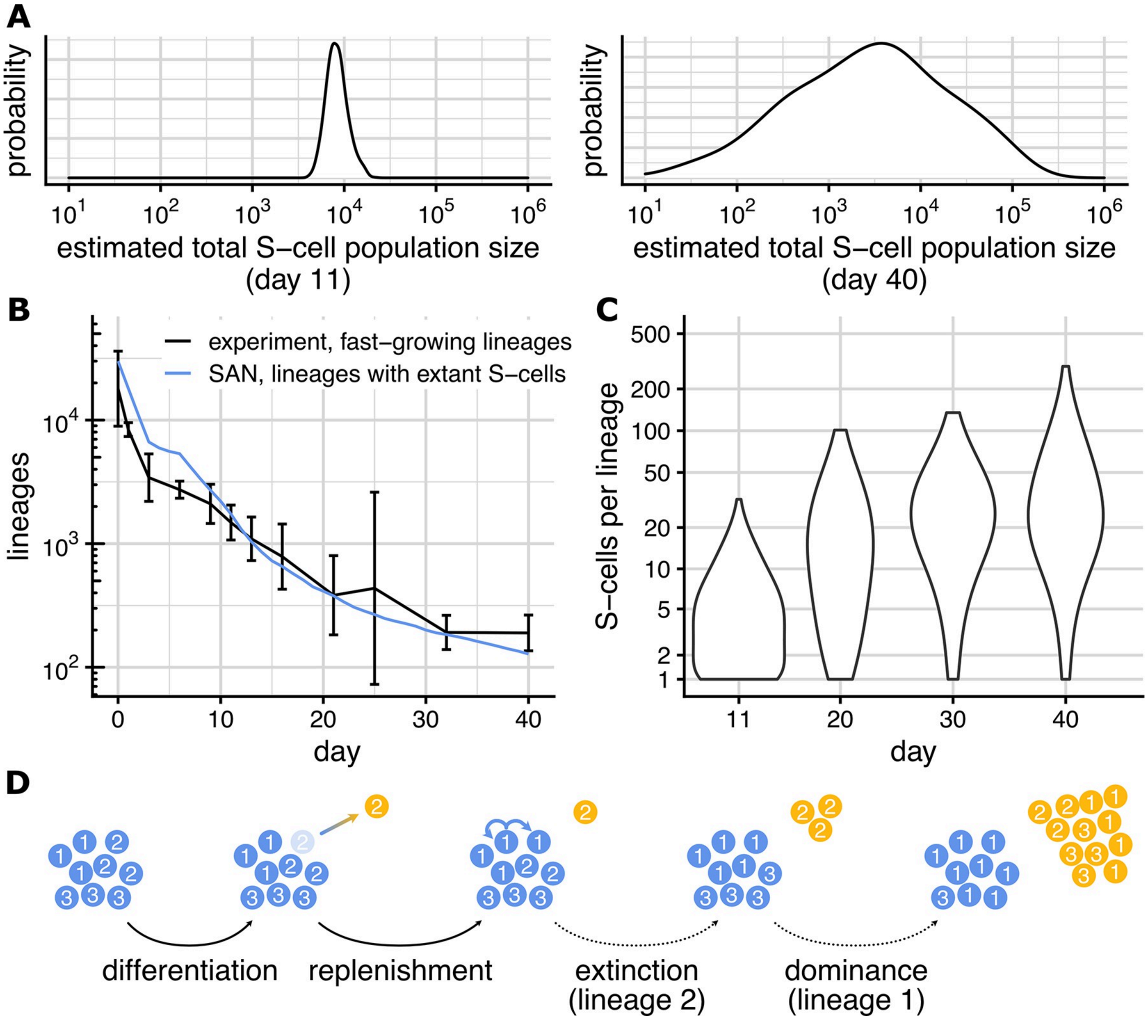

**Fig 3. The S-cell population.** **(A)** Posterior distribution of total S-cell population size on day 11 (left) and day 40 (right). Posterior distribution obtained with MCMC, see Methods. **(B)** Number of fast-growing lineages (black; from Fig 1E) number of lineages with extant S-cells as predicted 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.

https://doi.org/10.1371/journal.pcbi.1012054.g003

## Predicted S-cell population size

The MAP rate estimates (Table 1) predict that organoids contain ≈9,500 S-cells on day 11 and ≈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 computed the posterior distribution of S-cell population sizes (Fig 3A). We find that the total S-cell population size on day 11 shows a (relatively) sharp peak at about 10,000 cells. On day 40, the estimates are more dispersed, owing to cumulative stochastic effects and to the large effect of small changes to the rates over days. Yet while we cannot pin-point the exact population size for day 40, the posterior distribution indicates that likely values range from roughly 100 to 100,000. In particular, we find that organoids very likely contain at least some S-cells on day 40.

## 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 at day 40 that number has dropped to ≈200 (Fig 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 (Fig 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 (Fig 3B). This implicates S-cells as being the main driver of lineage growth. As stated above (A-cell output) a single differentiating S-cells 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

In linages with extant S-cells, the average number of S-cells grows continuously. At the same time, the spread between these 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) grows as well. 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* (Fig 3D) amongst S-cells, a term introduced to describe the population dynamics of stem cells within intestinal crypts [1].

Qualitatively, the dynamics of an organoid's S-cell population under neutral competition mimic the population-genetic Moran model [11] 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 (Fig 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 (Fig 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 [1]. 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 simulated 50,000,000 lineages using the SAN model with the MAP rate estimates given in Table 1. We then stratified the 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 $T_S = 13$ (Fig 4A left) respectively day $T_S = 25$ (Fig 4A middle) show

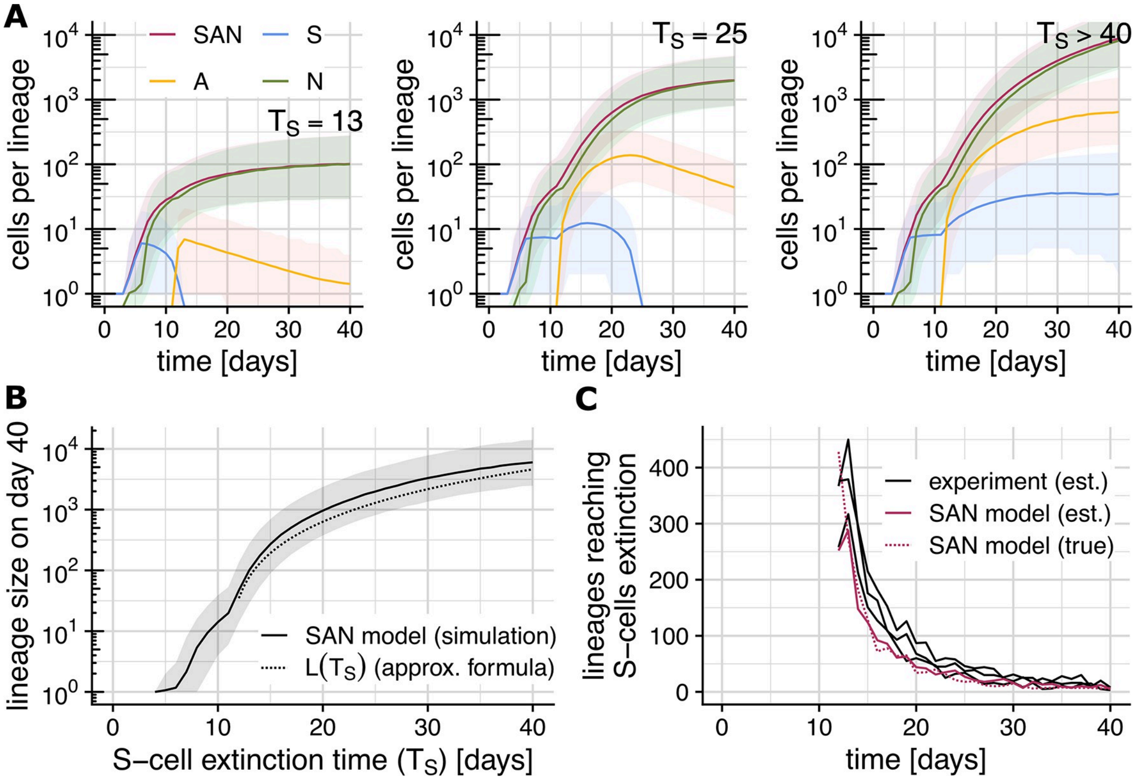

**Fig 4. Lineage-specific S-cell extinction time determines linage size.** (Shaded areas show the range between the 2.5% and 97.5% quantile across 5 billion simulations). **(A)** Lineage-specific growth trajectories under the SAN model stratified by the lineage's S-cell extinction time $T_S$. Plot show 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.

https://doi.org/10.1371/journal.pcbi.1012054.g004

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 (Fig 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 (Fig 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

$$L(\Delta T_S) = \left(\frac{1}{3}s_0\Delta T_S + \frac{r_{S\to SS}}{4\sqrt{3}}\Delta T_S^2\right)r_{S\to A}\left(1 + \frac{r_{A\to AN}}{r_{A\to N}}\right) \qquad \text{(Eq1)}$$

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 $\Delta T_S$ days after day 11. 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 (Fig 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 simulated 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 (Fig 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. Our relationship $L(\Delta T_S)$ between lineage size and S-cell extinction time together with results about the distribution of S-cell extinction times $\Delta T_S$ [12] can be used to mathematically infer the lineage size distribution expected under the SAN model. We find that lineage sizes under the SAN are indeed expected to follow a (truncated) Zipfian law, but only provided that the S-cell population is long-lived and roughly constant in size. Furthermore, in this case the predicted value $\alpha = 0.5$ for the Zipfian parameter closely matches the empirically observed value $\alpha \approx 0.46$ (see "Mathematical Analysis" in the methods section for details). The mathematical analysis of the SAN model thus independently supports the conclusion that S-cells divide and differentiate with very similar rates that we drew earlier from the MCMC-based rate estimates (Fig 2).

# Discussion

We have empirically observed lineages in developing cerebral organoids to initially grow fast and roughly uniformly until some stopping time, at which growth slows down significantly or ceases altogether. The stopping time is different for different lineages, and distributed such that the sizes of slow or non-growing lineages approximately follow a Zipfian power law with exponent $-1/\alpha, \alpha = 0.46$. While the destructive nature of sequencing-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 consider the cellular SAN model. The model is intentionally coarse to keep the number of parameters tractable. Despite omitting many known details about differentiation trajectories, the SAN model still recapitulates the observed lineage size distributions well. In particular, it recapitulates the experimentally observed emergence of a truncated power-law for the lineage size distribution with an α of about ½. This emergency is observed not only numerically; it also follows from our approximate expression (Eq 1) for lineage sizes, and we recently were able to prove it using rigorous methods for a simplified version of the SAN model [13].

The SAN model also recapitulates the small deviations from the analytically predicted truncated power-law observed in the data (Figs 1D and 2D). There are two likely causes of these deviations. The first are random lineage size differences that arise between days 6 and 11 when the number of S-cells shrinks. Since the population size is not constant during that time, no power-law will be produced. The second likely cause is that the analytical prediction holds only in the limit of large times when lineage sizes have reached their eventual maximum. At earlier times, the A → AN, A → N mechanisms will not have produced the final number of N-cells, which can cause deviations from the predicted power law.

Under the SAN model the apparently lineage-wide switch of growth regime from fast growth to slow or no growth 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 what determines how long a lineage grows and is therefore a major determinant of lineage size.

While neutral competition is the main source of lineage sizes variation, N-cell output of individual A-cells is also stochastic under the SAN model. This stochasticity is the result of A-cells choosing between continued asymmetric division and differentiation into an N-cell after each division. This stochasticity of the number of N-cells produced by an A-cell quantitatively matches the observation of Llorca *et al.* [3] that the number of neurons produced by an RGC is stochastic and varies over one to two orders of magnitude between lineages.

The relationship between a lineage's survival times within the organoid's S-cell population and the size it eventually attains can be expressed by a formula (Eq 1). Inverting this formula allows the history of the organoid's 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, the posterior distributions of these rates indicates that the rates of S-cell division and differentiation are almost identical. This is corroborated by a sensitivity analysis which shows that individual modifications of these rates drastically alter organoid size and lineage size distribution. The effect of these almost identical rates is that the S-cell population stays small, yet survives until day 40. Such a careful tuning of the rates of distinct events indicates the existence of an active control mechanism; without such a mechanism the rates would be expected to diverge over the course of 40 days. The simplest such mechanism may appear to be a direct, deterministic link between division and differentiation events, for example that one of the offspring of an S-cell always differentiates into an A-cell. Such a deterministic link, however, would not allow for neutral competition and thus not explain the large observed range of lineage sizes. The link between division and differentiation must thus exist on the organoid (or sub-population) level, not the level of individual cells. In the terminology of Simons and Clevers [14], 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 [1], we conjecture that similar structures located within proliferation centers called *neural rosettes* [5] 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 mechanisms 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. We explored a few reasonable perturbations of the SAN model to see how well such perturbations can be resolved by observing the total organoid size and linage size distribution. While not all perturbations can be expected to be identifiable through their effect on the linage size distribution alone, we do observe different effects of different types of perturbations in these simulations (S4 Fig). How accurate the predictions offered by the SAN model are in the case of perturbed organoids is an open question. We expect that responses that amount to temporal modulation of cell division and differentiation rates are modelled well, provided that the rate modulations can be estimated with sufficient accuracy. Responses of a more qualitative nature, such as a changing the mechanisms that controls the S-cell population size (e.g. to *single-cell asymmetric* in the terminology of Ref. [14]) may not be reflected accurately just by changing model parameters, however.

The finding that lineage sizes follow highly dispersed power-law distributions, and that S-cell survival times vary strongly between lineages has potentially profound consequences for many applications of neural organoids. Even though organoids are grown from tens of thousands of cells, the number of lineages that significantly contribute to an organoid will typically be much smaller. And as early as day 11, the number of lineages with an extant population of S-cells is likely to drop to around 2,000 (Fig 3B). This should be considered when organoids are used as a model system. It is particularly relevant if organoids are grown from a genetically heterogenous population of cells as is often done in organoid-based genetic perturbation screens. The SAN model is thus useful both when planning such 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.

Currently, our model and hence these applications are limited to neural organoids grown according to a protocol similar to that of Esk *et al.* [5]. This restriction is a result of data availability. While lineage tracing data is available for a range of model systems, these datasets commonly only detected a small subset of lineages or a small subset of cells, making them unsuitable for the kind of analysis we performed in this study. However, given data which like Ref. [5] resolves even single-cell lineages in most cases, we expect the SAN model to be applicable to other systems.

Neutral competition between S-cells has so far been mainly studied for tissues in homeostasis [1,7]. The SAN model demonstrates that during growth, neutral competition permanently shapes the composition of the resulting tissue. This is potentially relevant to both organoid models of other types of tissue, and to developmental processes *in vivo*.

Our simulation and parameter estimation strategies are not strongly tied to the specific structure of the SAN model. If necessary to model the growth of other types of tissue, the model could be expanded to include additional types of cells, additional differentiation paths, or more time intervals for which rates are estimated. We do assume, however, that it is appropriate to consider lineages to grow mostly independently of each other. While this is a common assumption in e.g. genetic perturbation screens, it may not be appropriate under all circumstances. In situations where lineages cannot be assumed to be mostly independent, more extensive changes to our simulation and parameter estimation methods would be required.

The SAN model also currently disregards the cell cycle, and assumes that multiple division and/or differentiation events can in principle occur in quick succession within the same cell (although the probability of that is low). While this assumption is computationally convenient because it makes the SAN model a Markov process, during times where actual inter-generation times are fairly deterministic (such as periods of fast growth) the model might over-estimate the amount of stochasticity. Future improvements of the model could remedy this using a more biologically realistic inter-generation time distribution, but doing so would come at the cost of making model simulations more computationally demanding.

To facilitate the adoption of the SAN model, we offer an R package which contains code to simulate the SAN model, to estimate parameters, and the datasets discussed in this study in pre-processed form (http://github.com/Cibiv/SANjar).

## Methods

### Experimental data

The experimental data used comprises the lineage tracing data of Esk *et al.* [5], the data for the additional replicates presented in S3 Fig (see *Sequencing data processing*), and the total organoid sizes measured using a combination of fluorescence-activated cell sorting (FACS) and image-based measurements (see *Total organoid sizes*). Fig 5 depicts the flow of data through the different processing steps described below.

### Total organoid sizes

For days 0 through 21, organoid sizes (in numbers of cells) were small enough to be measured using fluorescence-activated cell sorting (FACS). For days 11 through 40, organoid volumes were estimated from microscopy images. Organoid volumes were translated into cell counts based on the average number of cells per volume computed for days 10 from FACS and volume data (FACS data for day 10 was interpolated using days 9 and 13). The number of replicates was larger than the number of organoids eventually sequenced; S4 Table contains the raw FACS- and volume-based cell counts.

### Sequencing data processing

The pre-processed 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 underlying assumption is confirmed by the good agreement between the sum of absolute lineage sizes and the FACS and area-derived estimates of total organoid size. The data for the additional replicates shown in S3 Fig was error-corrected and filtered as described by Esk *et al.* [5]. The additional replicates were not sequenced deeply enough to reliably detect the mode in the lineage size distribution that corresponds to single-cellular lineages. To covert relative lineage sizes (measured in number of reads) to absolute lineage sizes (measured in number of cells) for these samples, we thus estimated the number of reads per cell by comparing the library size to the total organoid sizes shown in Fig 2B. The normalized version of both the data of Esk *et al.* [5] and the additional replicates shown in S3 Fig is available as part of our R package *SANjar* (http://github.com/Cibiv/SANjar).

### Pareto index and fast-slow threshold estimation

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

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

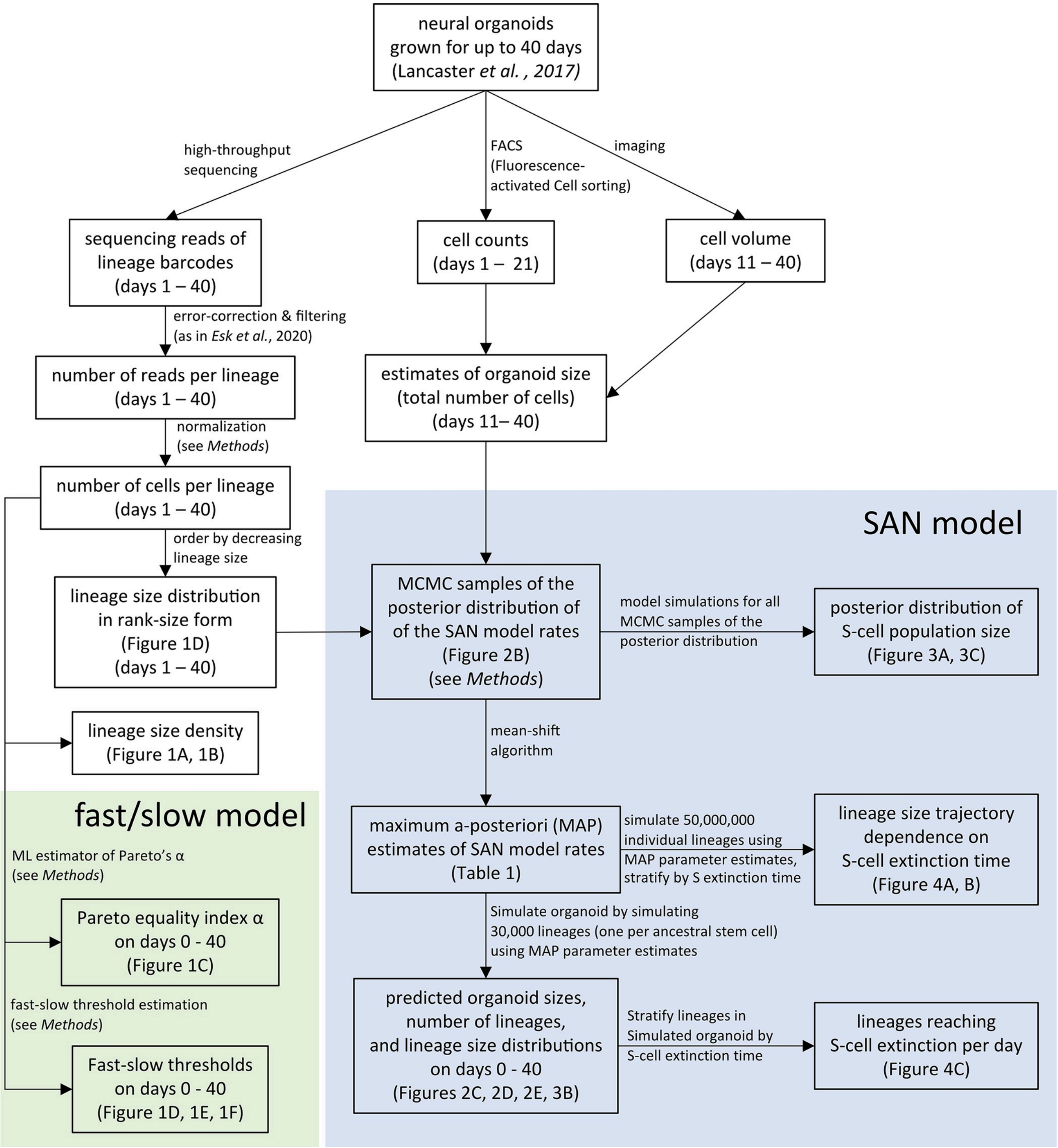

**Fig 5. Computational Flowchart.** Data sources and data flow through the difference processing steps described in the Methods section.

https://doi.org/10.1371/journal.pcbi.1012054.g005

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 $d^{\text{Pareto}}$ such that the Pareto-induced rank-size powerlaw $\log_{10} L^{\text{Pareto}}(r) =$

$-\bar{\alpha}^{-1}\log_{10} r + d^{\text{Pareto}}$ 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 $L^{\text{Pareto}}(r)$. We then fit a separate log-log-linear model $\log_{10} L^{\text{Large}}(r) = k\log_{10} r + d^{\text{Large}}$ to lineages with ranks $1, \ldots, \sqrt{R}$ (which we assume are surely not governed by the Pareto law), and set $l_{\text{Th}}$ to the size at which the two laws intersect (meaning $l_{\text{Th}} = L^{\text{Pareto}}(r) = L^{\text{Large}}(r)$).

## SAN model

In the SAN model, the clonal composition of a neural organoid at time $t$ is described by the number of S-, A- and N-cells in each of the lineages that comprise an organoid. We now focus on a single lineage. The state of a single lineage at a particular point in time is described by an integer-valued vector $(s,a,n)$ where $s$ denotes the number of S-cells, $a$ the number of S-cells and $n$ the number of N-cells. Cells can divide (producing new cells), differentiate (changing to a different cell type) or a removed as depicted in Fig 2A. For a particular cell, these events occur stochastically with event-specific and potentially time-dependent rates $r_{S\to SS}, r_{S\to\emptyset}, r_{S\to A}, r_{S\to N}, r_{A\to AN}$, and $r_{A\to N}$. Since the SAN model is stochastic, the appropriate mathematical description is in terms of a function $p(s,a,n,t)$ which represents the probability of the lineage comprising a certain number of S-, A- and N-cells at a specific time. Initially, at time $t = 0$, a lineage contains a single S-cell, and thus $p(1,a,n,0) = 1$ (and $p(s,a,n,0) = 0$ for all other values of $s$). For other time points, $p(s,a,n,t)$ is, at least in principle, fully defined by a so-called *master equation*. The master equation is found by accounting for all ways in which the system can enter or leave a specific state over an infitesimal time interval. For the SAN model the master equation for $p(s,a,n,t)$ is

$$\begin{aligned}\frac{dp(s,a,n,t)}{dt} &= r_{S\to SS}\cdot(s-1)\cdot p(s-1,a,n,t)\\ &+r_{S\to\emptyset}\cdot(s+1)\cdot p(s+1,a,n,t)\\ &+r_{S\to A}\cdot(s+1)\cdot p(s+1,a-1,n,t)\\ &+r_{S\to N}\cdot(s+1)\cdot p(s+1,a,n-1,t)\\ &+r_{A\to AN}\cdot a\cdot p(s,a,n-1,t)\\ &+r_{A\to N}\cdot(a+1)\cdot p(s,a+1,n-1,t)\\ &-(r_{S\to SS}+r_{S\to\emptyset}+r_{S\to A}+r_{S\to N})\cdot s\cdot p(s,a,n,t)\\ &-(r_{A\to AN}+r_{A\to N})\cdot a\cdot p(s,a,n,t).\end{aligned}$$

While this master equation unambiguously defines the SAN model mathematically, it is not a convenient tool for numerical simulations. The numerical algorithm we use for simulations is described under *SAN model simulation* below.

The full, stochastic SAN model is relevant when we consider the behavior of individual lineages. When we consider the total cell counts of a whole organoid, however, stochastic effects average out. In that case, the so-called *deterministic* SAN model is useful. It describes the expected total number of S-, A- and N-cells across a large number of lineages, each governed by the stochastic SAN model. The ordinary differential equations (ODEs) for the deterministic

SAN model are

$$\dot{s} = (r_{S\to SS} - r_{S\to\emptyset} - r_{S\to A} - r_{S\to N}) \cdot s,$$

$$\dot{a} = r_{S\to A} \cdot s - r_{A\to N} \cdot a,$$

$$\dot{n} = r_{S\to N} \cdot s + (r_{A\to AN} + r_{A\to N}) \cdot a.$$

## SAN model simulation

The total number of S-, A- and N-cells that an organoid is predicted to comprise at time $t$ is computed by evaluating the analytical solution of the deterministic SAN model. This is done separately over each interval in which the 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 $\Delta t$, 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 $\Delta_e$ undergoing event e is chosen from a Poisson distribution. Specifically,

$$\Delta_{S\to SS} \sim \text{Poisson}(r_{S\to SS} S_i(t)\Delta t), \Delta_{S\to\emptyset} \sim \text{Poisson}(r_{S\to\emptyset} S_i(t)\Delta t),$$

$$\Delta_{S\to A} \sim \text{Poisson}(r_{S\to N} S_i(t)\Delta t), \Delta_{S\to N} \sim \text{Poisson}(r_{S\to N} S_i(t)\Delta t),$$

$$\Delta_{A\to AN} \sim \text{Poisson}(r_{A\to AN} A_i(t)\Delta t), \Delta_{A\to N} \sim \text{Poisson}(r_{A\to N} A_i(t)\Delta t),$$

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

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

$$a_i(t+\Delta t) = a_i(t) + \Delta_{S\to A} - \Delta_{A\to N,}$$

$$n_i(t+\Delta t) = n_i(t) + \Delta_{S\to N} + \Delta_{A\to AN} + \Delta_{A\to N.}$$

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

While continuous-time methods exist to simulate the SAN model exactly, these methods have the drawback that that they simulate each event separately. Their runtime thus strongly depends on the total number of cells produced, and hence on the parameter values. Since the MCMC strives to fully explore the parameter space, continuous-time algorithms perform much worse than our discrete-time algorithm, to the extent that they make running MCMC to convergence impractical. To ensure that using a discrete-time algorithm did not affect our results, we compared our simulation results against an exact algorithm and verified that the differences were negligible.

## Simulation of sequencing-based lineage tracing

The effect of PCR amplification and sequencing on the observed lineage sizes was simulated using a stochastic model of PCR amplification and sequencing [15] 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* $Wl_i/\Sigma_i l_i$ 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 ($W/\Sigma_i l_i$).

## SAN rate estimation

To account for the major phases of the neural organoid protocol of Lancaster *et al*., (2017), we sub-divide days 0–40 into the four time-intervals 0–3, 3–6, 6–11 and 11–40. Briefly, until day 3, cells are expected to acclimatize and not divide. Until day 6, embryoid body (EB) formation is ongoing, and lineages may still be lost if not all cells end up as part of the EB. Between days 6 and 11, only symmetric divisions are expected to occur. On day 11, cells start to differentiate, and asymmetric divisions start to occur.

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 $S \rightarrow S\ S$ and $S \rightarrow N$ between 6 and 11, and $S \rightarrow S\ S$, $S \rightarrow A$, $A \rightarrow A\ N$ and $A \rightarrow N$ between days 11 and 40) we used a Markov-Chain Monte Carlo (MCMC) method to compute their posterior distribution given the experimental observations. The experimental observations comprised the total organoid sizes $\hat{c}^{(\text{day } t)}$ (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 the ranked lineage sizes $\hat{l}_{(r)}^{(\text{day } 11)}$, $\hat{l}_{(r)}^{(\text{day } 40)}$ (on days 11 and 40, for ranks $r \in \mathcal{R} = \{1, 2, 5, 10, 15, 25, 40, 60, 100, 150, 250, 400, 600, 1000, 1500, 2500, 4000, 6000, 10000, 15000, 25000\}$). For the total organoid size, we included all available experimental data to strongly penalize against parameters values which yield unrealistic growth behavior. For the ranked lineage sizes, we chose to include only the endpoints of the two time-intervals (days 6–11 and 11–40). This leaves the days in between available for model validation. To account for biological differences between replicates we assumed that experimental observations are log-normally distributed. The log-likelihood of the rate vector $\theta$ (comprising the 6 rates mentioned above) given the experimental data is thus

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

Here, $c^{(\text{day } t)}$ and $l_{(r)}^{(\text{day } 11)}$, $l_{(r)}^{(\text{day } 40)}$ are the SAN model predictions for total organoid size and ranked lineage sizes given parameter values $\theta$, and $\mu[\cdots]$ and $\sigma[\cdots]$ 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 $\theta$ 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 [16–18]. Each chain was initialized with random parameters drawn uniformly from the interval [0,4], and each chain was run until it had accepted 1000 moves. To verify convergence of the MCMC chains, we computed the Gelman-Rubin (GR) diagnostic [10] for each estimated parameter (plus the net S rate, i.e. the difference between the rate of S-cell production and differentiation). Our choice of uniformly distributed initial values satisfies the requirement over being over-dispersed relative to the posterior distribution, making the GR diagnostic applicable. For all rates except A → N, the 95% confidence intervals for the GR diagnostic lay below 1.2. For the rate A → N, the upper bound of the CI was 1.24 and the point estimate 1.19 (S1 Table). We concluded that the MCMC chains have explored the parameter space sufficiently deeply for the posterior distribution to be accurate. To find point estimates of the parameter, we computed the maximal mode of the (joint) posterior distribution with the mean-shift algorithm to obtain the MAP estimates (Table 1). The code used for MCMC sampling and mode-finding is available as part of our R package.

## Mathematical analysis

If we consider only S-cells, the SAN model corresponds to the well-known birth-death process [12]. We consider the diffusion approximation of this process and restrict our mathematical treatment to day 11 and later where the rates of symmetric division ($r_{S\rightarrow SS}$; birth) and of differentiation ($r_{S\rightarrow A}$;amounts to death since we consider only S-cells) are similar enough to be considered identical ($r_{S\rightarrow SS} = r_{S\rightarrow A} = \lambda/2$). 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 [19,20] it is possible to find the most probable trajectory of a linage that contains $S_0$ cells at $t = 0$ and loses its last S-cell $\Delta T_S$ days later [13],

$$s_{\text{ext}}(t|s_0, \Delta T_S) = s_0\left(1 - \frac{t}{\Delta T_S}\right)\left(1 + \rho\frac{t}{\Delta T_S}\right) \text{ where } \rho = \Delta T_S\frac{\lambda}{s_0}\frac{\sqrt{3}}{4} - 1.$$

On average, a lineage grows by $r_{S\rightarrow A}$ A-cells per S-cell and per day, and over its lifetime every A-cell will eventually produce $r_{A\rightarrow AN}/r_{A\rightarrow N}$ additional N-cells through asymmetric division. Eventually, a lineage that starts out with $S_0$ S-cells and loses its last S-cell $\Delta T_S$ days later will thus approximately grow to size

$$L(\Delta T_S) = \int_0^{\Delta T_s} s_{\text{ext}}(t|s_0, \Delta T_S)r_{S\rightarrow A}\left(1 + \frac{r_{A\rightarrow AN}}{r_{A\rightarrow N}}\right)dt.$$

Integration of this expression and setting $\lambda = 2r_{S\rightarrow SS}$ yields Eq (1).

For an S-cell population governed by the SDE $ds(t) = \sqrt{\lambda s(t)}dW(t)$, the probability that $\Delta T_S \geq t$, is $\exp(-2S_0/\lambda t)$, and for sufficiently large $t$ the probability that $\Delta T_S \approx t$ is thus approximately $2S_0t^{-2}/\lambda$. By using Eq (1) to translate this distribution of $\Delta T_S$ into a distribution of lineage sizes we find that the distribution of lineage sizes with large extinction times $\Delta T_S$ should follow a Pareto distribution with $\alpha = 1/2$. A more detailed mathematical analysis of the SAN model with an emphasis on criticality can be found in Ref. [13].

## Supporting information

**S1 Supplemental Methods. . Additional details about some procedures.**
(PDF)

**S1 Fig. Relative lineage size frequencies for days 1 through 40.** Relative frequencies of different lineage sizes on day 40 vs. Pareto power law with equality index $\bar{\alpha} = 0.46$.
(PDF)

**S2 Fig. Predicted lineage sizes without PCR and sequencing effects.** Rank-size distributions observed experimentally (black) and predicted by the SAN model (red). Other than in Fig 2D, the prediction shows the number of cells predicted by the model, not the number of cells observed through high-throughput sequencing.
(PDF)

**S3 Fig. Replicate Experiments.** Replicate experiments based on the same organoid protocol show similar lineage size distributions as the data from Esk et al. (2020). Plots show three (days 6 and 11) respectively two (days 1 and 40) replicates; the replicates not shown were excluded due to sequencing quality issues. Ranks of the Esk et al. data and SAN model predictions were scaled to account for a 1.7-fold increase in the number of detected lineages in the replicate experiments. See S1 Supplemental Methods for experimental details.
(PDF)

**S4 Fig. Perturbation Simulations.** Rate of asymmetric division after day 11 reduced to one-half (blue) and one-tenth (black) of its original value. (B). Rate of symmetric division after day 11 reduced to one-half (blue) and one-tenth (black) of its original value. Rate of S -> A differentiations reduced accordingly to keep the net S-cell growth rate unchanged (C). Immediate differentiation of S-cells on day 11, no symmetric divisions.
(PDF)

**S5 Fig. Sensitivity analysis of rates for days 6–11.** Model response to modified rates for days 11 to 40. MAP estimates (Table *1*) are increased/decreased two standard deviations of the posterior (0.94 ± 0.28 for $S \rightarrow S\ S$; 1.14 ± 0.35 for $S \rightarrow N$, see also S6 Table). The leftmost plot shows the total organoid size (red), S-cells (blue), A-cells (yellow), and N-cells (green). The plots on the right show the lineage size distribution. Solid lines represent the unmodified model. Dashed lines represent the two modified models, the area in between is shaded.
(PDF)

**S6 Fig. Sensitivity analysis of rates for days 11–40.** Model response to modified rates for days 11 to 40, including net S production $S \rightarrow S\ S$ minus $S \rightarrow A$. MAP estimates (Table *1*) are increased/decreased two standard deviations of the posterior (1.68 ± 1.04 for $S \rightarrow S\ S$; 1.69 ± 1.18 for $S \rightarrow A$; -0.022 ± 0.128 for net S; 0.71 ± 1.56 for $A \rightarrow AN$; 0.072 ± 1.27 for A -> N; see also S6 Table). The leftmost plot shows the total organoid size (red), S-cells (blue), A-cells (yellow), and N-cells (green). The plots on the right show the lineage size distribution. Solid lines represent the unmodified model. Dashed lines represent the two modified models, the area in between is shaded.
(PDF)

**S1 Table. Gelman-Rubin diagnostic for MCMC results.** Values of the point estimate and upper bound of a 95%-CI of the Gelman-Rubin diagnostic.
(XLSX)

**S2 Table. Fast/slow regime thresholds.** Estimates threshold size ($l_{Th}$) and threshold rank ($n_{Th}$) for each replicate. The column $\mu$ contains the mean across the replicates for each day, $\mu-2\sigma$ and $\mu-2\sigma$ contain the mean plus/minus two standard deviations.
(XLSX)

**S3 Table. Number of observed lineages.** Number of observed lineages for each replicate. The column $\mu$ contains the mean across the replicates for each day, $\mu-2\sigma$ and $\mu-2\sigma$ contain the mean plus/minus two standard deviations.
(XLSX)

**S4 Table. Organoid sizes.** Organoids sizes measured using fluorescence-activated cell sorting (FACS) and using an image-based method.
(XLSX)

**S5 Table. Pareto equality indices.** Estimates Pareto equality indices $\alpha$ for each replicate. The column $\mu$ contains the mean across the replicates for each day, $\mu-2\sigma$ and $\mu-2\sigma$ contain the

mean plus/minus two standard deviations.
(XLSX)

**S6 Table. Summary statistics of the posterior distribution.** Different summary statistics for the 1,000 samples of the posterior distribution we obtained using MCMC.
(XLSX)

## Acknowledgments

We thank Egor Kiselev for many fruitful discussions about criticality.

## Author Contributions

**Conceptualization:** Florian G. Pflug, Christopher Esk, Arndt von Haeseler.

**Data curation:** Florian G. Pflug, Simon Haendeler, Christopher Esk, Dominik Lindenhofer.

**Formal analysis:** Florian G. Pflug.

**Funding acquisition:** Jürgen A. Knoblich, Arndt von Haeseler.

**Investigation:** Florian G. Pflug, Simon Haendeler.

**Project administration:** Jürgen A. Knoblich, Arndt von Haeseler.

**Software:** Florian G. Pflug.

**Supervision:** Jürgen A. Knoblich, Arndt von Haeseler.

**Validation:** Florian G. Pflug, Christopher Esk, Dominik Lindenhofer.

**Visualization:** Florian G. Pflug.

**Writing – original draft:** Florian G. Pflug.

**Writing – review & editing:** Florian G. Pflug, Simon Haendeler, Christopher Esk, Dominik Lindenhofer, Arndt von Haeseler.

## References

1. Snippert HJ, van der Flier LG, Sato T, van Es JH, van den Born M, Kroon-Veenboer C, et al. Intestinal Crypt Homeostasis Results from Neutral Competition between Symmetrically Dividing Lgr5 Stem Cells. Cell. 2010; 143: 134–144. https://doi.org/10.1016/j.cell.2010.09.016 PMID: 20887898
2. Klingler E, Jabaudon D. Do progenitors play dice? eLife. 2020; 9: e54042. https://doi.org/10.7554/eLife.54042 PMID: 31951199
3. Llorca A, Ciceri G, Beattie R, Wong FK, Diana G, Serafeimidou-Pouliou E, et al. A stochastic framework of neurogenesis underlies the assembly of neocortical cytoarchitecture. eLife. 2019; 8: e51381. https://doi.org/10.7554/eLife.51381 PMID: 31736464
4. Lancaster MA, Corsini NS, Wolfinger S, Gustafson EH, Phillips AW, Burkard TR, et al. Guided self-organization and cortical plate formation in human brain organoids. Nature Biotechnology. 2017; 35: 659–666. https://doi.org/10.1038/nbt.3906 PMID: 28562594
5. Esk C, Lindenhofer D, Haendeler S, Wester RA, Pflug F, Schroeder B, et al. A human tissue screen identifies a regulator of ER secretion as a brain-size determinant. Science. 2020; 370: 935–941. https://doi.org/10.1126/science.abb5390 PMID: 33122427
6. Zechner C, Nerli E, Norden C. Stochasticity and determinism in cell fate decisions. Development. 2020; 147: dev181495. https://doi.org/10.1242/dev.181495 PMID: 32669276
7. Corominas-Murtra B, Scheele CLGJ, Kishi K, Ellenbroek SIJ, Simons BD, van Rheenen J, et al. Stem cell lineage survival as a noisy competition for niche access. Proceedings of the National Academy of Sciences. 2020; 117: 16969–16975. https://doi.org/10.1073/pnas.1921205117 PMID: 32611816

8. Pinto CMA, Mendes Lopes A, Machado JAT. A review of power laws in real life phenomena. Communications in Nonlinear Science and Numerical Simulation. 2012; 17: 3558–3578. https://doi.org/10.1016/j.cnsns.2012.01.013

9. Adamic LA, Huberman B. Zipf's law and the Internet. Glottometrics. 2002; 3: 143–150.

10. Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences. Statistical Science. 1992; 7: 457–472. https://doi.org/10.1214/ss/1177011136

11. Moran PAP. Random processes in genetics. Mathematical Proceedings of the Cambridge Philosophical Society. 1958; 54: 60–71. https://doi.org/10.1017/S0305004100033193

12. Feller W. Die Grundlagen der Volterraschen Theorie des Kampfes ums Dasein in wahrscheinlichkeitstheoretischer Behandlung. Acta Biotheoretica. 1939; 5: 11–40. https://doi.org/10.1007/BF01602932

13. Kiselev EI, Pflug F, von Haeseler A. Critical Growth of Cerebral Tissue in Organoids: Theory and Experiments. Phys Rev Lett. 2023; 131: 178402. https://doi.org/10.1103/PhysRevLett.131.178402 PMID: 37955473

14. Simons BD, Clevers H. Strategies for homeostatic stem cell self-renewal in adult tissues. Cell. 2011; 145: 851–862. https://doi.org/10.1016/j.cell.2011.05.033 PMID: 21663791

15. Pflug FG, von Haeseler A. TRUmiCount: correctly counting absolute numbers of molecules using unique molecular identifiers. Bioinformatics. 2018; 34: 3137–3144. https://doi.org/10.1093/bioinformatics/bty283 PMID: 29672674

16. Beaumont MA. Estimation of population growth or decline in genetically monitored populations. Genetics. 2003; 164: 1139–1160. https://doi.org/10.1093/genetics/164.3.1139 PMID: 12871921

17. Andrieu C, Roberts GO. The pseudo-marginal approach for efficient Monte Carlo computations. Annals of Statistics. 2009; 37: 697–725. https://doi.org/10.1214/07-AOS574

18. Warne DJ, Baker RE, Simpson MJ. A practical guide to pseudo-marginal methods for computational inference in systems biology. Journal of Theoretical Biology. 2020; 496: 11025. https://doi.org/10.1016/j.jtbi.2020.110255 PMID: 32223995

19. Onsager L, Machlup S. Fluctuations and irreversible processes. Physical Review. 1953; 91: 1505–1512. https://doi.org/10.1103/PhysRev.91.1505

20. Dürr D, Bach A. The Onsager-Machlup function as Lagrangian for the most probable path of a diffusion process. Communications in Mathematical Physics. 1978; 60: 153–170. https://doi.org/10.1007/BF01609446