nature

https://doi.org/10.1038/s41586-020-2503-6

Supplementary information

Single-cell lineage tracing unveils a role for TCF15 in haematopoiesis

In the format provided by the authors and unedited

Alejo E. Rodriguez-Fraticelli, Caleb Weinreb, Shou-Wen Wang, Rosa P. Migueles, Maja Jankovic, Marc Usart, Allon M. Klein, Sally Lowell & Fernando D. Camargo✉

Supplementary Methods

Theory and data analysis for serial clonal data

This supplementary methods include (1) the statistical model used to reject the hypothesis that HSCs are equipotent based on their clonal dynamics following serial transplantation; and (2) considerations and methods for error correction of clonal barcodes in scRNA-Seq data sets.

CONTENTS

I. CLONE SIZE STATISTICS OF EQUIPOTENT STEM CELLS

A. Model definition

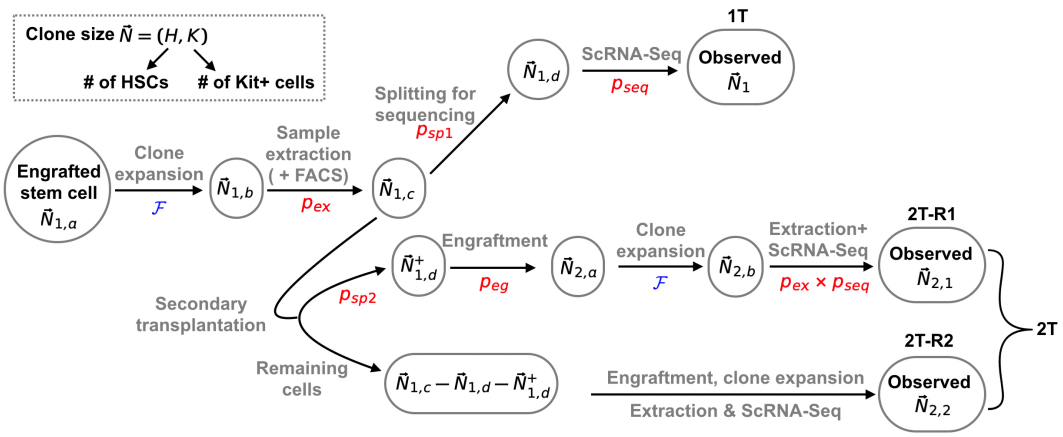
In our experiment, genetically barcoded HSCs are transplanted into the first host mouse, and part of the clonal output from the first transplantation are further transplanted into the secondary mouse. Both the clonal output from the first transplantation (1T) and secondary transplantation (2T) are profiled via single-cell sequencing. Our goal is to construct a distribution over the expected outcomes of such an experiment, assuming a null model in which each HSC behaves independently and has equal potency to self-renew and to generate differentiated cells after engraftment. By 'equal potency' we specifically mean that the behavior of two sister HSCs should be as different as that of two randomly selected HSCs. The self-renewal and differentiation of each cell can still be highly variable. To relate such a model to experimental observables, we must account for the statistical processes in this experiment, including the random sampling of cells, uncertainty in the engraftment of HSCs, and variation in their clonal expansion. We incorporate all of these uncertainties using the formalism of a probabilistic directed graphical model (PDGM) (Supplementary Methods Fig. 1), which explicitly encodes the conditional dependence between different stochastic variables and lends itself to rapid numerical calculation of the Likelihood of the observed data [1].

Referring to Supplementary Methods Fig. 1, the PDGM defines a distribution over nine latent variables and three observed variables, representing the size of each clone at subsequent stages of the experiment, according to the notation in Supplementary Methods Fig. 1, and as summarized in Supplementary Methods Table I. We introduce a compact notation for the clone size: $\vec{N} = (H, K)$, where H is the number of HSCs in a clone and K is the number of Kit+ progenies. Referring to Supplementary Methods Fig. 1, the nested structure of the PDGM reflects the successive steps occurring after a single HSC successfully engrafts in the primary transplantation, assuming that all HSCs are equipotent at the time of transplantation. Specifically, keeping in mind the equipotency assumption of this model, the successive steps of the PDGM are:

---

1. A single HSC engrafts per clone $[\vec{N}_{1,a} = (1, 0)]$ in the primary host.
2. Each engrafted HSC expands for three months to give rise to a clone of size $\vec{N}_{1,b}$. The clone size distribution after expansion $\mathcal{F}(\vec{N}_{1,b})$ is estimated empirically as described shortly.
3. Only a fraction of cells are successfully isolated from the bone marrow of the primary transplantation host, resulting in a clone of smaller size $\vec{N}_{1,c}$. Here, sample extraction also includes the following step of FACS designed for enriching HSCs. Survival of the extraction process is assumed to be i.i.d. for each HSC across all clones.

4. Then, a fraction of the cells are binomially sampled for immediate scRNA-Seq analysis ($\vec{N}_{1,d}$ cells are sampled). Sampling outcome is an i.i.d. Bernoulli process for all HSCs.
5. Only a fraction of the cells sampled for sequencing are successfully barcoded in the inDrop device, giving the final observed number of cells at 1T as $\vec{N}_1$ cells. Survival of scRNA-Seq is an i.i.d. Bernoulli process for all HSCs.
6. Of the remaining $(\vec{N}_{1,c} - \vec{N}_{1,d})$ cells, a fraction of HSCs $(\vec{N}^+_{1,d})$ are transplanted into one secondary mouse (R1) and the remaining into another secondary mouse (R2). As the splitting is random and i.i.d., the distribution between the two secondary hosts is multinomial.
7. For each of the R1 and R2 secondary host mice, only a subset of transplanted HSCs will successfully engraft ($\vec{N}^{(i)}_{2,a}$; $i = 1, 2$). We assume that engraftment outcome is i.i.d. for all HSCs across all clones.
8. The engrafted HSCs expand in the secondary hosts for 3 months to give rise to a clone of size ($\vec{N}^{(i)}_{2,b}$; $i = 1, 2$). The expansion of each individual HSC in each clone is assumed to follow the same distribution as in the primary host, $\mathcal{F}$.
9. These cells are extracted and profiled by scRNA-Seq, giving the final observed clone sizes $\vec{N}_{2,1}$ in 2T-R1, and $\vec{N}_{2,2}$ in 2T-R2. Survival of extraction and purification follows the same i.i.d. processes as in steps 4,5.



$$P(\vec{N}_{1,a}, \ldots, \vec{N}^{(2)}_{2,b}) = \delta_{H_{1,a},1}\delta_{K_{1,a},0} \times P_b(\vec{N}_{1,b}|\vec{N}_{1,a}) \times P_c(\vec{N}_{1,c}|\vec{N}_{1,b}) \times P_d(\vec{N}_{1,d}|\vec{N}_{1,c})$$
$$\times P_e(\vec{N}_1|\vec{N}_{1,d}) P_f(\vec{N}^+_{1,d}|\vec{N}_{1,c} - \vec{N}_{1,d}) \times P_g(\vec{N}^{(1)}_{2,a}|H^+_{1,d}) \times P_b(\vec{N}^{(1)}_{2,b}|H^{(1)}_{2,a})$$
$$\times P_h(\vec{N}^{(1)}_2|\vec{N}^{(1)}_{2,b}) \times P_g(\vec{N}^{(2)}_{2,a}|H_{1,c} - H_{1,d} - H^+_{1,d}) \times P_b(\vec{N}^{(2)}_{2,b}|H^{(2)}_{2,a}) \times P_h(\vec{N}^{(2)}_2|\vec{N}^{(2)}_{2,b})$$

Supplementary Methods Fig. 1. **A probabilistic directed graphical model (PDGM) of the serial transplantation experiment for a single clone.** Each node represents the clone size $\vec{N} = (H, K)$, where H is the number of HSCs and K is the number of Kit+ cells. The model defines the joint distribution shown below its diagrammatical representation, with distributions $P_a \ldots P_h$ defined in Supplementary Methods Table I. Symbols in red are success rates for the corresponding sampling steps modeled as binomial processes (see Supplementary Methods Table I and section I A for more details). The empirical clonal expansion distribution $\mathcal{F}$ is described in the text.

This model structure can be parameterized by five probabilities, most of which are independently estimated from the experimental design. The parameters are: the HSC engraftment probability p_{eg} (step 7), the probability of cell extraction and recovery p_{ex} (steps 3 and 9), the cell fraction sampled for scRNA-Seq after the primary transplantation p_{sp1} (step 4), the cell fraction transplanted into each of the secondary host mice p_{sp2} (step 6), and the fraction of cells successfully captured during scRNA-Seq p_{seq} (steps 5 and 9). In addition, the model accepts as input a clonal expansion distribution $\mathcal{F}$ (steps 2 and 8), whose structure is defined in the following section. For reference in the remaining discussion, the PDGM variables and their distributions are summarized in Supplementary Methods Table I.

B. PDGM parameter values

The parameter values are summarized in Supplementary Methods Table II. All parameters can be constrained from independent empirical sources. Specifically: the values of $p_{sp,1}$ and $p_{sp,2}$ were both 0.5, because in each of the

Variable	Description	Distribution
$\vec{N}_{1,a}$	Clone size initially transplanted (1 HSC)	$\delta_{H,1}\delta_{K,0}$
$\vec{N}_{1,b}$	Clone size after 3 months primary transplant	$P_b \sim \mathcal{F}$
$\vec{N}_{1,c}$	Clone size surviving BM extraction and FACS	$P_c \sim \mathrm{Bin}(\vec{N}_{1,b}, p_{ex})$
$\vec{N}_{1,d}$	Number of cells sampled for scRNA-Seq after primary transplant	$P_d \sim \mathrm{Bin}(\vec{N}_{1,c}, p_{sp1})$
$\vec{N}_1$	Observed clone size in primary scRNA-Seq	$P_e \sim \mathrm{Bin}(\vec{N}_{1,d}, p_{seq})$
$\vec{N}_{1,c} - \vec{N}_{1,d}$	Number of cells transferred for tranplantation into secondary hosts	-
$\vec{N}^+_{1,d}$	Number of cells transplanted into secondary host R1	$P_f \sim \mathrm{Bin}(\vec{N}_{1,c} - \vec{N}_{1,d}, p_{sp2})$
$\vec{N}_{1,c} - \vec{N}_{1,d} - \vec{N}^+_{1,d}$	Number of cells transplanted into secondary host R2	-
$\vec{N}^{(i)}_{2,a}$	Cells successfully engrafted into secondary host Ri ($i = 1, 2$)	$P_g \sim (\mathrm{Bin}(H_{1,d}, p_{eg}), 0)$
$\vec{N}^{(i)}_{2,b}$	Clone size after 3 months secondary transplant in host Ri ($i = 1, 2$)	$P_b \sim \mathcal{F}^{H_{2,a}}$
$\vec{N}_{2,i}$	Clone size detected after extraction and scRNA-Seq from host Ri ($i = 1, 2$)	$P_h \sim \mathrm{Bin}(\vec{N}_{2,b}, p_{seq}p_{ex})$
$\vec{N}_2$	$\vec{N}_{2.1} + \vec{N}_{2.2}$	-

Supplementary Methods Table I. **Summary of model variables for the PDGM shown in Supplementary Methods Fig. 1.** In the table, $\mathrm{Bin}(N, p)$ is the Binomial distribution with N trials and probability of success p, $\mathrm{Bin}(\vec{N}, p) = (\mathrm{Bin}(H, p), \mathrm{Bin}(K, p))$, and $\mathcal{F}^k$ is a convolution of $\mathcal{F}$ with itself k times.

respective splitting steps half of the volume of a well-mixed cell suspension was transferred to sequencing ($p_{sp,1}$) or into each of the secondary host mice ($p_{sp,2}$). The value of p_{seq} reflects the performance of the inDrop device, and is reflected in the ratio of the number of observed single cell transcriptomes to the number of cells loaded into the device ($p_{seq} = 70\%$). The value of p_{ex} reflects the product of the efficiencies of HPC isolation from mice, and the survival of flow cytometry (FACS). To estimate these survival fractions, we measured the number of nucleated hematopoietic cells (HPCs) immediately after bone marrow extraction ($\sim 3.5 \times 10^8$). As the number of nucleated HPCs in the same strain of mouse is relatively stable and is estimated to be around 5×10^8 [2], the extraction efficiency is approximately 70%. The process of enriching for HSCs through flow cytometry (FACS) has an efficiency of 80% based on the difference between the number of cells recorded by the FACS machine and the final number of cells counted in suspension. Hence, the compound probability in this step is $p_{ex} = 0.56$. Finally, we define the HSC engraftment probability as the probability that an HSC gives rise to a surviving clone after 3 months post-transplantation. We have performed primary transplantation for several different mouse, and the ratio of the observed clone numbers to the transplanted HSCs, which is a rough approximation for p_{eg}, ranges from 0.07 to 0.22. Given this broad range, we selected a value of $p_{eg} = 0.137$ that best reproduced the observed ratio of clone number between 2T and 1T, i.e., $133/414 = 0.32$ (Supplementary Methods Fig. 2**B**).

Parameter	Value	Justification
p_{ex}	0.56	See text
$p_{sp,1}$	0.5	Half volume of well-mixed cells are sent for seqeuncing
$p_{sp,2}$	0.5	Half volume of remaining well-mixed cells are sent for secondary transplantation in R1
p_{seq}	0.7	Capture efficiency measured in inDrops
p_{eg}	0.137	By fitting to the observed ratio between 2T and 1T clone number, and constrained by the empirical range

Supplementary Methods Table II. **Summary of PDGM parameter values.**

C. Generating the clonal expansion distribution $\mathcal{F}$

Though the distribution $\mathcal{F}$ over the number of progeny from each HSC is not measured directly, it can be numerically estimated through Bayesian inference under the equipotency assumption of the model. With $q = p_{ex}p_{sp1}p_{seq}$ and

$\{\vec{N}_1\} = \{\vec{N}_1^{(1)}, \ldots, \vec{N}_1^{(N)}\}$ being the list of observed clone sizes after sequencing the probability distribution for the output of a single engrafted HSC is inferred to be,

$$\mathcal{F}(\vec{N}_{1,b}|\left\{\vec{N}_1\right\}, q) = \alpha\frac{1}{N}\sum_k P(H_{1,b}|H_1^{(k)}, q)P(K_{1,b}|K_1^{(k)}, q) + (1-\alpha)P(H_{1,b}|0,q)P(K_{1,b}|0,q), \tag{1}$$

where

$$P(X_{1,b}|X_1, q) = \begin{cases} \binom{X_{1,b}}{X_1} q^{X_1+1}(1-q)^{X_{1,b}-X_1} & \text{if } X_1 \geq 0,\ X_{1,b} \geq X_1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

and α is the probability of detecting at least one cell from an engrafted HSC in the 1T measurement. The value of α is determined self-consistently from the PDGM (discussed below). The sum in Eq. (1) averages $\mathcal{F}$ over all observed clone sizes, with the final term correcting the distribution to account for clones that were not observed due to technical drop-outs during cell sampling for analysis (as *observed* clones satisfy the requirement $H_1^{(j)} + K_1^{(j)} > 0$).

To determine α, we note that the self-consistent equation for the probability not to observe a clone in this model is

$$1 - \alpha = \sum_{\vec{N}_{1,b}} \mathcal{F}(\vec{N}_{1,b}|\left\{\vec{N}_1\right\}, q)\mathrm{Bin}(0; q, H_{1,b})\mathrm{Bin}(0; q, K_{1,b}), \tag{3}$$

where $\mathrm{Bin}(0; q, X)$ is the Binomial probability to sample 0 cells from X cells with the success rate q. Substituting in $\mathcal{F}$ from Eq. (1), after simplification we obtain

$$\alpha = \frac{1-\beta_0}{1+\beta_1-\beta_0}, \tag{4}$$

where $\beta_{0,1}$ take the forms:

$$\beta_0 = [P_0(0,q)]^2, \quad \beta_1 = \frac{1}{N}\sum_j P_0(H_1^{(j)}, q)P_0(K_1^{(j)}, q), \tag{5}$$

and with $P_0(X_1, q)$ being:

$$P_0(X_1, q) = \sum_{X_{1,b}} P(X_{1,b}|X_1, q)\mathrm{Bin}(0; q, X_{1,b}) = \begin{cases} \sum_{X_{1,b}>=X_1} \binom{X_{1,b}}{X_1} q^{X_1+1}(1-q)^{2X_{1,b}-X_1} & \text{if } X_1 \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Note that $P_0(X_1, q)$ is negligible when $X_1 q \gg 1$. This property can be used to speed up the numerical computation of β_1 and β_2. In Eq. (4), we can see that when $\beta_1 \ll 1$ then $\alpha \approx 1$. This result states that an observed clone is likely to be re-observed if re-sampled from the bone marrow, which agrees well with our intuition. Using the 414 clones observed in 1T in the experiment, we inferred that $\alpha = 0.887$.

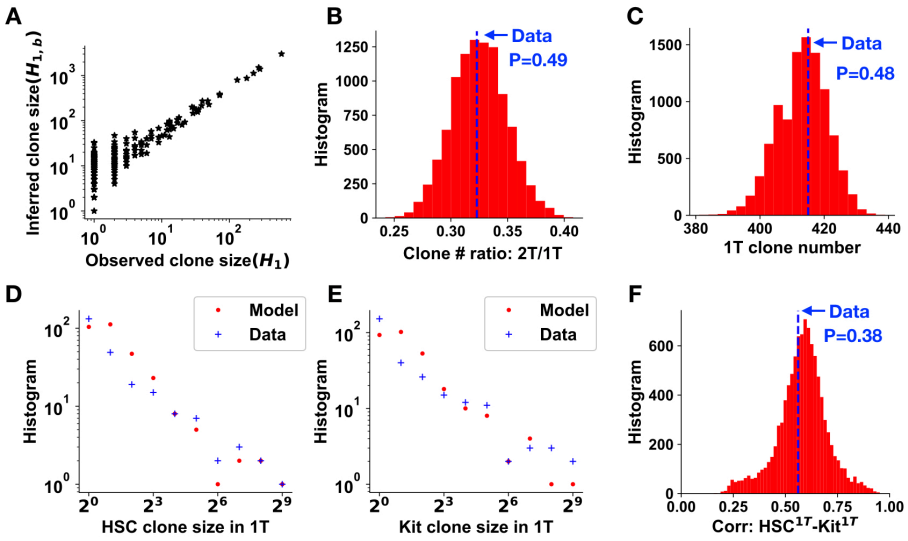
To derive Eq. (2), we invoke Bayes' law, which relates the desired distribution $P(X_{1,b}|X_1, q)$ to the observed distribution $P(X_1|X_{1,b}, q)$ through the relation,

$$P(X_{1,b}|X_1, q) = \frac{P(X_1|X_{1,b}, q)P(X_{1,b})}{P(X_1|q)}.$$

From the structure of the PDGM (see Supplementary Methods Table I), $P(X_1|X_{1,b}, q)$ is a binomial distribution. We assume a naive prior $P(X_{1,b}) = \text{const}$, and the quantity $P(X_1|q)$ is a normalization constant. Using the binomial formula and normalizing the result, we obtain Eq. (2). Note that the distribution $P(X_{1,b}|X_1, q)$ in Eq. (2) takes the form of a negative binomial if we perform the transformation $X_1 \to X_1 - 1$ and $X_{1,b} \to X_{1,b} - 1$.

D. Model implementation

The PDGM defines the joint distribution over the number of cells derived from an engrafted HSC in 1T (Supplementary Methods Table I). Since only $\bar{N}_1$, $\bar{N}_{2,1}$ and $\bar{N}_{2,2}$ are observable in experiments, we need only to calculate the



Supplementary Methods Fig. 2. **Model implementation and comparison with 1T statistics.** (**A**), sampling of the actual HSC clone size ($H_{1,b}$) in T1 from the PDGM, given the observed HSC clone size (H_1). The inference is based on Eq. (1). (**B**), Histogram of the model-predicted ratio between the observed 2T and 1T clone number. (**C**), Histogram of the model-predicted 1T clone number. (**D**), Histogram of the observed *vs.* model-predicted HSC clone sizes (H_1) in 1T. (**E**), Histogram of the observed *vs.* model-predicted Kit clone sizes (K_1) in 1T. (**F**), Histogram of the model-predicted Pearson correlation between 1T HSC clone size in 1T Kit+ clone size. The blue dashed line in each panel indicates the corresponding value of experimental data, and the corresponding p value is also indicated.

marginal distribution over these three variables. The observed data set consists of $N = 485$ detected clones (414 clones in 1T and 133 in 2T, with 61 shared clones) for which $U_i = \{\vec{N}_1^{(i)}, \vec{N}_{2,1}^{(i)}, \vec{N}_{2,2}^{(i)}\}$ are observed for the i-th clone. We define the data set as the joint list of observed values $\Omega_{\text{obs.}} = \{U_1, \ldots, U_N\}$. The PDGM predicts a hyperdistribution over Ω expected from the equipotency model. We make use of a Monte Carlo approach to sample the PDGM, as follows:

1. We sample the PDGM to generate 468 clones. This number results in an average of 414 clones observed in 1T and 133 clones observed in 2T (Supplementary Methods Fig. 2**BC**). The sampling procedure proceeds as in Supplementary Methods Fig. 1: a value of $N_{1,b}$ is sampled, followed by a value for $N_{1,c}$, and so on. For each clone we retain only $U_i = \{\vec{N}_1, \vec{N}_{2,1}, \vec{N}_{2,2}\}$ for the i-th clone, discarding latent model variables. The 485 simulated U_i provide a single realization of Ω.

2. We repeat Step 1 $L = 10^4$ times to generate a distribution over Ω.

E. Comparison with experimental data

To explore which observed aspects of the data are (in)consistent with the equipotency model defined by the PDGM, we defined multiple test statistics summarized in Supplementary Methods Table III from the data set $\Omega_{\text{obs.}}$. For each, we then evaluated the fraction of PDGM-sampled data sets Ω that generate values at least as extreme (larger or smaller) as seen in the data $\Omega_{\text{obs.}}$. Formally our approach defines a two-tailed p-value for each test statistic. See Supplementary Methods Fig. 2**F** for an example of the distribution of PDGM-sampled values compared to the observed value.

The assessed test statistics comprise of correlations between different cell counts (each clone $U_i = \{\vec{N}_1^{(i)}, \vec{N}_{2,1}^{(i)}, \vec{N}_{2,2}^{(i)}\}$ is represented by six numbers, allowing for fifteen pairwise correlations). In addition, we calculated higher-order relationships between the observed quantities by considering correlations betwen *clonal expansion* E_j and *clonal activity* A_j, for the first and second transplantation ($j = 1, 2$ respectively), defined as follows:

$$A_j = \frac{K_j + \epsilon}{H_j + \epsilon} \quad \text{for } j \in \{1, 2\}, \tag{7}$$

$$E_1 = \frac{N_1 + \epsilon}{1 + \epsilon}, \qquad E_2 = \frac{N_2 + \epsilon}{H_1 + \epsilon}, \qquad E_{2.j} = \frac{N_{2.j} + \epsilon}{H_1 + \epsilon} \text{ for } j \in \{1, 2\}. \tag{8}$$

Here, ϵ is a pseudocount used to avoid large contributions from clones with very small or no HSCs, and is set to be 1. We also assessed whether the 1T activity predicts HSC (Kit+ cells) specific expansion in 2T, which are defined as follows:

$$E_j^H = \frac{H_j + \epsilon}{H_1 + \epsilon}, \qquad E_j^K = \frac{K_j + \epsilon}{H_1 + \epsilon}, \text{ for } j \in \{2, 2.1, 2.2\}. \tag{9}$$

To this end, we partitioned clones according to their activities into active clones (top 40% values of A_1) and inactive clones (bottom 60% values of A_1), and evaluated the statistical difference of the corresponding T2 expansion (denoted as E_j^{ac} for active clones and E_j^{in} for inactive ones) for these two groups of clones, in terms of both the difference of average expansion and of the standard deviation of expansion (Supplementary Methods Table III).

Comparison	Test statistic	Experimental value	P value
HSC^{1T}–HSC^{2T}	$C(H_1, H_2)$	0.74	0.44
HSC^{1T}–Kit^{2T}	$C(H_1, K_2)$	0.54	0.13
Kit^{1T}–Kit^{2T}	$C(K_1, K_2)$	0.24	0.14
Kit^{1T}–HSC^{2T}	$C(K_1, H_2)$	0.27	0.16
1T–2T	$C(H_1 + K_1, H_2 + K_2)$	0.44	0.13
HSC^{2T}–Kit^{2T}	$C(H_2, K_2)$	0.91	0.12
Exp^{1T}–Exp^{2T}	$C(E_1, E_2)$	0.03	0.22
Act^{1T}–Act^{2T}	$C(A_1, A_2)$	0.04	0.08
Exp^{1T}–Act^{2T}	$C(E_1, A_2)$	0.58	< 0.0001
Act^{1T}–Exp^{2T}	$C(A_1, E_2)$	-0.08	< 0.0001
HSC^{R1}–Kit^{R1}	$C(H_{2.1}, K_{2.1})$	0.78	0.44
HSC^{R1}–HSC^{R2}	$C(H_{2.1}, H_{2.2})$	0.83	0.027
HSC^{R1}–Kit^{R2}	$C(H_{2.1}, K_{2.2})$	0.76	0.053
Kit^{R1}–Kit^{R2}	$C(K_{2.1}, K_{2.2})$	0.91	0.0004
R1–R2	$C(H_{2.1} + K_{2.1}, H_{2.2} + K_{2.2})$	0.91	0.0048
Exp^{R1}–Exp^{R2}	$C(E_{2.1}, E_{2.2})$	0.67	0.0013
$\mathsf{Exp}^{2T}_{inactive}$–$\mathsf{Exp}^{2T}_{active}$	$\mathsf{Mean}(E_2^{in})/\mathsf{Mean}(E_2^{ac})$	3.7	0.049
$\mathsf{Exp}^{2T}_{inactive}$–$\mathsf{Exp}^{2T}_{active}$	$\mathsf{S.t.d}(E_2^{in})/\mathsf{S.t.d.}(E_2^{ac})$	36	0.003
$(\mathsf{HSC\text{-}Exp})^{2T}_{inactive}$–$(\mathsf{HSC\text{-}Exp})^{2T}_{active}$	$\mathsf{Mean}(E_2^{H,in})/\mathsf{Mean}(E_2^{H,ac})$	2.5	0.056
$(\mathsf{HSC\text{-}Exp})^{2T}_{inactive}$–$(\mathsf{HSC\text{-}Exp})^{2T}_{active}$	$\mathsf{S.t.d}(E_2^{H,in})/\mathsf{S.t.d.}(E_2^{H,ac})$	19	0.017
$(\mathsf{Kit\text{-}Exp})^{2T}_{inactive}$–$(\mathsf{Kit\text{-}Exp})^{2T}_{active}$	$\mathsf{Mean}(E_2^{K,in})/\mathsf{Mean}(E_2^{K,ac})$	1.9	0.24
$(\mathsf{Kit\text{-}Exp})^{2T}_{inactive}$–$(\mathsf{Kit\text{-}Exp})^{2T}_{active}$	$\mathsf{S.t.d}(E_2^{K,in})/\mathsf{S.t.d.}(E_2^{K,ac})$	22	0.022

Supplementary Methods Table III. **List of comparisons between the equipotency model and observed data.** The p-value gives the fraction of PDGM samples of the data set Ω that gives a test statistic at least as extreme as observed in experiment (two-tailed test).

F. Supplemental discussion of PDGM results

In this section we provide a brief statistical interpretation of the analytical results, extending on the discussion in the manuscript. As can be seen from Supplementary Methods Table III, the equipotency model reproduces many features of the observed clone size distribution (represented by non-siginificant p-values), while significantly differing in several important ways. We found that the p-values were robust to variation in the PDGM parameter values.

Among the nontrivial predictions involving 2T clonal behavior, we tested how activity in 1T or 2T may predict expansion in 1T or 2T. Two couplings were found statistically significant (Supplementary Methods Table III, second block): 1) 1T expansion and 2T activity ($C = 0.58, p < 0.0001$); 2) 1T activity and 2T expansion ($C = -0.08, p < 0.0001$). More detailed analysis revealed that the high correlation between 1T expansion and 2T activity is contributed mainly by a single clone with significantly large expansion in 1T and activity in 2T, thus not reflecting a general trend

in the data. The negative relationship between clonal activity in 1T and expansion in 2T is robust, as shown by the statistical difference of the cell-type specific expansion between active and inactive clones (Supplementary Methods Table III, bottom block).

II. ERROR CORRECTION OF CLONAL BARCODES

In this supplemental note we describe the characterization and correction of errors in clonal barcodes.

In this study, each cell profiled by scRNA-Seq is detected to express a set of clonal barcodes stably integrated following delivery by lentiviral infection. A cell may express more than one barcode reflecting multiplicity of lentiviral infection. In ideal data, all cells from the same clone will express the exact same set of barcode sequences. In practice, there are three sources of error in clonal barcodes (hereafter 'BCs'): 1) the readout of a given BC sequence may contain sequencing errors; 2) for a given cell, certain BCs may fail to be detected by scRNA-Seq; and 3) cell doublets or droplet emulsion instability can also lead to an artificial set of BCs for a single cell. The problem of doublets is partially addressed by Scrublets [3] in early steps of data analysis.

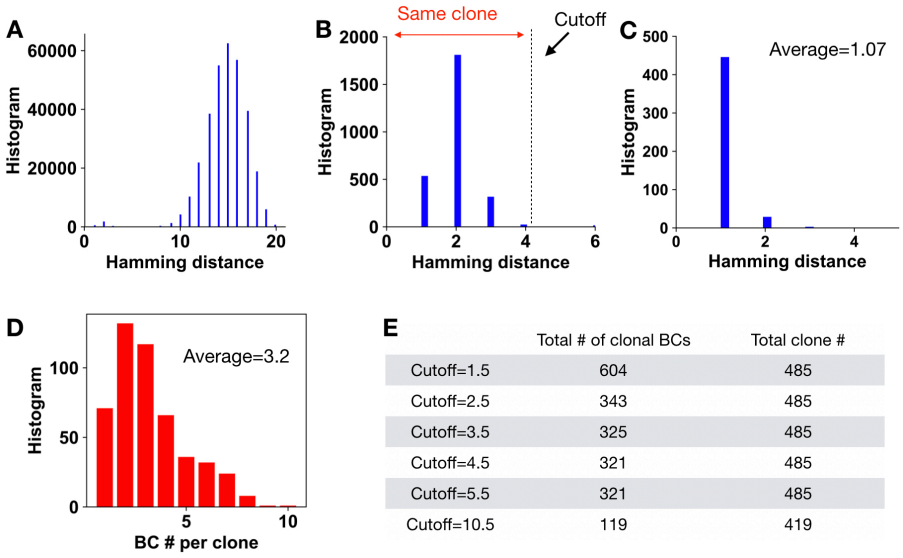
To overcome the first error (sequencing errors), we assigned cells to the same clone if they expressed BCs differing by a Hamming distance $D \leq 4$. This is justified by considering the Hamming distance histogram between any pair of distinct BCs in our 1T-2T dataset (Supplementary Methods Fig. 3**A,B**). For randomly-generated barcodes, such a histogram is expected to follow a unimodal binomial distribution with a peak at $0.75N$ (with N being the length of the random barcode, which is 29 here). We detected a second minor mode of the distribution at $D = 2$, which is understood to arise from sequencing error. A clear separation between the two modes occurs at $D = 4$. Further, for each of the clones defined in this manner, we identified a single BC sequence with the most UMI-corrected read counts. We expected this BC sequence to represent the error-free (or original) sequence, from which other sequences in the clone were derived. Consistent with this expectation, we found that the average distance between original BCs and other BCs within the each clone is 1.07 (Supplementary Methods Fig. 3**C**), indicating that most BC errors corresponding to just over one sequencing error on average. By contrast, the average distance between two random BCs within each same clone is larger 1.94 (Supplementary Methods Fig. 3**B**). This correction procedure collapsed 799 observed unique BCs into 321 clones.

To address the second error (barcode drop-out), we take the following two approaches. Although the identified clone number differs greatly, the pattern of clonal behavior across 1T and 2T remains the same.

- **No correction:** only cells with a fully identical set of BCs are classified as belonging to the same clone (after error correction as discussed above). This approach leads to 485 clones. The number of cells and shared clones for each sample set is summarized in Supplementary Methods Table IV, which is used in our paper. This data can be accessed in the attached *data.txt* file, which include, for each clone, the number of HSCs and Kit+ cells in T1, T2-R1, and T2-R2.
- **Dropout correction:** we classify two cells as belonging to the same clone if they share a significant number of BCs. Specifically, if cell A has N_A different BCs and cell B has N_B different BCs, these two cells are classified as from the same clone if the shared BC number N_S satisfies $N_S/\max(N_A, N_B) > 0.65$. Although the total number of clones drops to 258 (Supplementary Methods Table IV), very similar p values are obtained for different comparisons (Supplementary Methods Table V), demonstrating the robustness of our results to different clustering schemes. The insensitivity of the results to dropout correction can be understood by the fact that this clustering only affects small clones, which are more vulnerable to BC dropout or other sampling issues.

Although the number of unique clonal BCs is quite sensitive to the threshold of Hamming distance, we found that the number of identified clones is actually quite robust over a range of Hamming distance threshold (Supplementary Methods Fig. 3**E**). This is because each cell may carry multiple BCs, and errors in identifying individual BCs may not translate to the errors in identifying individual clones. Particularly, setting Hamming distance threshold to be 3 or 4 gives the same clonal annotation (Supplementary Methods Table IV), hence also all the clonal correlations and their respective p values (Supplementary Methods Table V).

[1] Christopher M Bishop. *Pattern recognition and machine learning.* springer, 2006.



E

	Total # of clonal BCs	Total clone #
Cutoff=1.5	604	485
Cutoff=2.5	343	485
Cutoff=3.5	325	485
Cutoff=4.5	321	485
Cutoff=5.5	321	485
Cutoff=10.5	119	419

Supplementary Methods Fig. 3. **Clustering BCs into the same clone**. (**A**), Histogram of Hamming distances between all pairs of distinct BCs. (**B**), The zoom-in version of **A** at small distances. BCs within a Hamming distance 4.5 are clustered within the same clone. (**C**), Histogram of Hamming distances between the identified "true" clonal BCs and other BCs within the same clone. (**D**), Histogram of the number of BCs per clone (each cell may carry several BCs). (**E**), Effect of the cutoff Hamming distance on the identified clonal BC number and clone number. No dropout correction is made here. While the number of clonal BCs are sensitive to this cutoff, the actual clone number is rather robust, as each clone may carry several BCs.

[2] GA Colvin, JF Lambert, M Abedi, CC Hsieh, JE Carlson, FM Stewart, and PJ Quesenberry. Murine marrow cellularity and the concept of stem cell competition: geographic and quantitative determinants in stem cell biology. *Leukemia*, 18(3):575, 2004.

[3] Samuel L Wolock, Romain Lopez, and Allon M Klein. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell systems*, 2019.

Cell number

	HSC number	Kit+ cell number	Total cell number
Primary transplantation (1T)	2812	4630	7442
Secondary transplantation (2T)	8818	9992	18810
Secondary transplantation (2T-R1)	4868	3881	8749
Secondary transplantation (2T-R2)	3950	6111	10061

Number of shared clones (no correction)

	1T	2T	2T-R1	2T-R2
1T	414	62	47	42
2T	—	133	98	77
2T-R1	—	—	98	42
2T-R2	—	—	—	77

Number of shared clones (dropout correction)

	1T	2T	2T-R1	2T-R2
1T	239	51	37	37
2T	—	70	51	47
2T-R1	—	—	51	28
2T-R2	—	—	—	47

Supplementary Methods Table IV. **Summary of the serial/parallel transplantation data.** Only cells carrying BCs are shown here, and these barcoded cells only represent 76% of all sequenced cells. In the middle and lower panel, the diagonal terms indicate the identified clone number of the corresponding sample.

Comparison	Test Statistic	Exp. value (A)	P-value (A)	Exp. value (B)	P-value (B)
HSC^{1T}–HSC^{2T}	$C(H_1, H_2)$	0.74	0.44	0.74	0.41
HSC^{1T}–Kit^{2T}	$C(H_1, K_2)$	0.54	0.13	0.53	0.10
Kit^{1T}–Kit^{2T}	$C(K_1, K_2)$	0.24	0.14	0.22	0.11
Kit^{1T}–HSC^{2T}	$C(K_1, H_2)$	0.27	0.16	0.25	0.13
1T–2T	$C(H_1 + K_1, H_2 + K_2)$	0.44	0.13	0.43	0.11
HSC^{2T}–Kit^{2T}	$C(H_2, K_2)$	0.91	0.12	0.90	0.16
Exp^{1T}–Exp^{2T}	$C(E_1, E_2)$	0.03	0.22	0.02	0.50
Act^{1T}–Act^{2T}	$C(A_1, A_2)$	0.04	0.08	0.03	0.12
Exp^{1T}–Act^{2T}	$C(E_1, A_2)$	0.58	< 0.0001	0.58	< 0.0001
Act^{1T}–Exp^{2T}	$C(A_1, E_2)$	-0.08	< 0.0001	-0.10	0.0001
HSC^{R1}–Kit^{R1}	$C(H_{2.1}, K_{2.1})$	0.78	0.44	0.77	0.50
HSC^{R1}–HSC^{R2}	$C(H_{2.1}, H_{2.2})$	0.83	0.027	0.82	0.03
HSC^{R1}–Kit^{R2}	$C(H_{2.1}, K_{2.2})$	0.76	0.053	0.75	0.066
Kit^{R1}–Kit^{R2}	$C(K_{2.1}, K_{2.2})$	0.91	0.0004	0.91	0.0005
R1–R2	$C(H_{2.1} + K_{2.1}, H_{2.2} + K_{2.2})$	0.91	0.0048	0.90	0.0067
Exp^{R1}–Exp^{R2}	$C(E_{2.1}, E_{2.2})$	0.67	0.0013	0.67	0.0022
$\mathsf{Exp}^{2T}_{inactive}$–$\mathsf{Exp}^{2T}_{active}$	Mean(E_2^{in})/Mean(E_2^{ac})	3.7	0.049	5.6	0.03
$\mathsf{Exp}^{2T}_{inactive}$–$\mathsf{Exp}^{2T}_{active}$	S.t.d(E_2^{in})/S.t.d.(E_2^{ac})	36	0.003	43	0.0035
$(\mathsf{HSC\text{-}Exp})^{2T}_{inactive}$–$(\mathsf{HSC\text{-}Exp})^{2T}_{active}$	Mean($E_2^{H,in}$)/Mean($E_2^{H,ac}$)	2.5	0.056	3.4	0.046
$(\mathsf{HSC\text{-}Exp})^{2T}_{inactive}$–$(\mathsf{HSC\text{-}Exp})^{2T}_{active}$	S.t.d($E_2^{H,in}$)/S.t.d.($E_2^{H,ac}$)	19	0.017	21	0.026
$(\mathsf{Kit\text{-}Exp})^{2T}_{inactive}$–$(\mathsf{Kit\text{-}Exp})^{2T}_{active}$	Mean($E_2^{K,in}$)/Mean($E_2^{K,ac}$)	1.9	0.24	2.8	0.15
$(\mathsf{Kit\text{-}Exp})^{2T}_{inactive}$–$(\mathsf{Kit\text{-}Exp})^{2T}_{active}$	S.t.d($E_2^{K,in}$)/S.t.d.($E_2^{K,ac}$)	22	0.022	26	0.024

Supplementary Methods Table V. **List of comparisons between the equipotency model and observed data.** The p-value gives the fraction of PDGM samples of the data set Ω that gives a test statistic at least as extreme as observed in experiment (two-tailed test). Columns labeled (A) show results without barcode dropout correction (reproduced from Supplementary Methods Table III for comparison). Columns labeled (B) show results after barcode dropout correction (see supplemental text).