Supplemental section

Stem cell dynamics in mouse hair follicles: a story from cell division counting and single cell lineage tracing

Ying V. Zhang, Brian S. White, David I. Shalloway, and Tudorita Tumbar

Department of Molecular Biology and Genetics,

Cornell University, Ithaca, NY 14853, USA

I. ANALYSIS OF H2B-GFP DIVISION DATA

A. Extraction of H2B-GFP division data

From the raw FACS data for mouse skin sample i ($1 \leq i \leq N_{\text{skin samples}} = 3$) we use a semi-automated approach to extract the proportion $p^i_{t_0 \to t_1, n}$ of labeled bulge cells at time t_1 (following doxycycline induction at time t_0) that have divided n times (Fig S1): we select live CD34$^+$/α6-integrin$^+$ cells as described [1] and export the GFP fluorescence values for each of the N_{events} cells in the selected subpopulation from FlowJo (Tree Star, Inc.), define $x^i_{t_0 \to t_1, j}$ ($1 \leq j \leq N_{\text{events}}$) by applying a logicle [2] transformation (having a linear regime near zero and a logarithmic region away from zero) to each of the exported values, and determine the $\vec{p}^{\,i}_{t_0 \to t_1}$ as the mixing coefficients of a Gaussian mixture model fit to the $\vec{x}^{\,i}_{t_0 \to t_1}$ via expectation-maximization (EM) [3]. In this case, the likelihood maximized is a sum evaluating a mixture of k Gaussian probability distribution functions $n[\cdot|\mu^i_n, (\sigma^i_n)^2]$ ($0 \leq n \leq k-1 \equiv \text{dim}$) with mean μ^i_n and variance $(\sigma^i_n)^2$ at each of the logicle-transformed H2B-GFP data values $x^i_{t_0 \to t_1, j}$:

$$L(\vec{p}^{\,i}_{t_0 \to t_1} | \vec{x}^{\,i}_{t_0 \to t_1}) = \sum_{j=1}^{N_{\text{events}}} \sum_{n=0}^{k-1} p^i_{t_0 \to t_1, n}\, n[x^i_{t_0 \to t_1, j} | \mu^i_n, (\sigma^i_n)^2] \,. \tag{S1}$$

The gaussians are assumed ordered according to their means, with μ^i_0 the largest mean fluorescence and $\mu^i_{k-1} \equiv \mu^i_{\text{dim}}$ the smallest mean fluorescence, corresponding to the "dim" peak (see below). Following standard practice [4–6], we initialize the EM algorithm using k-means [7], whose number k and location of clusters are themselves manually initialized by visualizing a histogram of the logicle-transformed H2B-GFP data. This computational approach is a simplified, univariate version of techniques for analyzing the full multivariate FACS data space by fitting mixtures of Gaussian [4–6, 8], skew normal [6], t [6, 8], or skew t [6] distributions using EM [4, 6, 8] or Markov chain Monte Carlo (MCMC) [5] algorithms, where the number of components k in the mixture may be determined by automated model selection criteria [3], such as Akaike Information Criterion (AIC) [6], Bayesian Information Criterion (BIC) [5, 6, 8], or related approaches [6].

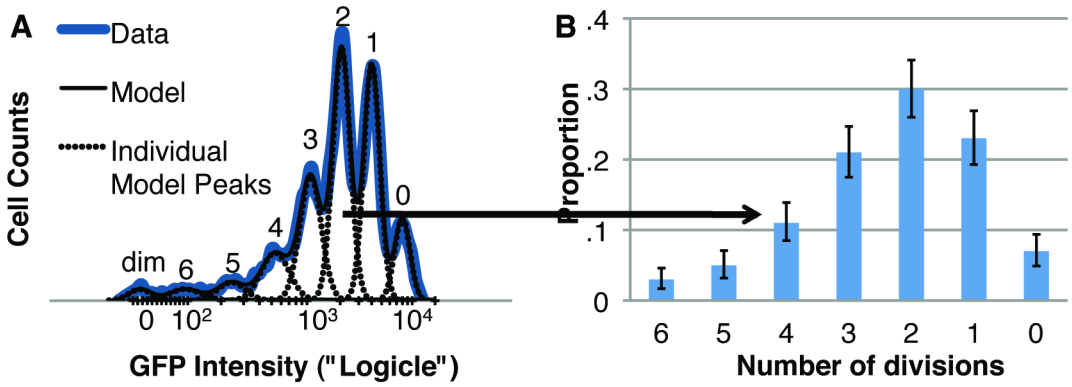


FIG. S1: Extracting data from GFP experiments. (a) Raw data (heavy blue line) is fit to a Gaussian mixture (thin black line). Numbers above peaks correspond to number of divisions. Dim indicates highly proliferative cells or unlabeled cells (see text). (b) Individual Gaussians within the mixture [dashed lines in (a)] represent a population of cells that have undergone the same number of divisions. Error bars are 90% credible sets.

B. Likelihood function for H2B-GFP division data

A likelihood function for the H2B-GFP division data needs to reflect two sources of potential variation—that between skin samples and that due to error fitting the Gaussian mixture model. In the vicinity of the maximum likelihood parameter estimates, the likelihood function of Eq. (S1), governing the fit of the Gaussian mixture model to the data, may be approximated by the multivariate Gaussian distribution [9, 10]

$$L(\vec{p}^{\,i}_{t_0\to t_1}|\vec{x}^{\,i}_{t_0\to t_1}) \propto \exp[-(\vec{p}^{\,i}_{t_0\to t_1} - \vec{p}^{\,i*}_{t_0\to t_1}) \cdot (\Sigma^i)^{-1} \cdot (\vec{p}^{\,i}_{t_0\to t_1} - \vec{p}^{\,i*}_{t_0\to t_1})/2] \,, \tag{S2}$$

where $\vec{p}^{\,i*}_{t_0\to t_1}$ are the maximum likelihood estimates (MLEs) of the $\vec{p}^{\,i}_{t_0\to t_1}$ determined from the fit and the covariance matrix

$$\Sigma^i = -\, H^{-1}\{\log[L(\vec{p}^{\,i}_{t_0\to t_1}|\vec{x}^{\,i}_{t_0\to t_1})]\}\big|_{\vec{p}^{\,i}_{t_0\to t_1}=\vec{p}^{\,i*}_{t_0\to t_1}}$$

is calculated in terms of the inverse of the Hessian matrix H of the log likelihood function

$$H\{\log[L(\vec{p}^{\,i}_{t_0\to t_1}|\vec{x}^{\,i}_{t_0\to t_1})]\}_{qr} \equiv \frac{\partial^2 \log[L(\vec{p}^{\,i}_{t_0\to t_1}|\vec{x}^{\,i}_{t_0\to t_1})]}{\partial p^i_{t_0\to t_1,q}\partial p^i_{t_0\to t_1,r}} \,.$$

Eq. (S2) characterizes the error due to the fit, but since it allows negative proportions it is clearly an approximation to the likelihood, which would not allow negative proportions.

We seek as an overall likelihood function $L(\vec{p}_{t_0\to t_1}|X_{t_0\to t_1})$ for the data $X_{t_0\to t_1}$ across all skin samples a compound distribution that accommodates both inter- and intra-skin sample variation, where the latter is captured by Eq. (S2). However, Eq. (S2) is an approximation that, in principle, is inadequate because it allows logically-incoherent negative proportions. An intuitive appeal to a Gaussian distribution in the *logarithm* of the proportions is still inadequate: though it could represent the intra-skin sample variation of non-negative proportions, it would remain to introduce inter-skin sample variation about these values, which could again force them below zero. As a recourse and for mathematical convenience, we use a Dirichlet distribution to approximate both sources of variation while respecting the non-negativity constraints. We assume that this Dirichlet distribution approximates the true error model, in which the realized $\vec{p}_{t_0\to t_1}$ result from fitting error (from the Gaussian mixture) introduced to a set of proportions sampled from an unknown distribution. As such, it can not be conveniently represented in analytical form. Instead we generate samples consistent with the unknown error model and fit a Dirichlet distribution to them. Given our limited knowledge of the true proportion distribution, we make no assumptions other than that it yields one of the experimentally-realized $\vec{p}^{\,i*}_{t_0\to t_1}$, each with equal probability. Therefore, to generate samples from the error model $p^{\text{true}}(\vec{p}_{t_0\to t_1})$ we choose one of the experimentally-realized $\vec{p}^{\,i*}_{t_0\to t_1}$ with equally probability and then introduce fitting error by sampling the corresponding Gaussian distribution of Eq. (S2) with mean $\vec{p}^{\,i*}_{t_0\to t_1}$. [We discard any negative proportions sampled from the Gaussian distribution.] In order to fit the approximate error model represented by a Dirichlet distribution $\widetilde{p}(\vec{p}_{t_0\to t_1}|\vec{\alpha}_{t_0\to t_1}) \equiv Dir(\vec{p}_{t_0\to t_1}|\vec{\alpha}_{t_0\to t_1})$ to these samples [assumed drawn from the unknown, but fixed error model $p^{\text{true}}(\vec{p}_{t_0\to t_1})$], we minimize the Kullback-Leibler divergence [3] between the true and approximate error models with respect to the $\vec{\alpha}_{t_0\to t_1}$

$$KL(p^{\text{true}}||\widetilde{p}) \equiv \int p^{\text{true}}(\vec{p}_{t_0\to t_1})\, \log \frac{p^{\text{true}}(\vec{p}_{t_0\to t_1})}{\widetilde{p}(\vec{p}_{t_0\to t_1}|\vec{\alpha}_{t_0\to t_1})}\, d\vec{p}_{t_0\to t_1}$$

to obtain $\vec{\alpha}^{*}_{t_0\to t_1}$ and then define $L(\vec{p}_{t_0\to t_1}|X_{t_0\to t_1}) = Dir(\vec{p}_{t_0\to t_1}|\vec{\alpha}^{*}_{t_0\to t_1})$. Since $p^{\text{true}}(\vec{p}_{t_0\to t_1})$ is fixed, minimizing the Kullback-Leibler divergence is equivalent to maximizing

$$\int p^{\text{true}}(\vec{p}_{t_0\to t_1})\, \log[\widetilde{p}(\vec{p}_{t_0\to t_1}|\vec{\alpha}_{t_0\to t_1})]\, d\vec{p}_{t_0\to t_1}\ .$$

We can approximate this integral via importance sampling (see Section II)

$$\int p^{\text{true}}(\vec{p}_{t_0\to t_1})\, \log[\widetilde{p}(\vec{p}_{t_0\to t_1}|\vec{\alpha}_{t_0\to t_1})]\, d\vec{p}_{t_0\to t_1} \approx \sum_m \log[\widetilde{p}(\vec{p}_{t_0\to t_1,m}|\vec{\alpha}_{t_0\to t_1})]$$

wherein the $\vec{p}_{t_0\to t_1,m}$ are sampled from the error model $p^{\text{true}}(\vec{p}_{t_0\to t_1})$ as described above. [Do not confuse the m^{th} sample k-vector of proportions $\vec{p}_{t_0\to t_1,m}$ with the (scalar) proportion of the n^{th} peak $p^i_{t_0\to t_1,n}$ (from skin sample i) or $p_{t_0\to t_1,n}$ (in general).] Therefore, in order to fit the Dirichlet distribution to the data sampled from the error model, we maximize the sum of the logarithm of the Dirichlet probability distribution function evaluated at those samples. We consider importance sampling converged once the mean-normalized difference between the current and previous estimates of the Dirichlet parameters is below 0.001. We draw random samples for $100,000 * N_{\text{skin samples}}$ iterations, where $N_{\text{skin samples}}$ is the number of skin samples, and then check for convergence every $10,000 * N_{\text{skin samples}}$ iterations.

C. Bulge fold change and average number of divisions

One difficulty in the fold change calculation is the peak indicated as "dim" in Fig. S1. Waghmare et al. [1] noted the presence of such a near-zero H2B-GFP intensity peak, even for mice that were induced with doxycycline too soon before being sacrificed to exhibit peaks of H2B-GFP diluted into that range. They surmised that this "unlabeled peak" was caused by mosaicism in transgene expression. However, given the apparent presence of a peak immediately abutting it and, in this case, representing six divisions, it is likely that the dim peak in Fig. S1 contains both cells that were never labeled (due to mosaicism) and those whose label was diluted to or below the detection threshold through repeated division. It would be possible to differentiate between these two classes of cells by introducing a dynamical model. Instead, we consider the two extreme cases that provide lower and upper bounds for $fc_{\text{no loss}}$, the fold change under the hypothesis of no bulge cell loss. This upper bound will also serve as an upper bound on fc, the biologically-realized fold change with cell loss, as in Eq. (3).

To derive bounds, we recognize that the percentage of cells in the dim peak is fixed at $p_{t_0\to t_1,\text{dim}}$. A lower bound is obtained if the dim peak has only unlabeled cells. Since the $p_{t_0\to t_1,n}$ are intended to represent H2B-GFP labeled cells, we must subtract off the fraction $p_{t_0\to t_1,\text{dim}}$ of unlabeled cells, and adjust the proportions of the remaining peaks so that they sum to one: $p_{t_0\to t_1,n\neq\text{dim}} \to p_{t_0\to t_1,n\neq\text{dim}}/(1-p_{t_0\to t_1,\text{dim}})$. Proving the lower bound

$$\frac{1-p_{t_0\to t_1,\text{dim}}}{\sum_{n\neq\text{dim}} p_{t_0\to t_1,n}\cdot\left(\frac{1}{2}\right)^n} < \frac{1}{\sum_{n=0} p_{t_0\to t_1,n}\cdot\left(\frac{1}{2}\right)^n} = fc_{\text{no loss}} \tag{S3}$$

involves some algebraic manipulation. We begin by writing the denominator of $fc_{\text{no loss}}$ as $\sum_{n=0} p_{t_0\to t_1,n}\cdot\left(\frac{1}{2}\right)^n = p_{t_0\to t_1,\text{dim}}\cdot\left(\frac{1}{2}\right)^{\text{dim}} + \sum_{n\neq\text{dim}} p_{t_0\to t_1,n}\cdot\left(\frac{1}{2}\right)^n$. Multiplying the numerator and denominator of $fc_{\text{no loss}}$ by $1-p_{t_0\to t_1,\text{dim}}$, expanding products, and collecting terms then establishes the lower bound

$$\begin{aligned}
fc_{\text{no loss}} &= \frac{1}{\sum_{n=0} p_{t_0\to t_1,n}\cdot\left(\frac{1}{2}\right)^n} \\
&= \frac{1-p_{t_0\to t_1,\text{dim}}}{\left(1-p_{t_0\to t_1,\text{dim}}\right)\left(p_{t_0\to t_1,\text{dim}}\cdot\left(\frac{1}{2}\right)^{\text{dim}} + \sum_{n\neq\text{dim}} p_{t_0\to t_1,n}\cdot\left(\frac{1}{2}\right)^n\right)} \\
&= \frac{1-p_{t_0\to t_1,\text{dim}}}{\sum_{n\neq\text{dim}} p_{t_0\to t_1,n}\cdot\left(\frac{1}{2}\right)^n + p_{t_0\to t_1,\text{dim}}\left(\left(\frac{1}{2}\right)^{\text{dim}} - p_{t_0\to t_1,\text{dim}}\left(\frac{1}{2}\right)^{\text{dim}} - \sum_{n\neq\text{dim}} p_{t_0\to t_1,n}\cdot\left(\frac{1}{2}\right)^n\right)} \\
&> \frac{1-p_{t_0\to t_1,\text{dim}}}{\sum_{n\neq\text{dim}} p_{t_0\to t_1,n}\cdot\left(\frac{1}{2}\right)^n}\ .
\end{aligned}$$

The inequality holds because

$$\begin{aligned}
-\sum_{n\neq\text{dim}} p_{t_0\to t_1,n}\cdot\left(\frac{1}{2}\right)^n &< \left(\frac{1}{2}\right)^{\text{dim}} - p_{t_0\to t_1,\text{dim}}\left(\frac{1}{2}\right)^{\text{dim}} - \sum_{n\neq\text{dim}} p_{t_0\to t_1,n}\cdot\left(\frac{1}{2}\right)^n \\
&< \left(\frac{1}{2}\right)^{\text{dim}} - \left(\frac{1}{2}\right)^{\text{dim}}\sum_n p_{t_0\to t_1,n} \\
&= \left(\frac{1}{2}\right)^{\text{dim}} - \left(\frac{1}{2}\right)^{\text{dim}} \\
&= 0\ .
\end{aligned}$$

Therefore, the denominator is less than $\sum_{n\neq\text{dim}} p_{t_0\to t_1,n}\cdot\left(\frac{1}{2}\right)^n$ (though positive), which proves the inequality used in the lower bound.

The upper bound leverages both the fixed $p_{t_0\to t_1,\text{dim}}$ and the fact that the per cell H2B-GFP fluorescence of the cells in this peak is unknown. Though it could vary between cells, a worst case and a mathematical upper bound occurs when all cells in the dim peak tend towards infinite divisions. In this case their individual H2B-GFP fluorescence tends towards zero: $\left(\frac{1}{2}\right)^n \to 0$ as $n\to\infty$. Therefore, the term involving $p_{t_0\to t_1,\text{dim}}$ effectively drops out of the sum, which establishes the upper bound

$$\begin{aligned}
fc_{\text{no loss}} &= \frac{1}{\sum_{n=0} p_{t_0\to t_1,n}\cdot\left(\frac{1}{2}\right)^n} \\
&< \lim_{\text{dim divisions}\to\infty} \frac{1}{p_{t_0\to t_1,\text{dim}}\cdot\left(\frac{1}{2}\right)^{\text{dim divisions}} + \sum_{n\neq dim} p_{t_0\to t_1,n}\cdot\left(\frac{1}{2}\right)^n} \\
&= \frac{1}{\sum_{n\neq dim} p_{t_0\to t_1,n}\cdot\left(\frac{1}{2}\right)^n}\ .
\end{aligned} \tag{S4}$$

Combining Eqs. (S3) and (S4) provides the bounds on $fc_{\text{no loss}}$

$$\frac{1-p_{t_0\to t_1,\text{dim}}}{\sum_{n\neq\text{dim}} p_{t_0\to t_1,n}\cdot\left(\frac{1}{2}\right)^n} < fc_{\text{no loss}} = \frac{1}{\sum_{n=0} p_{t_0\to t_1,n}\cdot\left(\frac{1}{2}\right)^n} < \frac{1}{\sum_{n\neq dim} p_{t_0\to t_1,n}\cdot\left(\frac{1}{2}\right)^n}\,. \quad \text{(S5)}$$

We caution that the notion of the number of divisions tending towards infinity is a valid mathematical approach even though it is biologically inconsistent. In particular, given that we are working under the hypothesis of no cell loss, it is biologically impossible to have a small percentage $p_{t_0\to t_1,\text{dim}}$ of cells that have divided many times, while the remaining cells divide significantly fewer times. Biological plausibility would instead require cells with final division states intermediate between the two proliferative extremes. In short, the fold change calculation effectively counts cells, not number of divisions. Mathematically, the former can remain finite though the latter tends to infinity. Biologically, and under the hypothesis of no cell loss, this is not possible. Nevertheless, the mathematical bound is valid in establishing an upper bound on the biologically-realized fold change. Further, assuming that the fraction $p_{t_0\to t_1,\text{dim}}$ of cells in the dim peak is small, the assumption of division tending towards infinity does not lead to loose bounds: the ratio between the upper and lower bounds, $1/\left(1-p_{t_0\to t_1,\text{dim}}\right)$, is close to one, and Fig. S2 shows that the difference between the lower ("unlabeled") and upper ("highly proliferative") bounds is small. Finally, a more careful bound accounting for the finiteness of the cell cycle duration would not significantly change the result. For example, imposing a cell cycle time of 24 hours on the PD21-35 data allows 14 divisions. The effect of using an H2B-GFP content of $\left(\frac{1}{2}\right)^{14}$ as opposed to $\lim_{n\to\infty}\left(\frac{1}{2}\right)^n = 0$ would be negligible and not worth the additional biological assumption of a particular cell cycle time.

The above arguments do not invalidate the bound $fc_{\text{no loss}} > fc$ established in Eq. (3) for any particular $p_{t_0\to t_1,n}$. In fact, Eq. (S5) allows us to establish a bound on fc (i.e., with cell loss) that accounts for the uncertainty in the dim peak:

$$\begin{aligned} fc &= \frac{1-\left(N_1^l/N_0^b\right)\cdot\sum_{n=0} p^l_{t_0\to t_1,n}\cdot\left(\frac{1}{2}\right)^n}{\sum_{n=0} p_{t_0\to t_1,n}\cdot\left(\frac{1}{2}\right)^n} \\ &< \frac{1}{\sum_{n=0} p_{t_0\to t_1,n}\cdot\left(\frac{1}{2}\right)^n} \\ &< \frac{1}{\sum_{n\neq dim} p_{t_0\to t_1,n}\cdot\left(\frac{1}{2}\right)^n}\,. \end{aligned} \quad \text{(S6)}$$

When the population fold change is known, Eq. (2) may be inverted to provide a lower

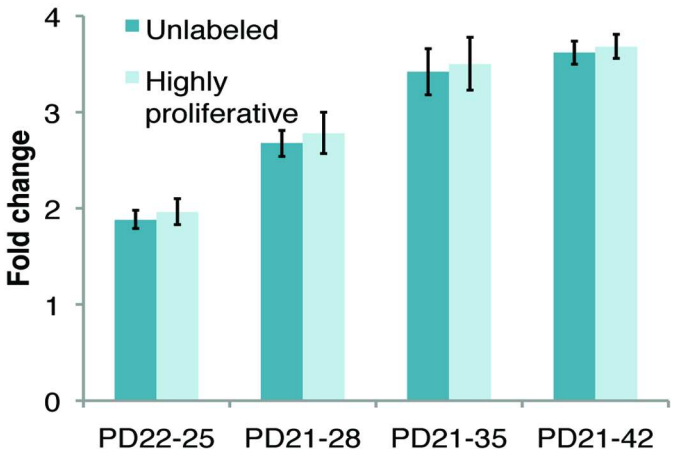


FIG. S2: Fold change calculated assuming the dim peak is comprised of unlabeled cells or of highly proliferative cells for time points indicated. Error bars are 90% credible sets.

bound on the fractional cell loss N_1^l/N_0^b from the bulge

$$\frac{N_1^l}{N_0^b} = \frac{1 - (N_1^b/N_0^b) \cdot \sum_{n=0} p_{t_0 \to t_1, n} \cdot \left(\frac{1}{2}\right)^n}{\sum_{n=0} p^l_{t_0 \to t_1, n} \cdot \left(\frac{1}{2}\right)^n} \,. \tag{S7}$$

When the fold change N_1^b/N_0^b and the *bulge* division probabilities $p_{t_0 \to t_1, n}$ are fixed, Eq. (S7) is minimized with $p^l_{t_0 \to t_1, 0} = 1$ (and hence $p^l_{t_0 \to t_1, n \neq 0} = 0$). Hence, the ratio of cells lost from the bulge with respect to the initial bulge population is bounded below by

$$\frac{N_1^l}{N_0^b} \geq 1 - (N_1^b/N_0^b) \cdot \sum_{n=0} p_{t_0 \to t_1, n} \cdot \left(\frac{1}{2}\right)^n \,. \tag{S8}$$

Multiplying by N_0^b/N_1^b instead gives the ratio with respect to the bulge population at the end of the chase (t_1)

$$\frac{N_1^l}{N_1^b} \geq (N_0^b/N_1^b) - \sum_{n=0} p_{t_0 \to t_1, n} \cdot \left(\frac{1}{2}\right)^n \,. \tag{S9}$$

The average number of divisions can also be calculated from the H2B-GFP data as

$$\sum_{n=0} n \, p_{t_0 \to t_1, n} \,.$$

Unlike the fold-change calculation, the average number of divisions is critically dependent on whether the dim peak contains unlabeled cells or highly replicative cells. For the purposes of this calculation, we assume that the dim peak contains only unlabeled cells. A more sophisticated calculation would incorporate a dynamical model (with its additional assumptions)

to infer the distribution of label within the dim peak, as mentioned above. The relative insensitivity of the fold change to this factor may make it a more biologically-meaningful statistic.

D. Estimating number of bulge cells via microscopy

Calculating the number of bulge cells per hair follicle is most straightforward before the cells begin their lateral migration. In that case, bulge cells form a closed cylinder around the base of the hair follicle and the number of bulge cells may be derived from simple geometry. We magnified a high-resolution image of a hematoxylin and eosin-stained bulge at PD25 (Fig. S5D in Ref. 11) to determine an inter-cell spacing of 5.32 μm, the number of layers of cells (14) in a bulge, and the layer-dependent bulge radii (which range from 4.25 μm at the base to 12 μm mid-bulge). Assuming that inter-cell spacing is isotropic (i.e., the same in all directions) allows the number of cells per layer to be calculated by dividing the bulge circumference at that layer by the inter-cell spacing. Summing across all layers gives a total number of bulge cells of ~ 150. Assuming bulge morphology and number of cells are similar between PD21 and PD25, we can multiply the fractional loss relative to number of bulge cells at PD21 (42%) by 150 to determine that ~ 63 bulge cells must have been lost during the first hair cycle.

II. ERROR ANALYSIS

We characterize the error in the experimentally-derived H2B-GFP peaks, the bulge fold change, and the average number of divisions using 90% credible sets [12]. Given data X and posterior distributions $\pi(\vec{\theta}|X)$ over parameters $\vec{\theta}$ and $\pi[h(\vec{\theta})|X] \equiv \pi(h|X)$ over a scalar function $h(\vec{\theta})$ of those parameters, a credible set A for $h(\vec{\theta})$ satisfies

$$\begin{aligned} P(h(\vec{\theta}) \in A|X) &= \int_A \pi(h|X)\, dh = \int_A \int \delta[h - h(\vec{\theta}')]\, \pi(h|X)\, dh\, d\vec{\theta}' \\ &= \int_B \pi(\vec{\theta}'|X)\, d\vec{\theta}' \,, \end{aligned}$$

where $\vec{\theta} \in B \rightarrow h(\vec{\theta}) \in A$ and $\delta(x)$ is the Dirac delta function. The set B defines an ellipsoid in parameter space, whereas the credible set A is a projection onto the one-dimensional space

spanned by $h(\vec{\theta})$. We choose this contiguous region such that $h(\vec{\theta})$ has the same probability of being above it as below it.

For situations in which the posterior distribution $\pi(\vec{\theta}|X)$ is known and its quantile function (i.e., the inverse of its cumulative distribution function) is simply calculated, the above approach is straightforward to apply. For our purposes, one or both of these conditions is frequently violated. Fortunately, in cases where the posterior distribution is unknown, we have access to a likelihood function $L(\vec{\theta}|X)$, which is related to the posterior distribution

$$\pi(\vec{\theta}|X) = \frac{L(\vec{\theta}|X)\, p(\vec{\theta})}{\int L(\vec{\theta}'|X)\, p(\vec{\theta}')\, d\vec{\theta}'} \equiv \frac{\tilde{\pi}(\vec{\theta}|X)}{\int \tilde{\pi}(\vec{\theta}'|X)\, d\vec{\theta}'}$$

through the prior distribution $p(\vec{\theta})$. Once the prior is specified, this allows us to use the general strategy of evaluating integrals via importance sampling [13].

Importance sampling is a computational technique for evaluating potentially high-dimensional integrals of the form

$$\langle h(\vec{\theta})\rangle = \int h(\vec{\theta})\, \pi(\vec{\theta}|X)\, d\vec{\theta} = \frac{\int h(\vec{\theta})\, \tilde{\pi}(\vec{\theta}|X)\, d\vec{\theta}}{\int \tilde{\pi}(\vec{\theta}'|X)\, d\vec{\theta}'}\,, \tag{S10}$$

involving distributions $\pi(\vec{\theta}|X)$ that can not be conveniently sampled. It instead relies on a sampling kernel $\mu(\vec{\theta})$ from which samples $\vec{\theta}_m$ can be drawn efficiently. Under mild conditions [13], a sum of the $h(\vec{\theta}_m)$ weighted by $w(\vec{\theta}_m) \equiv \tilde{\pi}(\vec{\theta}_m|X)/\mu(\vec{\theta}_m)$,

$$\bar{h}_N \equiv \frac{\sum_{m=1}^{N} h(\vec{\theta}_m)\, w(\vec{\theta}_m)}{\sum_{q=1}^{N} w(\vec{\theta}_q)}\,,$$

converges to Eq. (S10). Since convergence to the true value $\langle h(\vec{\theta})\rangle$ necessarily requires sampling the high-probability credible set A, the latter can be calculated as a side effect of the computation [13]. For example, the lower bound for a $1-\alpha$ credible set is any $h^{\rm lo}$ such that $\sum_{m:h(\vec{\theta}_m)\le h^{\rm lo}} w(\vec{\theta}_m)/\sum_{q=1}^{N} w(\vec{\theta}_q) \ge \alpha/2$ and $\sum_{m:h(\vec{\theta}_m)\ge h^{\rm lo}} w(\vec{\theta}_m)/\sum_{q=1}^{N} w(\vec{\theta}_q) \ge 1-\alpha/2$. We consider convergence obtained once the number of samples N is at least $100,000$ and exceeds the number required to maintain relative error below 1% at a 95% asymptotic confidence level [14]

$$N \ge \left(\frac{1.96}{0.005}\right)^2 \left(\frac{\sigma_{\bar{h}_N}}{\langle h(\vec{\theta})\rangle}\right)^2\,.$$

We approximate $\langle h(\vec{\theta})\rangle$ as $\bar{h}_N$ and the variance $\sigma^2_{\bar{h}_N}$ of the samples as [13, 15]

$$\hat{\sigma}^2_{\bar{h}_N} = \frac{\sum_{m=1}^{N}\left(h(\vec{\theta}_m) - \bar{h}_N\right)^2 w(\vec{\theta}_m)^2}{\left(\sum_{q=1}^{N} w(\vec{\theta}_q)\right)^2} \; .$$

We calculate credible sets for the division proportions $\vec{p}_{t_0\rightarrow t_1}$ and for bulge fold changes (fc) and average number of divisions based on the $\vec{p}_{t_0\rightarrow t_1}$ using the Dirichlet likelihood function $L(\vec{p}_{t_0\rightarrow t_1}|X_{t_0\rightarrow t_1}) = Dir(\vec{p}_{t_0\rightarrow t_1}|\vec{\alpha}^*_{t_0\rightarrow t_1})$ derived in Section I A. We take a uniform prior distribution $p(\vec{p}_{t_0\rightarrow t_1})$ so that the posterior distribution is a Dirichlet distribution: $L(\vec{p}_{t_0\rightarrow t_1}|X_{t_0\rightarrow t_1})\, p(\vec{p}_{t_0\rightarrow t_1}) \propto Dir(\vec{p}_{t_0\rightarrow t_1}|\vec{\alpha}^*_{t_0\rightarrow t_1}) = \pi(\vec{p}_{t_0\rightarrow t_1}|X_{t_0\rightarrow t_1}) \equiv \pi(\vec{\theta}|X)$. This is used as the importance sampling kernel $\mu(\vec{p}_{t_0\rightarrow t_1}) = Dir(\vec{p}_{t_0\rightarrow t_1}|\vec{\alpha}^*_{t_0\rightarrow t_1})$ in evaluating integrals such as

$$\frac{\int_B p_{t_0\rightarrow t_1,n}\, L(\vec{p}_{t_0\rightarrow t_1}|X_{t_0\rightarrow t_1})\, p(\vec{p}_{t_0\rightarrow t_1})\, d\vec{p}_{t_0\rightarrow t_1}}{\int L(\vec{p}_{t_0\rightarrow t_1}|X_{t_0\rightarrow t_1})\, p(\vec{p}_{t_0\rightarrow t_1})\, d\vec{p}_{t_0\rightarrow t_1}} = \int_B p_{t_0\rightarrow t_1,n}\, Dir(\vec{p}_{t_0\rightarrow t_1}|\vec{\alpha}^*_{t_0\rightarrow t_1})\, d\vec{p}_{t_0\rightarrow t_1}$$

and

$$\frac{\int_B fc(\vec{p}_{t_0\rightarrow t_1})\, L(\vec{p}_{t_0\rightarrow t_1}|X_{t_0\rightarrow t_1})\, p(\vec{p}_{t_0\rightarrow t_1})\, d\vec{p}_{t_0\rightarrow t_1}}{\int L(\vec{p}_{t_0\rightarrow t_1}|X_{t_0\rightarrow t_1})\, p(\vec{p}_{t_0\rightarrow t_1})\, d\vec{p}_{t_0\rightarrow t_1}} = \int_B fc(\vec{p}_{t_0\rightarrow t_1})\, Dir(\vec{p}_{t_0\rightarrow t_1}|\vec{\alpha}^*_{t_0\rightarrow t_1})\, d\vec{p}_{t_0\rightarrow t_1} \; .$$

---

[1] S. K. Waghmare, R. Bansal, J. Lee, Y. V. Yang, D. J. McDermitt, and T. Tumbar, EMBO J. **27**, 1309 (2008).

[2] D. R. Parks, M. Roederer, and W. A. Moore, Cytometry Part A **69A**, 541 (2006).

[3] C. M. Bishop, *Pattern recognition and machine learning* (Springer, New York, 2006).

[4] M. J. Boedigheimer and J. Ferbas, Cytometry Part A **73A**, 421 (2008).

[5] C. Chan, F. Feng, J. Ottinger, D. Foster, M. West, and T. B. Kepler, Cytometry Part A **73A**, 693 (2008).

[6] S. Pyne, X. Hu, E. Rossin, T. Lin, L. M. Maier, C. Baecher-Allan, G. J. McLachlan, P. Tamayo, D. A. Hafler, P. L. De Jager, et al., Proc. Natl. Acad. Sci. USA **106**, 8519 (2009).

[7] B. S. Everitt, S. Landau, and M. Leese, *Cluster Analysis* (Arnold, London, 2001).

[8] K. Lo, R. R. Brinkman, and R. Gottardo, Cytometry Part A **73A**, 321 (2008).

[9] M. H. Kalos and P. A. Whitlock, *Monte Carlo Methods, Volume I: Basics* (John Wiley & Sons, New York, 1986).

[10] P. M. Lee, *Bayesian Statistics: An Introduction* (Oxford University Press, New York, NY, 1989).

[11] Y. V. Zhang, J. Cheong, N. Ciapurin, D. J. McDermitt, and T. Tumbar, Cell Stem Cell **5**, 1 (2009).

[12] G. Casella and R. L. Berger, *Statistical Inference* (Duxbury Press, Australia, 2001), 2nd ed.

[13] J. Geweke, Econometrica **57**, 1317 (1989).

[14] T. Kloek and H. K. van Dijk, Econometrica **46**, 1 (1978).

[15] J. Geweke, *Contemporary Bayesian Econometrics and Statistics* (John Wiley & Sons, Hoboken, NJ, 2005).

Equations

The following equations appear in order as they appear for the first time in the text.

$$N_1^b$$

$$N_0^b$$

$$fc = \frac{N_1^b}{N_0^b}.$$

$$p_{t_0 \to t_1, n} \ (n \geq 0)$$

$$N_1^{tot}$$

$$N_1^l$$

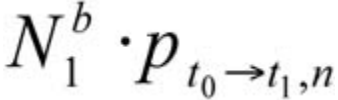
$$N_1^{tot} = N_1^b + N_1^l$$

$$N_1^b \cdot p_{t_0 \to t_1, 0}$$

$$N_1^b \cdot p_{t_0 \to t_1, 0} \cdot 1$$

$$N_1^b \cdot p_{t_0 \to t_1, 1}$$

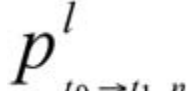
$$N_1^b \cdot p_{t_0 \to t_1, 1} \cdot \frac{1}{2}$$



$$\left(\frac{1}{2}\right)^{n}$$

$$N_1^b \cdot p_{t_0 \to t_1, n} \cdot \left(\frac{1}{2} \right)^n$$

$$N_1^b \cdot p_{t_0 \to t_1,0} \cdot 1 + N_1^b \cdot p_{t_0 \to t_1,1} \cdot \frac{1}{2} + N_1^b \cdot p_{t_0 \to t_1,2} \cdot \left(\frac{1}{2}\right)^2 + \ldots = N_1^b \cdot \sum p_{t_0 \to t_1,n} \cdot \left(\frac{1}{2}\right)^n \cdot$$



$$N_1^l \cdot \sum p^l_{t_0 \to t_1, n} \cdot \left(\frac{1}{2}\right)^n$$

$$N_0^b \cdot 1$$

$$N_0^b \cdot 1 = N_1^b \cdot \sum p_{t_0 \to t_1, n} \cdot \left(\frac{1}{2}\right)^n + N_1^l \cdot \sum p^l_{t_0 \to t_1, n} \cdot \left(\frac{1}{2}\right)^n$$

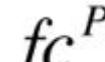
$$fc = \frac{N_1^b}{N_0^b} = \frac{1 - \left(N_1^l / N_0^b\right) \sum_n p^l_{t_0 \to t_1, n} \cdot \left(\frac{1}{2}\right)^n}{\sum_n p_{t_0 \to t_1, n} \cdot \left(\frac{1}{2}\right)^n}.$$

$$f_{no\ loss} = \frac{1}{\sum_n p_{t_0 \to t_1, n} \cdot \left(\frac{1}{2}\right)^n} > \frac{1 - \left(N_1^l / N_0^b\right) \sum_n p^l_{t_0 \to t_1, n} \cdot \left(\frac{1}{2}\right)^n}{\sum_n p_{t_0 \to t_1, n} \cdot \left(\frac{1}{2}\right)^n} = f_c$$

$$N_1^l = 0$$


fc
PD22–25
no loss


fc
PD21–28
no loss


fc
PD 21–35
no loss

$$N_1^l \,/\, N_0^b$$

$$\frac{N_1^l}{N_0^b} \geq 1 - \left(N_1^b / N_0^b\right) \cdot \sum_n p_{t_0 \to t_1, n} \cdot \left(\frac{1}{2}\right)^n .$$

$$N_1^b / N_0^b = 2$$

$$N_1^l / N_0^b > 0.42$$