doi:10.1038/nature13824

SUPPLEMENTARY METHODS

Estimation of total number of clones from single cell data. This supplement derives the expression provided in the main text (methods) for the probability, $P(N|b,c)$, of the peripheral blood (PB) being supported by N clones, given that we observed b distinct clonal barcodes among c cells randomly sampled. The general strategy is to first calculate the probability $P(b|N,c)$ of observing b barcodes assuming N clonal barcodes in total, and to then apply Bayes' theorem to obtain $P(N|b,c)$.

As noted in the methods section, we assume that clones are of uniform size (each consisting of approximately $1/N$ of the total number of cells in the PB). This approximation provides a lower limit on estimates of N. Formally, although the experiment involves sampling cells without replacement, we can make the approximation that sampling occurs with replacement since only ~100 cells are sampled out of 10^5-10^6 PB cells in each mouse. Enumerating the N clonal barcodes as 1,2,…,N, the probability of sampling x_1 cells with barcode 1, x_2 cells with barcode 2, etc, is a multinomial, $P(x_1, \ldots, x_N|N, c) = \frac{c!}{x_1! \cdots x_N!} N^{-c}$.

Since the aim of the analysis to relate the number of barcodes b observed in a given experiment to the total number of barcodes N, the precise identity of clonal barcodes (1,…,N) is not of interest. We therefore group together all permutations of {x_1,…,x_N}. These permutations all have the same probability $P(x_1, \dots, x_N|N, c)$, so if N_p is the number of permutations, then the probability of realizing the group in an experiment is $P(x_1, \dots, x_N|N, c)N_p$.

To obtain an expression for $P(b|N,c)$, it is useful to note that each permutation group {x_1,…,x_N} can be characterized by a distinct pattern of clonal counts {n_0,…,n_c}, where n_0 is the number of clonal barcodes that are found in the mouse but do not appear among the c cells of the experiment; n_1 clones appear in just one cell, n_2 clones appear in two cells, etc. The clonal counts satisfy two constraints $\sum_{k=0}^{c} n_k = N$ and $\sum_{k=0}^{c} k \cdot n_k = c$, so we only need to consider {n_2,…,n_c}, since $n_1 = c - \sum_{k=2}^{c} k \cdot n_k$ and $n_0 = N - \sum_{k=1}^{c} n_k$. With this notation, number of observed barcodes is $b = N - n_0 = \sum_{k=1}^{c} n_k$, and the number of permutations of {x_1,…,x_N} is $N_p = \frac{N!}{n_0! \cdots n_c!}$. The desired probability is thus,

$$P(n_1, \dots, n_c|N, c) = \frac{c!}{\prod_{k=1}^{c} (k!)^{n_k}} N^{-c} \times \frac{N!}{(N - \sum_{k=1}^{c} n_k)! n_1! \cdots n_c!}.$$

The first term in this expression is $P(x_1, \dots, x_N|N, c)$ for all permutations (x_1,…,x_N) with clonal counts {n_1,…,n_c}; the second term is N_p. As noted above, n_1 is not independent of {n_2,…,n_c}, but is included for clarity. For example, the probability of each of the c sampled cells arising from a distinct clone is,

$$P(n_1 = c, n_2 = 0, \dots, n_c = 0|N, c) = \frac{N!}{(N-c)!} N^{-c}.$$

This particular case is well known as the "birthday problem", which asks about the probability that *c* people in a room will have different birthdays (with *N* days per year).

With this result we are now set to apply Bayes' theorem to obtain the probability $P(N|c; n_1, \ldots, n_c)$ for the number of clones *N*, given data on the clone counts $\{n_2,\ldots,n_c\}$ from *c* cells. We use a uniform prior for *N*. We find,

$$P(N|c; n_1, \ldots, n_c) = \frac{P(n_1, \ldots, n_c|N, c)}{\sum_{M=0}^{\infty} P(n_1, \ldots, n_c|M, c)} = \frac{1}{Z} N^{-c} \frac{N!}{(N-b)!},$$

where $Z = \sum_{k=1}^{\infty} \frac{k!}{(k-b)!k^c}$, and $b = \sum_{k=1}^{c} n_k$ is the number of unique barcodes observed in the sample. Noting that $P(N|c; n_1, \ldots, n_c)$ depends only on *b* and not on the individual values of $(n_1,\ldots,n_c)$, we obtain the main result given in the methods section, $P(N|b, c) = \frac{1}{Z} \frac{N!}{(N-b)!N^c}$.