Open access • Posted Content • DOI:10.1101/674937

How heterogeneous thymic output and homeostatic proliferation shape naive T cell receptor clone abundance distributions — Source link

Renaud Dessalles, Maria R. D'Orsogna, Tom Chou

Institutions: University of California, Los Angeles

Published on: 19 Jun 2019 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Naive T cell, Clone (cell biology), T cell, T-cell receptor and Population

How heterogeneous thymic output and homeostatic proliferation shape naive T cell receptor clone abundance distributions

Renaud Dessalles[1], Maria R. D'Orsogna[1,2], Tom Chou[1,3],

1 Dept. of Biomathematics, UCLA, Los Angeles, CA 90095-1766
2 Dept. of Mathematics, CalState-Northridge, Los Angeles, CA 91330
3 Dept. of Mathematics, UCLA, Los Angeles, CA 90095-1555

Abstract

The set of T cells that express the same T cell receptor (TCR) sequence represent a T cell clone. The number of different naive T cell clones in an organism reflects the number of different T cell receptors (TCRs) arising from recombination of the V(D)J gene segments during T cell development in the thymus. TCR diversity and more specifically, the clone abundance distribution is an important factor in immune function. Specific recombination patterns occur more frequently than others while subsequent interactions between TCRs and self-antigens are known to trigger proliferation and sustain naive T cell survival. These processes are TCR-dependent, leading to clone-dependent thymic export and naive T cell proliferation rates. Using a mean-field approximation to the solution of a regulated birth-death-immigration model, we systematically quantify how TCR-dependent heterogeneities in immigration and proliferation rates affect the shape of clone abundance distributions (the number of different clones that are represented by a specific number of cells). By comparing predicted clone abundances derived from our heterogeneous birth-death-immigration model with experimentally sampled clone abundances, we quantify the heterogeneity necessary to generate the observed abundances. Our findings indicate that heterogeneity in proliferation rates is more likely the mechanism underlying the observed clone abundance distributions than heterogeneity in immigration rates.

Author Summary

The abundance distribution of different T cell receptors (TCRs) expressed on naive T cells depends on their rates of thymic output, homeostatic proliferation, and death. However, measured TCR count distributions do not match, even qualitatively, those predicted from a multiclone birth death-immigration process when constant birth, death, and immigration rates are used (a neutral model). We show how non-neutrality in the birth-death-immigration process, where naive T cells with different TCRs are produced and proliferate with a distribution of rates shape the predicted sampled clone abundance distributions (the clone counts). Using physiological parameters, we find that heterogeneity in proliferation rates, and not in thymic output rates, is the main determinant in generating the observed clone counts. These findings are consistent with proliferation-driven maintenance of the T cell population in humans.

Introduction

Naive T cells play a fundamental role in the immune system's response to pathogens, tumors, and other infectious agents. These cells are produced in the thymus, circulate through the blood, and migrate to the lymph nodes where they may be presented with different antigen proteins from various pathogens. Naive T cells mature in the thymus where the so-called V, D, and J segments of genes that code T cell receptors undergo rearrangement. Most T cell receptors (TCRs) are comprised of an alpha chain and a beta chain that are formed after VJ segment and VDJ segment recombination, respectively. The number of possible TCR gene sequences is extremely large, but while recombination is a nearly random process, not all TCRs are formed with the same probability. Before export to the periphery, T cells undergo a selection process, during which T cells with TCRs that react to self proteins are eliminated.

The unique receptors expressed on the cell surface of circulating TCRs enable them to recognize specific antigens; well known examples include the naive forms of helper T cells (CD4+) and cytotoxic T cells (CD8+). The set of naive T cells that express the same TCR are said to belong to the same T cell clone. Upon encountering the antigens that activate their TCRs, naive T cells turn into effector cells that assist in eliminating infected cells. Effector cells die after pathogen clearance, but some develop into memory T cells. Because of the large space of unknown pathogens, TCR clonal diversity is a key attribute in mounting an effective immune response. Recent studies also reveal that human TCR clonal diversity is implicated in healthy ageing, neonatal immunity, vaccination response and T cell reconstitution following haematopoietic stem cell transplantation [1, 2]. Despite the central role of the naive T cell pool in host defense, and broadly speaking in health and disease, TCR diversity is difficult to quantify. For example, the human body is known to host a large repertoire of T cell clones, however the actual distribution of clone sizes is not precisely known [3]. Only recently have experimental and theoretical efforts been devoted to understanding the mechanistic origins of TCR diversity [4–9]. The goal of this work is to formulate a realistic mathematical model that includes heterogeneous naive T cell generation and reproduction rates and that we will use to describe recent experimental results.

A well-established way to describe the T cell repertoire is by determining the clone abundance distribution or "clone count" $\hat{c}_k$ (for $k \geq 1$) that measures the number of distinct clones represented by exactly k T cells: $\hat{c}_k \equiv \sum_{i=1}^{\infty} \mathbb{1}(n_i, k)$, where n_i is the discrete number of T cells carrying TCR i. This distribution captures the entire pattern of the clonal populations. Several summary indices for T cell diversity such as Shannon's entropy, Simpson's index, or the whole population richness can be deduced from the distribution $\hat{c}_k$ [10]. Note that $\hat{c}_k$ counts only the number of clones of a specific population k and does not carry any TCR sequence or identity information.

Complete clone counts $\hat{c}_k$ are difficult to measure in humans due to the large number of naive T cells (the total number of naive T cells is $N \sim 10^{11}$ [11]). Nonetheless, high-throughput DNA sequencing on samples of peripheral blood containing T cells [12–15] have provided some insight into TCR diversity. A commonly observed feature is that all experimentally measured realizations of the clone counts $\hat{c}_k$ exhibit a power-law distribution in the clone abundance k [3, 4, 16–18]. Several authors have introduced stochastic models to explain the power law [4–7]. One of the proposed mechanisms is that T cells in different clones have TCRs that have different affinities for self-ligands that are necessary for peripheral proliferation [4–6], leading to clone specific replication rates. An alternative hypothesis [7] is that specific TCR sequences are more likely to arise in the V(D)J recombination process in the thymus [19] leading to a higher probability that these TCRs are produced. De Greef *et al.* [7] estimated the probability of production of a given TCR sequence by using the Inference and Generation of Repertoires (IGoR) simulation tool that quantitatively characterizes the statistics of receptor generation from both cDNA and gDNA data [19]. However, none of the previous studies have systematically incorporated heterogeneity in both immigration and replication rates, sampling, and comparison with measured TCR clone abundance distributions.

In this paper, we quantitatively analyze the effects of heterogeneity and sampling using a stochastic multiclone birth-death-immigration (BDI) process that allows for TCR-sequence dependent replication and immigration rates. Our model is based on a general continuous-time Markovian birth-death-immigration (BDI) process [10] where: (i) immigration represents the arrival of new clones from the thymus; (ii) birth describes homeostatic proliferation of naive T cells that yield newborn naive T cells with the same TCR as their parent; and (iii) death represents cell apoptosis. We also include a regulation, or "carrying capacity," mechanism through a total population-dependent death rate which may represent the global competition for cytokines, such as Interleukine-7 [20–24], needed for naive T cell survival and homeostasis [25, 26]. This homeostasis will be considered as clone-independent since these cytokine signals are TCR-independent [22]. Mathematically, the inclusion of a carrying capacity ensures a finite naive T cell population at homeostasis.

We derive analytical results of our heterogeneous BDI model that are applicable on the scale of the entire organism. Our calculations provide insight into how parameters describing the shape of the distribution of immigration and proliferation rates affect the shape of the expected clone counts. To compare with experimental measurements, we also quantify the random sampling process that describes actual measurements derived from blood draws. Comparison with available TCR clone abundance data shows that predicted thymic output rate heterogeneity cannot generate the qualitative shape of the observed clone count distribution $\hat{c}_k$. However, we show how a simple uniform proliferation rate distribution can yield the observed $\hat{c}_k$.

Materials and Methods

We focus on the clone counts $\hat{c}_k$ of the clone abundance distribution associated only with naive T cells, the first type of cells produced by the thymus that have not yet been activated by any antigen. Antigen-mediated activation initiates a one-way irreversible cascade of differentiation into effector and memory T cells that we can subsume into a death rate. Thus, we limit our analysis to birth, death, and immigration within the naive T cell compartment.

Non-neutral Birth-Death-Immigration model

The multiclone BDI process is depicted in Fig. 1. We define Q to be the number of possible functional naive

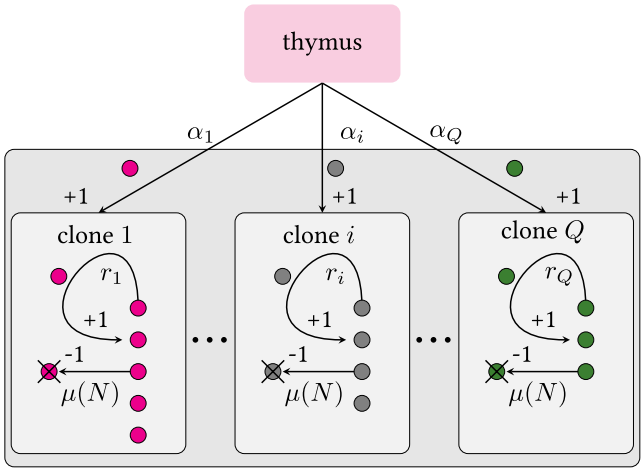

Fig 1. Schematic of a multiclone birth-death-immigration process. Clones are defined by distinct TCR sequences i. Theoretically, there are $Q \gtrsim 10^{15}$ [6] or more [27,28] possible viable V(D)J recombinations that lead to naive T cells. Due to finite and heterogeneous thymic output rates and and selection, the number of clones predicted in an entire organism is much less, on the order $10^6 - 10^7$ [1,6,29–31]. Each clone carries its own thymic output and peripheral proliferation rates, α_i and r_i, respectively. We assume all clones have the same population-dependent death rate $\mu(N)$. Since $Q \gg 1$, we impose a continuous distribution over the rates α and r.

T cell clones that can be generated by V(D)J recombination in the thymus and that survive selection. This quantity is a small fraction ($\sim 1\%$ [32]) of the total theoretical number of possible sequences, estimated to be $10^{15} - 10^{18}$ [6]. Thus, for each TCR chain, we choose $Q = 10^{15}$ as the number of different clones that can be exported by the thymus. Due to naive T cell death, not all possible clone types will be present in the organism, so we denote the number of clones actually present in the body (or "richness") by $C \ll Q$, where estimates of C range from $\sim 10^6 - 10^8$ in mice and human [1,6,29,30]. The total number of naive T cells in an individual is $N \sim 10^{10} - 10^{11}$ [11]. The discrete quantities Q, C, N are related to the discrete clone counts $\hat{c}_k$ through

$$C = \sum_{k \geq 1} \hat{c}_k = Q - \hat{c}_0 \qquad \text{and} \qquad N = \sum_{k \geq 1} k \hat{c}_k, \tag{1}$$

where $\hat{c}_0$ is the number of possible clones that are not expressed in the organism.

Each clone i (with $1 \leq i \leq Q$) is characterized by an immigration rate α_i and a per cell replication rate r_i. The immigration rate α_i is clone-specific because it depends on the preferential V(D)J recombination process; the replication rate r_i is also clone-specific due to the different interactions with self-peptides that trigger proliferation. Since the numbers of potential ($Q \gg 1$) and observed ($C \gg 1$) clones is extremely large, we can define a continuous density $\pi(\alpha, r)$ from which immigration and proliferation rates α and r are drawn. This means that the probability that a given clone has an immigration rate between α and $\alpha + \mathrm{d}\alpha$ and replication rate between r and $r + \mathrm{d}r$ is $\pi(\alpha, r)\mathrm{d}\alpha\mathrm{d}r$. Since Q is finite and countable, there are theoretical maximum values A and R for the immigration and proliferation rates, respectively, such that $\pi(\alpha, r) = 0$ for $\alpha \geq A$ and $r \geq R$. In the BDI process, the upper bound R on the proliferation rate prevents unbounded numbers of naive T cells and

is necessary for a self-consistent solution. Conversely, results are rather insensitive to the upper bound A on the immigration rate provided $\pi(\alpha, r)$ decays sufficiently fast such that the mean value of α is finite for all $r \leq R$. Thus, for simplicity, we henceforth take the $A \to \infty$ limit. The heterogeneity in the immigration and replication rates allow us to go beyond typical "neutral" BDI models, where both rates are fixed to a specific value for all clones.

Finally, we assume the per cell death rate $\mu(N)$ is clone-independent but a function of the total population N. This dependence represents the competition among all naive T cells for a common resource (such as cytokines), which effectively imposes a carrying capacity on the population [23, 28, 33]. We choose the linear form $\mu(N) = \mu_1 N$ but the specific form of the regulation will not qualitatively affect our findings.

Mean-Field Approximation

The exact steady-state probabilities over the discrete abundances $\hat{c}_k$ for a fully stochastic neutral BDI model with regulated death rate $\mu(N)$ were recently derived [10]. To incorporate TCR-dependent immigration and replication rates in a non-neutral model, we must consider distinct values of α_i and r_i for each clone i. In this case, an analytic solution for the probability distribution over $\hat{c}_k$, even at steady state, cannot be expressed in an explicit form. However, since the number of naive T cells ($N \sim 10^{11}$ [11]) is large, we can exploit a mean-field approximation to the non-neutral BDI model and derive expressions for the mean values of the discrete clone counts $\hat{c}_k$. We will show later that under realistic parameter regimes, the mean-field approximation is quantitatively accurate. Breakdown of the mean field approximation has been carefully analyzed in other studies [34].

Deterministic approximation for total population

To implement the mean-field approximation in the presence of a regulated death rate $\mu(N)$, we must first calculate the mean total steady-state population $N^* \gg 1$. We start by writing the deterministic, "mass-action" ODE for the mean number $n_{\alpha,r}(t)$ of cells with immigration rate α and proliferation rate r in a BDI process

$$\frac{\mathrm{d}n_{\alpha,r}(t)}{\mathrm{d}t} = \alpha + r n_{\alpha,r}(t) - \mu(N(t)) n_{\alpha,r}(t). \tag{2}$$

Since the number of clones that have an immigration rate between α and $\alpha + \mathrm{d}\alpha$ and a replication rate between r and $r + \mathrm{d}r$ is approximately $Q\pi(\alpha, r)\mathrm{d}\alpha\mathrm{d}r$, the total mean number $N(t)$ of naive T cells can be estimated as a weighted integral over all $n_{\alpha,r}(t)$

$$N(t) = Q \int_0^\infty \mathrm{d}\alpha \int_0^R \mathrm{d}r \, n_{\alpha,r}(t)\pi(\alpha, r). \tag{3}$$

At steady-state, the solution to Eq. 2 can be simply expressed as

$$n^*_{\alpha,r} = \frac{\alpha}{\mu(N^*) - r} \tag{4}$$

in which N^* is the steady-state value of $N(t)$. Thus, upon averaging Eq. 4 over α and r, we find

$$N^* = Q \int_0^R \mathrm{d}r \int_0^\infty \mathrm{d}\alpha \, \frac{\alpha\pi(\alpha, r)}{\mu(N^*) - r}, \tag{5}$$

a self-consistent equation whose solution N^* will be used in our mean-field approximation for the mean clone counts. Note that without the finite upper bound R of the density $\pi(\alpha, r)$, the integral in Eq. 5 diverges.

Mean-field model of clone counts

We now use the results for the mean steady-state population N^* to find the clone counts averaged over all realizations of the underlying stochastic process. The mean-field equations for the dynamics of these mean clone counts in the neutral model were derived in [34, 35] and are reviewed in Appendix A. In the neutral model, we assume that all effective clones Q are associated with the same rates α and r so that the mean field evolution equation for $c_k(\alpha, r)$ is [34, 35]

$$\frac{\mathrm{d}c_k(\alpha,r)}{\mathrm{d}t} = \alpha\left[c_{k-1}(\alpha,r) - c_k(\alpha,r)\right] + r\left[(k-1)c_{k-1}(\alpha,r) - kc_k(\alpha,r)\right] + \mu(N)\left[(k+1)c_{k+1}(\alpha,r) - kc_k(\alpha,r)\right], \tag{6}$$

along with the constraint $\sum_{k=0}^{\infty} c_k(\alpha,r) = Q$. This equation assumes that both $c_k(\alpha,r)$ and N are uncorrelated, allowing us to write the last term as a product of functions of the mean population $N = \sum_\ell \ell c_\ell$ and c_{k+1}, c_k.

At steady state we replace $\mu(N)$ with $\mu(N^*)$, where the mean steady-state population N^* is found from self-consistently solving Eq. 5. The solution follows a negative binomial distribution with parameters α/r and $r/\mu(N^*)$ [10,34]

$$c_k(\alpha,r) = Q\left(1 - \frac{r}{\mu(N^*)}\right)^{\alpha/r}\left(\frac{r}{\mu(N^*)}\right)^k \frac{1}{k!}\prod_{\ell=0}^{k-1}\left(\frac{\alpha}{r} + \ell\right). \tag{7}$$

Now, suppose the clones obey possibly different values of α and r, as depicted in Fig. 1. The total mean clone count can be obtained from the α, r-specific result in Eq. 7 by averaging over $\pi(\alpha,r)$:

$$c_k = \int_0^\infty \mathrm{d}\alpha \int_0^R \mathrm{d}r\, c_k(\alpha,r)\,\pi(\alpha,r), \tag{8}$$

where for notational simplicity, we refer to the realization- and $\pi(\alpha,r)$-averaged clones counts simply by c_k. Note that since it is an average over the mean-field expression for $c_k(\alpha,r)$, c_k defined in Eq. 8, is to be considered distinct from the actual $\hat{c}_k$, which is instead a discrete stochastic quantity that may be described via, *e.g.*, a Q-dimensional master equation. As mentioned earlier, the presence of heterogeneities makes computing the solution of the master equation impractical. Henceforth, we utilize the mean-field average c_k in Eq. 8 as a proxy for the discrete clone counts $\hat{c}_k$, and discuss in detail the regimes where the identification $\hat{c}_k \to c_k$ is justified; namely, cases where fluctuations may be considered small, and c_k can be compared to experimentally measured realizations. We also henceforth denote the mean field approximation to the total population and total richness by N and C, respectively. Since the population constraints are linear, these mean field quantities also obey the relationships in Eq. 1.

Results

We now analyze Eqs. 5 and 8 using different choices for $\pi(\alpha,r)$ to show the effect of clone-specific immigration and proliferation rates on the overall measured clone abundance distribution. To determine the clone abundance c_k we must first solve the fixed point equation (Eq. 5) and use Eqs. 7 and 8 to determine c_k. Rather than choosing a value for μ_1 in $\mu(N) = \mu_1 N$, we first fix N^* to observed values and then determine μ_1 that satisfies Eq. 5. The parameters used in our study are set according to

- The average total number of naive T cells is set to $N^* = 10^{11}$ [11].
- The total possible number of TCRs of either alpha or beta chains is set to $Q = 10^{15}$ [36].
- The average immigration rate per clone $\bar{\alpha}$ can be deduced from the total thymic output of all clones to be $\bar{\alpha}Q = 1.6 \times 10^7$/day [37]. Using $Q = 10^{15}$, the average per clone immigration rate is set to $\bar{\alpha} = 1.6 \times 10^{-8}$/day.
- The average proliferation rate $\bar{r}$ is estimated as $\bar{r} \sim 0.18$/year. We thus set $\bar{r} = 5 \times 10^{-4}$/day [37].

These parameters will be used in the rest of our analyses. We have verified that the predicted shapes of the clone counts are relatively insensitive to $Q \gg 1$ and $N^* \gg 1$.

Clone-specific immigration

To make analytical progress, we first assume that all clones have the same proliferation rate $\bar{r}$ and that heterogeneity arises only in the immigration rate α through $\pi(\alpha,r) = \pi_\alpha(\alpha)\delta(r-\bar{r})$. We analyze how differential V(D)J recombination probabilities affect the clone abundance distribution by drawing for each clone i $(1 \le i \le Q)$,

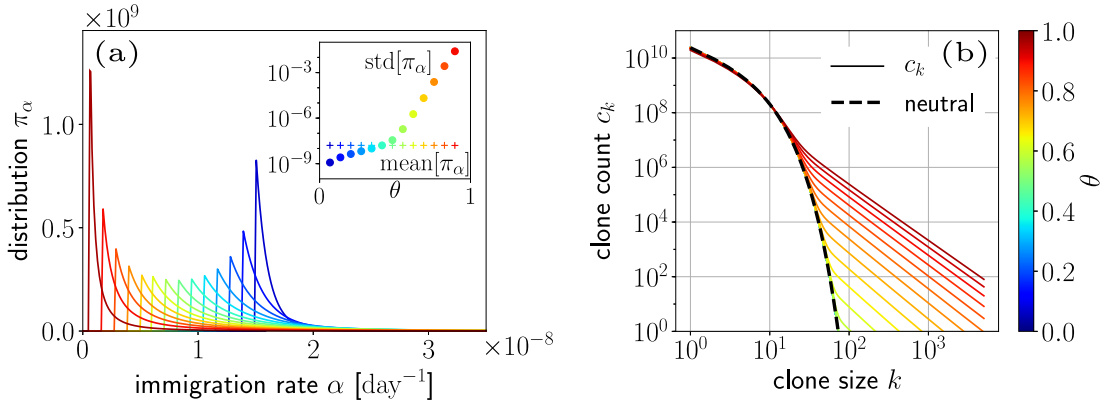

Fig 2. Effects of immigration rate heterogeneity. (a) Different Pareto distributions $\pi_\alpha(\alpha) \equiv \pi_\alpha(\alpha|\bar{\alpha}, \theta)$ of the immigration rate α plotted for the same value of the mean $\bar{\alpha}$ but different values of the shape parameter θ in Eq. 10. Increasing θ increases the heterogeneity in the immigration rate per clone α. Although the different distributions appear to have different mean values, the longer tails of the distributions with larger θ yield means that are invariant to θ. Inset: Standard deviation of $\pi_\alpha(\alpha)$ as a function of θ. (b) Expected clone counts c_k for different choices of θ (Eq. 9). The clone counts differ significantly from that of the BDI neutral model only when $\theta \gtrsim 0.35$, as can be seen from the dashed curve plotted above.

the immigration rate α_i from the probability density $\pi_\alpha(\alpha)$. By averaging the mean clone abundance (Eq. 8) over $\pi_\alpha(\alpha)$, we find

$$c_k = Q\left(\frac{\bar{r}}{\mu(N^*)}\right)^k \frac{1}{k!}\int_0^\infty \left(1-\frac{\bar{r}}{\mu(N^*)}\right)^{\alpha/\bar{r}} \prod_{\ell=0}^{k-1}\left(\frac{\alpha}{\bar{r}}+\ell\right)\pi_\alpha(\alpha)\mathrm{d}\alpha. \tag{9}$$

A specific form for $\pi_\alpha(\alpha)$ can be obtained from previous studies that predict V(D)J recombination frequencies associated with each TCR sequence. The statistical model for differential V(D)J recombination in humans is implemented in the Inference and Generation of Repertoires (IGoR) software [19]. In Appendix B.1, we estimate $\pi_\alpha(\alpha)$ by repeatedly running IGoR [19] and sampling a large number of TCRs. We assume that thymic selection is uncorrelated with V(D)J recombination so the relative probabilities of forming different TCRs provide an accurate representation of the ratios of the TCRs exported into the periphery. The observed frequency of each TCR shows that $\pi_\alpha(\alpha)$ can be approximated by a Pareto distribution with threshold parameter $\bar{\alpha}(1-\theta)$ and shape parameter θ:

$$\pi_\alpha(\alpha|\bar{\alpha},\theta) = \begin{cases} \dfrac{(\bar{\alpha}(1-\theta))^{1/\theta}}{\theta\alpha^{1+1/\theta}} & \text{if } \alpha \geq \bar{\alpha}(1-\theta) \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

The parameters of the Pareto distribution will first be constrained such that $\bar{\alpha}$ is set to the average per-clone immigration rate, $\bar{\alpha} = Q^{-1}\sum_{i=1}^{Q}\alpha_i \approx \int_0^\infty \alpha\pi_\alpha(\alpha)\mathrm{d}\alpha$.

The distributions $\pi_\alpha(\alpha)$ for different values of θ are shown in Fig. 2. An increase in the parameter θ will increase the variance of $\pi_\alpha(\alpha)$ (Fig. 2(a)-inset). In Fig. 2(b) we plot the mean clone count c_k derived in Eq. 9 with $\pi_\alpha(\alpha)$ specified by Eq. 10 for various values of θ. These curves are compared with those obtained when $\theta \to 0$ and $\pi_\alpha(\alpha) \to \delta(\alpha - \bar{\alpha})$ describing a fixed immigration rate $\bar{\alpha}$. The BDI model in this neutral limit predicts a log-series distribution for c_k [10]. As can be seen in Fig. 2(b), deviations from the predictions of the neutral model ($\theta \to 0$) are noticeable only when $\theta \gtrsim 0.35$. The estimates using IGoR fixes $\theta \approx 0.28$, suggesting that heterogeneity in thymic output is insufficient for producing the observed clone abundance distributions of naive T cells.

The shape of c_k as derived from Eqs. 9 and 10 exhibits two regimes. At small clone sizes k, regardless of the value of θ, c_k follows that of the neutral model. At larger sizes k, c_k converges to a power-law. The transition between these two regimes occurs at a characteristic size that decreases with θ, and with the standard deviation of $\pi_\alpha(\alpha)$. In Appendix B.2 and Fig. S2(b), we show that clones within these two regimes correspond to two

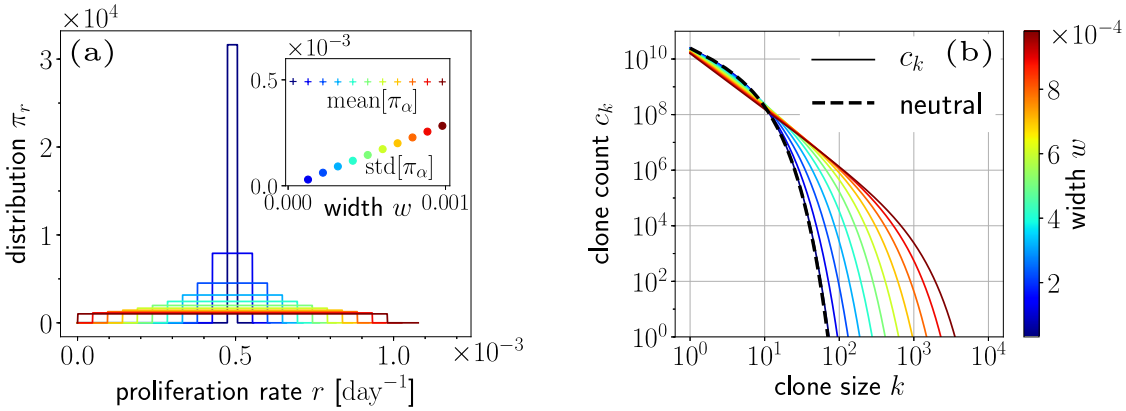

Fig 3. Effects of proliferation rate heterogeneity. (a) Different "box" distributions for $\pi_r(r)$ associated with different widths w (see Eq. 12). Larger w corresponds to higher heterogeneity across clones in the per-cell proliferation rate r. Inset: Standard deviation of $\pi_r(r)$ for different values of w. (b) Predicted clone abundances c_k from Eq. 11 for different choices of w. Wider distributions generate longer-tailed distributions.

sub-populations. The first, with smaller clone sizes, is characterized by immigration rates α smaller than the proliferation rate $\bar{r}$. For these clones, the production of new cells is due mostly to peripheral proliferation rather than from thymic immigration. The second sub-population with larger clone sizes corresponds to clones with immigration rates $\alpha > \bar{r}$. The production of new cells in these clones is driven mainly by thymic immigration rather than by proliferation. Intuitively, one can see that as θ increases, the proportion of clones with higher immigration rate α increases as well, leading to a larger representation of the second, high immigration rate sub-population.

Clone-specific proliferation

The other limit of clonal heterogeneity that can be readily analyzed is that of an immigration rate that is common to all clones and proliferation rates that are clone-dependent. The peripheral proliferation of T cells depends on the affinity of the corresponding TCR to self-antigens. TCR-antigen affinity in turn depends on the receptor amino acid sequence; thus, the rate at which a T cell proliferates can be clone-specific [28, 38]. The distribution of proliferation rates among all the Q possible clones is a mapping to the interactions between TCRs and low-affinity MHC/self-peptide complexes that has not been experimentally resolved.

For simplicity, we assume $\pi(\alpha, r) = \pi_r(r)\delta(\alpha - \bar{\alpha})$ where $\bar{\alpha}$ is the common immigration rate and where the probability that a clone has a replication rate between r and $r + dr$ is $\pi_r(r)\mathrm{d}r$. We previously introduced the biologically relevant constraint that $\pi(r) = 0$ for $r > R$; we now further assume that $R < \mu(N^*)$ so that a finite fixed point solution N^* exists for Eq. 5. This assumption implies that the death rate is larger than the fastest proliferation rate and allows us to write

$$c_k = \frac{Q}{k!}\int_0^R \left(1 - \frac{r}{\mu(N^*)}\right)^{\bar{\alpha}/r} \left(\frac{r}{\mu(N^*)}\right)^k \prod_{\ell=0}^{k-1}\left(\frac{\bar{\alpha}}{r} + \ell\right) \pi_r(r)\,\mathrm{d}r. \tag{11}$$

We illustrate in Fig. 3 the effect of a uniform proliferation rate distribution

$$\pi_r(r) = \begin{cases} 1/w & \text{if } |r - \bar{r}| < w/2 \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

where w represents the width of the uniform distribution centered about $\bar{r}$, as shown in Fig. 3(a). By construction, the maximum allowable immigration rate is $R = \bar{r} + w/2$. Fig. 3(b) shows that an increase in w leads to an increase in the proportion of large clones. It is important to note that the shape of the clone abundance c_k is particularly sensitive to the behavior of $\pi_r(r)$ near the upper bound R as can be seen in Fig. S3. This is because, regardless of their initial numbers, clones with $r \approx R$ will have a proliferation advantage over other clones and will eventually dominate the total naive T cell population, as shown in Fig. S3(c).

Sampling

Unless an animal is sacked and its entire naive T cell population sequenced, TCR clone distributions are typically measured from sequencing TCRs in a small blood sample. In such samples, low population clones may be missed. In order to compare our predictions with measured clone abundance distributions, we must revise our predictions to allow for random cell sampling.

We define η as the fraction of naive T cells in an organism that is drawn in a sample and assume that all naive T cells in the organism have the same probability η of being sampled. This is true only if naive T cells carrying different TCRs are not preferentially partitioned into different tissues and are uniformly distributed within an animal. Let us assume that a specific TCR is represented by ℓ cells in an organism. If $N\eta \gg \ell$, the probability that k cells from the same clone are sampled approximately follows a binomial distribution with parameters ℓ and η [39–41]

$$\mathbb{P}\left[k|\ell\right] \approx \binom{\ell}{k}\eta^k\left(1-\eta\right)^{\ell-k}, \quad k \leq \ell. \tag{13}$$

The associated mean *sampled* clone count c_k^{s} depends on the clone count c_ℓ in the body via the formula

$$c_k^{\mathrm{s}} \approx \sum_{\ell\geq k} c_\ell \mathbb{P}\left[k|\ell\right] = \sum_{\ell\geq k} c_\ell \binom{\ell}{k}\eta^k(1-\eta)^{\ell-k}, \tag{14}$$

where c_ℓ is found through Eq. 8.

In Fig. 4(a) we show the effects of sampling by plotting both the whole-organism c_k derived from Eq. 8 and the corresponding sampled distribution c_k^{s} calculated using Eq. 14. We show three cases: i) the c_ℓ distribution is derived from a fully neutral model with specific values of $\bar{\alpha}$ and $\bar{r}$; ii) the proliferation rate is fixed at $\bar{r}$ and a high heterogeneity in the immigration rate α is chosen with $\pi_\alpha(\alpha)$ given in Eq. 10 with $\theta \sim 1$, iii) the immigration rate is kept fixed at $\bar{\alpha}$ and proliferation rates r are highly heterogeneous with $w = 2\bar{r} = 9.8\times10^{-4}$/day in $\pi_r(r)$, as given in Eq. 12.

Experimental sampling will strongly affect the observed clone abundances c_k^{s} especially when $\eta \ll 1$. We set the sampling fraction to $\eta = 2\times10^{-3}$, which is typical for data on humans. In all three cases described above the observed clone abundances c_k^{s} are significantly reduced with respect to c_k, reducing the overall slopes of the power-law-like clone abundance distributions. The magnitude of the reduction in the exponent depends on the value of η and can be quite significant as can be seen in Fig. S4).

An interesting crossover feature can be seen, particularly when sampling in the case of heterogeneous immigration rates, as described in case ii) above. As can be seen from the blue curves in Fig. 4(a), two power-law regimes arise with a crossover size $k \sim 2-3$. To understand this behavior, consider the effect of sampling from a hypothetical power-law clone abundance $c_\ell \sim \ell^{-\lambda}, \lambda = 2.1$ as shown in the inset of Fig. 4(b). For a given sampled clone of size k, we can calculate its most probable size $\hat{\ell}(k)$ in the organism by evaluating the maximum likelihood of the binomial sampling

$$\hat{\ell}(k) = \max_{\ell\geq k}\left[c_\ell\binom{\ell}{k}\eta^k(1-\eta)^{\ell-k}\right], \quad c_\ell \sim \ell^{-2.1}. \tag{15}$$

The $\hat{\ell}(k)$ curve in this case is shown in Fig. 4(b). For clones of size $k \leq 2$ measured in the sample, the most likely size in the organism $\hat{\ell}(k)$ is also small, less than ~ 10 copies in the body. Conversely, sampled clones of size $k \geq 2$ are most likely to have originated from clones with $\gtrsim 10^3$ copies in the body. The two regimes observed in Fig. 4(b) can then be interpreted as a trade-off between the contributions of the large number of clones with small populations and the few number of clones with large populations, leading to a cross-over behavior in c_k. This cross-over arises in the heterogeneous immigration model (blue curve) in Fig. 4(a) and is a consequence of sampling and not the intrinsic kink in the whole-body c_k at $k \approx 30$. The sampling trade-off is not apparent in the neutral and heterogeneous proliferation model, case i) and iii) above, and depicted by the black and red curves in Fig. 4(a). These distributions decay too fast before the "sampling kink" reveals itself.

Comparison with measured clone abundances

We now compare our predictions for c_k^{s} with measured clone abundances from Oakes *et al.* [13]. These data were extracted by combining barcoding with TCR sequencing of either the alpha or beta chains of TCRs from

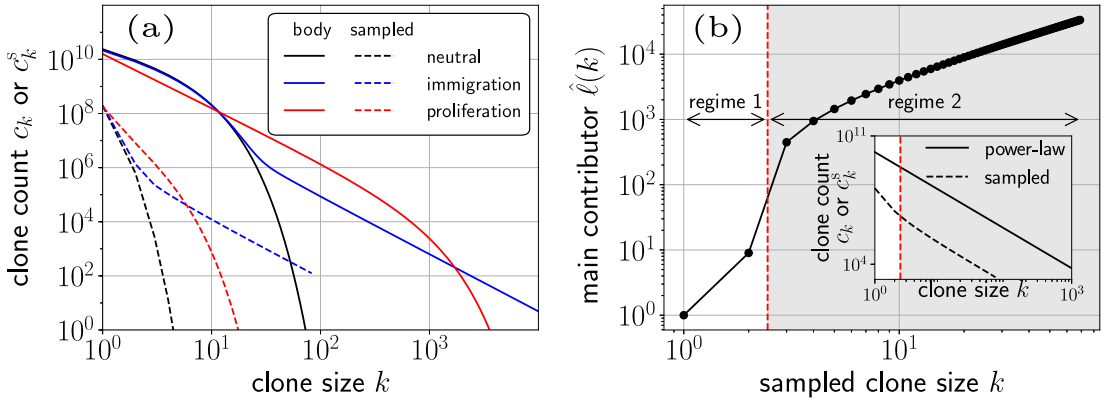

Fig 4. Effect of sampling on mean clone abundances c_k. (a) Whole-body mean clone abundances c_k (solid curves) and expected sampled (dashed curves) abundances c_k^s for the neutral (black), immigration rate heterogeneity (blue), and proliferation rate heterogeneity (red) models. The sampling fraction used is $\eta = 2 \times 10^{-3}$. (b) Illustration of two regimes in sampling. We considered sampling from an idealized pure power-law clone abundance $c_\ell \sim \ell^{-\lambda}, \lambda = 2.1$ (inset) that is consistent with that predicted by the heterogeneous proliferation rate model (solid red curve in (a)). For a sampled clone of size k, the main plot shows the most likely abundance of said clone in the whole body, $\hat{\ell}(k)$, defined in Eq. 15. Plotting $\hat{\ell}(k)$ indicates two regimes in the sampling. The main contributors to sampled clones of size $k \leq 2$ (regime 1) are smaller whole-body clones ($\ell \lesssim 10$), while the main contributors to clones of size $k \geq 3$ (regime 2) are clones of size $\ell \gtrsim 10^3$.

naive human CD4+ and CD8+ T cells, thereby eliminating PCR bias, especially for larger clones [15]. In Fig. 5 we plot the experimental clone abundances alongside our theoretical results for c_k^s as evaluated in Eq. 13. We focused our fitting to small- and intermediate-sized clones since the mean field approximation at large clone sizes k is not accurate [34], the frequencies at large k are low and fitting to a continuous expected distribution is not reliable, and the single large clones represented by the flat region in Fig. 5 contain memory T cells that have expanded [26, 42]. Moreover, we do not expect a model for an expected continuous clone count to accurately capture these isolated, discrete single clones (open circles) in the large-k regime where $c_k = 0, 1$. Since our model does not incorporate the emergence of memory T cells, and the large-k data is more likely to include a higher percentage of memory cells, we simply drop the single-clone counts and fit only those alpha chain or beta chain clones for which the clone count $c_k \geq 2$.

As shown by the black curve in Fig. 5, the expected sampled clone count c_k^s derived from the neutral model with fixed $\bar{\alpha}, \bar{r}$ cannot qualitatively fit the data. Heterogeneity in immigration rate α also does not yield a good match to data for realistic values of the sampling parameter η, except possibly at large clone sizes k where the clone counts are highly variable, and only for high degrees of heterogeneity $\theta \sim 0.9$. However, as estimated from IGoR, $\theta \approx 0.28 > 0.9$. At this value of θ, no significant difference between the neutral and the heterogeneous immigration rate cases are to be expected, especially after sampling. Thus, a heterogeneous immigration rate model cannot be consistent with both measured clone abundances [13] and the level immigration heterogeneity estimated from IGoR [19].

On the other hand, the red curves in Fig. 5 show that the c_k^s predicted using heterogeneity in the proliferation rate match the data for reasonable values of the clone size k, using realistic sampling fractions η. In this context, since a specific TCR is composed of one alpha chain and one beta chain, and since the distributions of alpha and beta chains are quantified separately, the heterogeneous proliferation rates should be interpreted as those of an individual alpha or beta sequence, and not as the heterogeneity in functional TCR-self-antigen avidity.

In Fig. S4, we show additional comparisons for different values of the sampling fractions, confirming that $\eta \sim 2 \times 10^{-3}$ provides the optimal fitting with measured clone counts. In summary, it is evident that the neutral model cannot be made to fit the data without using unrealistic sampling rates $\eta \gtrsim 0.1$. The model with large immigration heterogeneity also fails to fit some of the data, in particular, those derived from sequencing different beta chains on CD4+ cells.

We have also tried to fit our model to data from mice, such as from Zarnitsyna *et al.* [31]. These preliminary data did not filter out PCR errors and yielded biphasic clone abundance distributions with $\lambda \approx 1.76, 1.1$. The magnitudes of these exponents are too small to be accurately fit by a heterogeneous BDI process, even when

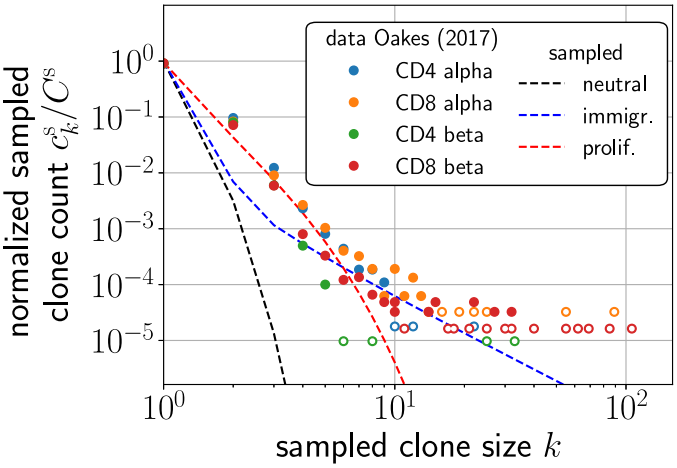

Fig 5. Comparison of the predicted mean sampled clone abundances c_k^s (dashed) with measured human TCR alpha and beta chain seuence abundances (dots) in which PCR bias was avoided [13]. The best-fit model (for reliable measurements at modest sizes k) model includes proliferation rate heterogeneity. In the dataset shown, TCRs observed in memory cells are not subtracted. Data points corresponding to single clones are denoted by the open circles.

unrealistic distributions in r are used and the effects of sampling are neglected.

Discussion and Conclusions

Here, we review and justify a number of critical biological assumptions and mathematical approximations used in our analysis. The effects of relaxing our approximations are also discussed.

Mean field approximation. Our mean-field approximation is embodied in Eq. 6, where correlations between fluctuations in the total population $N = \sum_k kc_k$ in the regulation term $\mu(N)$ and the explicit c_k terms are neglected. This approximation has been shown to be accurate for $k \lesssim N^*$ when $\bar{\alpha}Q^2 > \mu(N^*)$ [34]. When the total T cell immigration rate is extremely small, the above constraint is violated and a single large clone arises due to competitive exclusion [34, 43, 44]. In this case, an accurate approximation for the steady-state clone abundance c_k can be obtained using a variation of the two-species Moran model [34]. Thus, except under such extreme scenarios, the parameters associated with the human adaptive immune system satisfy the condition for the mean-field approximation to be accurate.

For large k the number of cells contributing to c_k is also large so demographic stochasticity is relatively small and results in small uncertainties in the value of k, and not in the magnitude of c_k. However, the mean field approximation to large-k clone counts is not accurate. For very large clone sizes, the total population constraint forces c_k to rapidly approach zero for $k \gtrsim N^*$, a feature that is not accurately reflected in the mean-field approximation [34]. Finally, large clones likely include memory T cells that have been produced after antigen stimulation of specific clones. Thus, the data at large k is likely to contain T cells in the memory pool. Thus, in our analysis, we perform qualitative "fits" to the data only for modest values of k.

Estimation of $\pi(\alpha, r)$. Given an explicit mean-field solution to $c_k(\alpha, r)$ from Eq. 7, distributions $\pi(\alpha, r)$ detailing the immigration and proliferation rates α and r are then required to compute the expected clone abundances in c_k from Eq. 8. The main difficulty is accurately representing immigration and proliferation heterogeneity. Even if we assume that the immigration and proliferation rates are uncorrelated and that $\pi(\alpha, r)$ is factorizable, $\pi(\alpha, r) = \pi_\alpha(\alpha)\pi_r(r)$, there have been no direct experimental measurements of either $\pi_\alpha(\alpha)$ or $\pi_r(r)$.

We refer to a study on the statistical inference of the probability of TCR sequences to estimate immigration heterogeneity. In particular, we used the IGoR (Inference and Generation of Repertoires) computational tool which predicts recombination events [19]. We selected approximately 10^8 such IGoR recombination events,

and created a frequency histogram on the observed sequences. The resulting frequency distribution closely follows Zipf's law [45, 46], from which we constructed a self-consistent Pareto distribution for $\pi_\alpha(\alpha)$, with a mean $\bar{\alpha} \sim 1.6 \times 10^{-8}$/day [37]. As mentioned, we assume that selection and recombination are uncorrelated and that the results from IGoR can be used to infer the heterogeneity of the thymic output after selection. While IGoR allows us to reconstruct $\pi_\alpha(\alpha)$ there are no reliable data on the *in vivo* distribution of TCR-dependent homeostatic proliferation rates r.

Previous studies modeled the proliferation rate of each TCR as being directly proportional to the average interaction times between the TCR itself and a large number of self-ligands [4, 6]. Each TCR was assumed to interact with multiple self-antigens through cross-reactivity and the interaction of each TCR–self-antigen pair is sampled from a given Gaussian [4] or log-normal [6] distribution. However averages were taken over a large number of TCR–self-ligand interactions sampled from the same distribution and may underestimate the true proliferation heterogeneity. In this work we assumed a simple box distribution for $\pi_r(r)$ centered around the mean value $\bar{r} \sim 5 \times 10^{-4}$/day [37] and analyzed the predicted the clone abundances as the width w of the box is varied.

Qualitative description of clone abundances c_k. Here, we offer a qualitative interpretation of the observed clone abundance profiles c_k as a function of k, as seen in Fig. 4. The arguments we use closely resemble those presented to describe the two regimes observed under heterogeneous immigration, case ii), and are more thoroughly discussed in Appendix B.2 and Fig. S2(b). We begin by dissecting the behavior of $c_k(\alpha, r)$, which, as described above, defines a negative binomial distribution with parameters $u = \alpha/r$ and $v = r/\mu(N^*)$. In Fig. 6 we plot the mean clone size k,

$$\frac{\sum_{k=1}^{\infty} k c_k(\alpha, r)}{\sum_{k=1}^{\infty} c_k(\alpha, r)} = \frac{uv}{1-v} = \frac{\alpha}{\mu(N^*) - r}, \tag{16}$$

as a function of $u = \alpha/r$. Two different regimes emerge:

- If $u \to 0$ while $v = r/\mu(N^*)$ remains constant (lower left in Fig. 6), $c_k(\alpha, r)$ (conditioned on $k > 0$) converges to a log-series distribution.
- If $u \to \infty$ while the averaged clone size given by $uv/(1-v) = \alpha/(\mu(N^*) - r)$ (Eq. 16) remains constant, $c_k(\alpha, r)$ converges to a Poisson distribution with parameter $uv/(1-v)$.

These results follow from analogous considerations made in Appendix B.2. Finally, we show paths in the $u = \alpha/r$ and $uv/(1-v) = \alpha/(\mu(N^*) - r)$ plane over which the constrained "line integrals" correspond to the distributions $\pi(\alpha, r)$ for the neutral, heterogeneous immigration, and heterogeneous proliferation rate models.

Results are displayed in Fig. 6. As can be seen, both the heterogeneous proliferation and the neutral cases fall in the log-series regimes; here the expected clone counts result from an integration of different log-series distributions over $\pi(\alpha, r)$. The heterogeneous immigration case includes both log-series and Poisson distributions, generating the two qualitatively different regimes. More details are discussed in Appendix B.2).

Correlated immigration α and proliferation r. Hitherto, we have considered independent immigration and proliferation, and assumed a factorizable rate distribution $\pi(\alpha, r) = \pi_\alpha(\alpha)\pi_r(r)$. However, immigration and proliferation rates may be correlated for certain clones. For example, a frequent realization of V(D)J recombination may also result in a TCR that is more likely to be activated for proliferation. In this case, α would be positively correlated with r. In Fig. S5 we consider the effect of correlated $\pi(\alpha, r)$. For $\bar{r}/2 \leq r \leq 2\bar{r}$, we considered normalized, positively/negatively correlated box distributions as shown in Fig S5(a):

$$\begin{aligned} \text{Positively correlated}: \quad & \pi(\alpha, r) = \frac{1}{\bar{r}} \delta\left(\alpha - \frac{\bar{\alpha}}{\bar{r}} r\right), \\ \text{Negatively correlated}: \quad & \pi(\alpha, r) = \frac{1}{\bar{r}} \delta\left(\alpha - \bar{\alpha}\left(2 - \frac{r}{\bar{r}}\right)\right). \end{aligned} \tag{17}$$

Within our mean field model, these correlated distributions $\pi(\alpha, r)$ result in very similar expected clone abundance distributions c_k (Fig S5(b)). This insensitivity to correlations between immigration and proliferation can be qualitatively understood by considering the "line integral" over dominant paths of $\pi(\alpha, r)$ in the

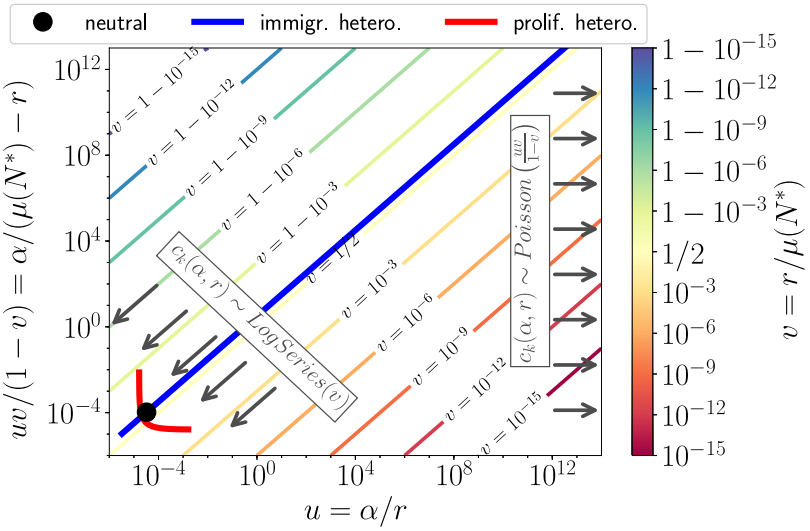

Fig 6. Parameter regimes of different distributions. For a clone with immigration rate α and proliferation rate r, (whose clone size follows a negative binomial distribution of parameters $u = \alpha/r$ and $v = r/\mu(N^*)$, see Eq. 7, the diagram shows the average clone size $uv/(1-v) = \alpha/(\mu(N^*) - r)$ as a function of $u = \alpha/r$. When $v > 0$ is held constant and $u \to 0$, the clone counts (conditioned to $k > 0$) converges to a log series distribution. If $uv/(1-v) = \alpha/(\mu(N^*) - r)$ remains constant while $u \to \infty$, the clone size converges to a Poisson distribution. The "line integral" of the three models of Fig. 4 are shown: both the neutral model, and the proliferation model are in the log-series regime; the immigration regime covers both the log-series and the Poisson regime.

$uv/(1-v) = \alpha/(\mu(N^*) - r)$ *vs.* $u = \alpha/r$ diagram. As shown in Fig. S5(c), both line integrals remain in the log-series distribution regime, indicating that the clone abundance distributions are qualitatively similar to that predicted by a model with proliferation heterogeneity alone.

Steady state assumption. In this study, we only considered the steady state of our birth-death-immigration model in Eq. 7 because this limit allowed relatively easy derivations of analytical results. This was also the strategy for previous modeling work [4, 6, 7, 34, 35]. However, the per clone immigration and proliferation times may be on the order of months or years, a time scale over which thymic output can change significantly as an individual ages [33, 37, 47, 48] and as thymic output diminishes [33, 37]. Moreover, clone abundance distributions have been shown to show specific patterns as a function of age [49–51].

From the dynamics of $N(t)$ in a neutral model with fixed $\bar{\alpha}$ and $\bar{r}$, the timescale for relaxation to steady state can be estimated to be $\sim 15 - 20$ years [33]. Besides the total population, the different subpopulations of specific sizes described by their number c_k relax to steady-state in a spectrum of time scales depending on the clone sizes k [52].

Thus, the steady-state may not be strictly reached and should be considered as an approximation, especially given time-dependent perturbations, including aging, to the adaptive immune system of an individual.

These timescales can be approximated by the eigenvalues of linearized forms of Eqs. 6. Besides time-dependent changes in α, more subtle time-inhomogeneities such as changes in proliferation and and death rates have been demonstrated [47, 48]. Thus, our steady-state assumption should be relaxed by incorporation of time-dependent perturbations to the model parameters $\mu(N)$ and/or $\pi(\alpha, r)$. Longitudinal measurements of clone abundances or experiments involving time-dependent perturbations would provide significant insight into the overall dynamics of clone abundances.

General conclusions. We developed a heterogeneous multispecies birth-death-immigration model and analyzed it in the context of T cell clonal heterogeneity; the clone abundance distribution is derived in the mean-field limit. Unlike previous studies [4], our modeling approach incorporated sampling statistics and provided simple formulae, allowing us to predict clone abundances under different rate distributions for arbitrarily large systems ($N \sim 10^{11}$), without the need for simulation.

The heterogeneous BDI model produces mean clone count distributions that follow power laws over a range of sizes k, with varying exponents that depend on the immigration and proliferation rate heterogeneity and sampling size η. We were able to compare results from our model to measured clone abundance distributions in alpha and beta TCR chains. Based on recombination statistics inferred from cDNA and gDNA sequences [19], we derived a Pareto distribution quantifying the heterogeneity in TCR immigration rates. The derived distribution for $\pi_\alpha(\alpha)$, given in Eq. 10 with $\theta = 0.28$, however, is not large enough to generate a sampled clone abundance distribution c_k^{s} consistent with observations. In fact, the predicted distributions for reasonable sampling fractions η are nearly indistinguishable from those arising from a simple neutral model in which all clones have the same immigration rate. Conversely, proliferation heterogeneity within our model yields significantly different clone abundances that match those seen in experimental samples. Even under modest proliferation rate heterogeneity, larger clones become significantly more numerous at steady state since, although the number of TCR clones with large proliferation rates r may be small, such clones proliferate more rapidly contributing to higher clone counts at larger sizes. In particular, we found that the expected clone abundance is sensitive to the behavior of $\pi_r(r \approx R)$, the proliferation distribution near the maximum proliferation rate R.

Our work leads to the conclusion that proliferation heterogeneity is the more likely mechanism driving the emergence of the power law distributions as observed in [13]. These results are consistent with the findings that peripheral selection leads to contraction of T cell diversity [49] and that naive T cells in humans are maintained by proliferation rather than thymic output [9]. Since we have only explored the effects of a uniform distribution for $\pi_r(r)$, further studies using more complex shapes of $\pi(\alpha, r)$ can be easily explored using our modeling framework. Different parameter values and rate distributions appropriate for mice, in which naive T cells are maintained by thymic output, should also be explored within our modeling framework.

Acknowledgements

This work was supported by grants from the NIH through grant R01HL146552 and the NSF through grants DMS-1814364 (TC) and DMS-1814090 (MD). The authors also thank the Collaboratory in Institute for Quantitative and Computational Biosciences at UCLA for support to RD.

References

1. Qi Q, Liu Y, Cheng Y, Glanville J, Zhang D, Lee JY, et al. Diversity and clonal selection in the human T-cell repertoire. Proceedings of the National Academy of Sciences (USA). 2014;111:13139–13144.

2. van den Broek T, Borghans JAM, van Wijk F. The full spectrum of human naive T cells. Nature Reviews Immunology. 2018;18:363–373.

3. Laydon DJ, Bangham CRM, Asquith B. Estimating T-Cell Repertoire Diversity: Limitations of Classical Estimators and a New Approach. Phil Trans R Soc B. 2015;370(1675):20140291. doi:10.1098/rstb.2014.0291.

4. Desponds J, Mora T, Walczak AM. Fluctuating Fitness Shapes the Clone-Size Distribution of Immune Repertoires. Proceedings of the National Academy of Sciences, USA. 2016;113(2):274–279. doi:10.1073/pnas.1512977112.

5. Desponds J, Mayer A, Mora T, Walczak AM. Population Dynamics of Immune Repertoires. arXiv preprint arXiv:170300226. 2017;.

6. Lythe G, Callard RE, Hoare RL, Molina-París C. How Many TCR Clonotypes Does a Body Maintain? Journal of Theoretical Biology. 2016;389:214–224.

7. de Greef PC, Oakes T, Gerritsen B, Ismail M, Heather JM, Hermsen R, et al. V(D)J Recombination Shapes the Distribution of TCR Chains in the Naive T-Cell Repertoire. Proceedings of the National Academy of Sciences. 2018;Preprint.

8. Koch H, Starenki D, Cooper SJ, Myers RM, Li Q. powerTCR: A Model-Based Approach to Comparative Analysis of the Clone Size Distribution of the T Cell Receptor Repertoire. PLOS Computational Biology. 2018;14(11):e1006571. doi:10.1371/journal.pcbi.1006571.

9. den Braber I, Mugwagwa T, Vrisekoop N, Westera L, Mögling R, Bregje de Boer A, et al. Maintenance of Peripheral Naive T Cells Is Sustained by Thymus Output in Mice but Not Humans. Immunity. 2012;36(2):288–297. doi:10.1016/j.immuni.2012.02.006.

10. Dessalles R, D'Orsogna M, Chou T. Exact Steady-State Distributions of Multispecies Birth–Death–Immigration Processes: Effects of Mutations and Carrying Capacity on Diversity. Journal of Statistical Physics. 2018;doi:10.1007/s10955-018-2128-4.

11. Jenkins MK, Chu HH, McLachlan JB, Moon JJ. On the Composition of the Preimmune Repertoire of T Cells Specific for Peptide–Major Histocompatibility Complex Ligands. Annual Review of Immunology. 2010;28(1):275–294. doi:10.1146/annurev-immunol-030409-101253.

12. Mora T, Walczak AM, Bialek W, Callan CG. Maximum Entropy Models for Antibody Diversity. Proceedings of the National Academy of Sciences, USA. 2010;107(12):5405–5410.

13. Oakes T, Heather JM, Best K, Byng-Maddick R, Husovsky C, Ismail M, et al. Quantitative Characterization of the T Cell Receptor Repertoire of Naïve and Memory Subsets Using an Integrated Experimental and Computational Pipeline Which Is Robust, Economical, and Versatile. Frontiers in Immunology. 2017;8. doi:10.3389/fimmu.2017.01267.

14. Aguilera-Sandoval CR, O Yang O, Jojic N, Lovato P, Chen DY, Boechat MI, et al. Supranormal Thymic Output Up to Two Decades After HIV-1 Infection. AIDS (London, England). 2016;30(5):701–711. doi:10.1097/QAD.0000000000001010.

15. Gerritsen B, Pandit A, Andeweg AC, de Boer RJ. RTCR: A Pipeline for Complete and Accurate Recovery of T Cell Repertoires from High Throughput Sequencing Data. Bioinformatics. 2016;32(20):3098–3106. doi:10.1093/bioinformatics/btw339.

16. Burgos JD, Moreno-Tovar P. Zipf-Scaling Behavior in the Immune System. Biosystems. 1996;39(3):227–232. doi:10.1016/0303-2647(96)01618-8.

17. Weinstein JA, Jiang N, White RA, Fisher DS, Quake SR. High-Throughput Sequencing of the Zebrafish Antibody Repertoire. Science. 2009;324(5928):807–810. doi:10.1126/science.1170020.

18. Naumov YN, Naumova EN, Hogan KT, Selin LK, Gorski J. A Fractal Clonotype Distribution in the CD8+ Memory T Cell Repertoire Could Optimize Potential for Immune Responses. The Journal of Immunology. 2003;170(8):3994–4001. doi:10.4049/jimmunol.170.8.3994.

19. Marcou Q, Mora T, Walczak AM. High-Throughput Immune Repertoire Analysis with IGoR. Nature Communications. 2018;9(1):561.

20. Tan JT, Dudl E, LeRoy E, Murray R, Sprent J, Weinberg KI, et al. IL-7 Is Critical for Homeostatic Proliferation and Survival of Naïve T Cells. Proceedings of the National Academy of Sciences, USA. 2001;98(15):8732–8737. doi:10.1073/pnas.161126098.

21. Schluns KS, Kieper WC, Jameson SC, Lefranccois L. Interleukin-7 Mediates the Homeostasis of Naïve and Memory CD8 T Cells *in Vivo*. Nature Immunology. 2000;1(5):426–432. doi:10.1038/80868.

22. Ciupe SM, Devlin BH, Markert ML, Kepler TB. The Dynamics of T-Cell Receptor Repertoire Diversity Following Thymus Transplantation for DiGeorge Anomaly. PLOS Computational Biology. 2009;5(6):e1000396. doi:10.1371/journal.pcbi.1000396.

23. Reynolds J, Coles M, Lythe G, Molina-Par'is C. Mathematical model of naive T cell division and survival IL-7 thresholds. Frontiers in Immunology. 2013;4:434.

24. Silva SL, Sousa AE. Establishment and Maintenance of the Human Naïve CD4+ T-Cell Compartment. Frontiers in Pediatrics. 2016;4:119.

25. Surh CD, Sprent J. Homeostasis of naive and memory T cells. Immunity. 2008;29(6):848–862.

26. Farber DL, Yudanin NA, Restifo NP. Human memory T cells: generation, compartmentalization and homeostasis. Nature Reviews Immunology. 2014;14:24–35.

27. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, et al. Comprehensive assessment of T-cell receptor β-chain diversity in $\alpha\beta$ T cells. Blood. 2009;114(19):4099–4107.

28. Mayer A, Zhang Y, Perelson AS, Wingreen NS. Regulation of T cell expansion by antigen presentation dynamics. Proceedings of the National Academy of Sciences, USA. 2019;116:5914–5919.

29. Arstila TP, Casrouge A, Baron V, Even J, Kanellopoulos J, Kourilsky P. A Direct Estimate of the Human $\alpha\beta$ T Cell Receptor Diversity. Science. 1999;286(5441):958–961.

30. Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. Genome Research. 2011;21:790–797.

31. Zarnitsyna V, Evavold B, Schoettle L, Blattman J, Antia R. Estimating the Diversity, Completeness, and Cross-Reactivity of the T Cell Repertoire. Frontiers in Immunology. 2013;4. doi:10.3389/fimmu.2013.00485.

32. Yates AJ. Theories and quantification of thymic selection. Frontiers in Immunology. 2014;5:13.

33. Lewkiewicz S, Chuang YL, Chou T. A mathematical model predicting decay of naive T-cell diversity with age. In Press: Bulletin of Mathematical Biology. 2018;.

34. Xu S, Chou T. Immigration-Induced Phase Transition in a Regulated Multispecies Birth-Death Process. Journal of Physics A: Mathematical and Theoretical. 2018;51(42):425602. doi:10.1088/1751-8121/aadcb4.

35. Goyal S, Kim S, Chen ISY, Chou T. Mechanisms of blood homeostasis: lineage tracking and a neutral model of cell populations in rhesus macaques. BMC Biology. 2015;13:85.

36. Davis MM, Bjorkman PJ. T-Cell Antigen Receptor Genes and T-Cell Recognition. Nature. 1988;334(6181):395. doi:10.1038/334395a0.

37. Westera L, van Hoeven V, Drylewicz J, Spierenburg G, van Velzen JF, de Boer RJ, et al. Lymphocyte Maintenance during Healthy Aging Requires No Substantial Alterations in Cellular Turnover. Aging Cell. 2015;14(2):219–227. doi:10.1111/acel.12311.

38. Min B, Foucras G, Meier-Schellersheim M, Paul WE. Spontaneous proliferation, a response of naïve CD4 T cells determined by the diversity of the memory cell repertoire. Proceedings of the National Academy of Sciences, USA. 2004;101(11):3874–3879.

39. Xu S, Kim S, Chen ISY, Chou T. Modeling large fluctuations of thousands of clones during hematopoiesis: The role of stem cell self-renewal and bursty progenitor dynamics in rhesus macaque. PLoS Computational Biology. 2018;14:e1006489.

40. Ferrarini M, Molina-París C, Lythe G. Sampling from T Cell Receptor Repertoires. In: Graw F, Franziska Matthäus JP, editors. Modeling Cellular Systems. Springer; 2017. p. 67–79.

41. Lythe G, Molina-París C. Some deterministic and stochastic mathematical models of naive T-cell homeostasis. Immunological Reviews. 2018;285:206–217.

42. Gossel G, Hogan T, Cownden D, Seddon B, Yates AJ. Memory CD4 T cell subsets are kinetically heterogeneous and replenished from naive T cells at high levels. eLife. 2017;6:e23013.

43. Hardin G. The competitive exclusion principle. Science. 1960;131:1292–1297.

44. Hutchinson GE. The paradox of the plankton. American Naturalist. 1961;95:137–145.

45. Zipf GK. Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. New York: Addison-Wesley; 1949.

46. Aitchison L, Corradi N, Latham PE. Zipf's Law Arises Naturally When There Are Underlying, Unobserved Variables. PLOS Computational Biology. 2016;12(12):1–32. doi:10.1371/journal.pcbi.1005110.

47. Hogan T, Gossel G, Yates AJ, Seddon S. Temporal fate mapping reveals age-structured heterogeneity in naive CD4 and CD8 T lymphocyte populations in mice. Proceedings of The National Academy of Sciences USA. 2015;112:E6917–E6926.

48. Rane S, Hogan T, Seddon B, Yates AJ. Age is not just a number: Naive T cells increase their ability to persist in the circulation over time. PLoS Computational Biology. 2018;16:e2003949.

49. Johnson P, Yates A, Goronzy J, Antia R. Peripheral selection rather than thymic involution explains sudden contraction in naive CD4 T-cell diversity with age. Proceedings of The National Academy of Sciences USA. 2012;109:21432–21437.

50. Britanova OV, Shugay M, Merzlyak EM, Staroverov DB, Putintseva EV, Turchaninova MA, et al. Dynamics of Individual T Cell Repertoires: From Cord Blood to Centenarians. The Journal of Immunology. 2016;doi:10.4049/jimmunol.1600005.

51. Egorov ES, Kasatskaya SA, Zubov VN, Izraelson M, Nakonechnaya TO, Staroverov DB, et al. The changing landscape of naive T cell receptor repertoire With human aging. Frontiers in Immunology. 2018;9:1618.

52. Lewkiewicz SM, Chuang YL, Chou T. Dynamics of T Cell Receptor Distributions Following Acute Thymic Atrophy and Resumption. arXiv. 2019; p. arXiv:1905.12179.

S1: Mathematical Appendices

A Neutral model

Here, we review the neutral model to provide insight into the properties of our heterogeneous BDI model. When there is no heterogeneity in either proliferation or immigration rates, $\pi(\alpha, r) = \delta(\alpha - \bar{\alpha})\delta(r - \bar{r})$. Upon inserting this expression for $\pi(\alpha, r)$ in Eqs. 7 and 8, we find that the clone abundance c_k follows a negative binomial distribution [10]:

$$c_k = Q \left(1 - \frac{\bar{r}}{\mu(N^*)}\right)^{\bar{\alpha}/\bar{r}} \left(\frac{\bar{r}}{\mu(N^*)}\right)^k \frac{1}{k!} \prod_{k'=0}^{k-1} \left(\frac{\bar{\alpha}}{\bar{r}} + k'\right). \tag{S1}$$

We can also express c_k/C, the clone abundance distribution normalized by the mean richness C in the body as

$$\frac{c_k}{C} = \frac{c_k}{\sum_{\ell \geq 1} c_\ell} \tag{S2}$$

which is a negative binomial distribution of parameters $\bar{\alpha}/\bar{r}$ and $\bar{r}/\mu(N^*)$. Using $\bar{\alpha} \approx 1.6 \times 10^{-8}$/day, $\bar{r} \sim 5 \times 10^{-4}$/day, and $\mu(N^*) \approx 6.4 \times 10^{-4}$, we find $\bar{\alpha}/\bar{r} \ll \bar{r}/\mu(N^*)$. In this regime, c_k/C, for $k \geq 1$, can be approximated by a log-series distribution with parameter $\bar{r}/\mu(N^*)$. Fig. S2(a) shows that the exact solution for $c_k(\bar{\alpha}, \bar{r})$ is indistinguishable from the log-series approximation.

To mathematically show that c_k/C converges to a log-series distribution when $\bar{\alpha}/\bar{r} \to 0$, consider a random variable X that follows a negative binomial distribution of parameters $\bar{\alpha}/\bar{r}$ and $\bar{r}/\mu(N^*)$

$$\mathbb{P}[X = k] = \left(1 - \frac{\bar{r}}{\mu(N^*)}\right)^{\bar{\alpha}/\bar{r}} \left(\frac{\bar{r}}{\mu(N^*)}\right)^k \frac{1}{k!} \prod_{\ell=0}^{k-1} \left(\frac{\bar{\alpha}}{\bar{r}} + \ell\right). \tag{S3}$$

Note that the probability mass function of X is given by c_k/Q as can be seen from Eq. S1, the clone abundance distribution for all possible Q clones, which includes c_0, the number of all clones that are not represented in the organism. To find the clone abundance distribution c_k/C, for all the C clones present in the organism, we must exclude the case $k = 0$ by marginalizing the distribution of X over all $X > 0$:

$$\mathbb{P}[X = k | X > 0] = \frac{\mathbb{P}[X = k]}{\sum_{\ell \geq 1} \mathbb{P}[X = \ell]} = \frac{c_k/Q}{\sum_{\ell \geq 1} c_\ell/Q} = \frac{c_k}{C}. \tag{S4}$$

What remains is to show that the distribution of converges to a log-series distribution of parameter $\bar{r}/\mu(N^*)$ when $\bar{\alpha}/\bar{r} \to 0$. Consider the moment generating function of $X|X > 0$ given by

$$\mathbb{E}\left[e^{\xi X} | X > 0\right] = \frac{\mathbb{E}\left[e^{\xi X}\right] - \mathbb{E}\left[e^{\xi X} | X = 0\right] \mathbb{P}[X = 0]}{\mathbb{P}[X > 0]}. \tag{S5}$$

Since the moment generating function of a negative binomial distribution $\mathbb{E}\left[e^{\xi X}\right]$ is known, and since $\mathbb{P}[X > 0] = 1 - \mathbb{P}[X = 0]$ (see Eq. S3), we can write

$$\mathbb{E}\left[e^{\xi X} | X > 0\right] = \frac{\left(\frac{1 - \bar{r}/\mu(N^*)}{1 - e^{\xi}\bar{r}/\mu(N^*)}\right)^{\bar{\alpha}/\bar{r}} - \left(1 - \frac{\bar{r}}{\mu(N^*)}\right)^{\bar{\alpha}/\bar{r}}}{1 - \left(1 - \frac{\bar{r}}{\mu(N^*)}\right)^{\bar{\alpha}/\bar{r}}}. \tag{S6}$$

For any $x > 0$, the limit of $\bar{\alpha}/\bar{r} \to 0$, yields

$$x^{\bar{\alpha}/\bar{r}} = 1 + \frac{\bar{\alpha}}{\bar{r}} \log x + o\left(\frac{\bar{\alpha}}{\bar{r}}\right).$$

If we apply this result to Eq. S6 for $\mathbb{E}\left[e^{\xi X} | X > 0\right]$, we find

$$\mathbb{E}\left[e^{\xi X}|X>0\right] = \frac{1+\frac{\bar{\alpha}}{\bar{r}}\log\left(\frac{\mu(N^*)-\bar{r}}{\mu(N^*)-e^{\xi}\bar{r}}\right)-\left(1+\frac{\bar{\alpha}}{\bar{r}}\log\left(1-\frac{\bar{r}}{\mu(N^*)}\right)\right)+o\left(\frac{\bar{\alpha}}{\bar{r}}\right)}{-\frac{\bar{\alpha}}{\bar{r}}\log\left(1-\frac{\bar{r}}{\mu(N^*)}\right)+o\left(\frac{\bar{\alpha}}{\bar{r}}\right)}$$
$$= \frac{\log\left(1-e^{\xi}\frac{\bar{r}}{\mu(N^*)}\right)}{\log\left(1-\frac{\bar{r}}{\mu(N^*)}\right)}+o(1),$$

which we recognize as the moment generating function of a log series distribution of parameter $\bar{r}/\mu(N^*)$. Thus, we finally have

$$\lim_{\bar{\alpha}/\bar{r}\to 0} c_k = \frac{C}{\log\left(\frac{1}{1-\bar{r}/\mu(N^*)}\right)}\frac{1}{k}\left(\frac{\bar{r}}{\mu(N^*)}\right)^k. \tag{S7}$$

B Heterogeneous immigration

Here, we provide technical details of the derivation of a Pareto distribution for the immigration rates α.

B.1 Determination of immigration rate distribution $\pi_\alpha(\alpha)$

Marcou et al. [19] developed a computational tool, Inference and Generation of Repertoires (IGoR), which infers the probabilities of the recombination events that lead to the generation of a sequence. Using this tool, we generated 10^8 different trials (for producing the human beta chain) and determined the frequency of occurrence of the corresponding sequence. Most of the sampled sequences arose only once, but others occurred twice or more in the sample. If n_i denotes the number of times each distinct sequence i arises in the 10^8-trial simulation, its frequency is $f_i = n_i/10^8$. In Fig. S1 we plot the IGoR-generated frequencies f_i in descending order $f_1 \geq f_2 \geq \ldots \geq f_Q$.

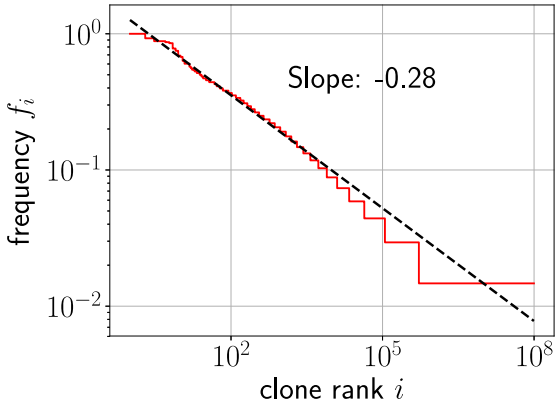

Fig S1. Clone rank versus frequency of sequences produced by IGoR (sampled for the human beta chain). Fitting this data to a Zipf distribution $f_i \sim i^{-\theta}$ yields the best-fit parameter $\theta \approx 0.28$.

By identifying the sequences generated by IGoR, we show that the distribution of thymic output rates follows a Zipf law (a discrete power-law distribution $f_i \sim 1/i^\theta$). This correspondence is not surprising since our procedure is analogous to determining word frequencies [45] from a given corpus of words in natural language, which is also known to follow a discrete power law. We can thus posit

$$f_i = \frac{1}{H_{\theta,Q} i^\theta} \tag{S8}$$

where $H_{\theta,Q} = \sum_{i=1}^{Q} i^{-\theta}$ is a generalized harmonic number that serves as a normalizing constant. Using regression on the 10^8 counts generated from IGoR, we determined that the slope of the Zipf distribution in our problem is $\theta \approx 0.28$.

Note that in principle we should probe a sufficient number of events to adequately represent the total number of distinct TCR sequences, on the order $Q \sim 10^{15}$. However, since this is costly from a computational perspective, we generated 10^8 recombination events (a "trial") and assumed that the frequency-rank distribution would remain Zipf-like, with parameter $\theta \approx 0.28$, even under a higher number of trials. In the main text, we tested different values of θ in case this exponent is modified by choosing a larger number of trials.

Based on Zipf's Law (Eq. S8), we reconstructed the statistics of immigration rates and $\pi_\alpha(\alpha)$ by associating f_i to the probability that each trial using IGoR generates a clone of sequence i. The total immigration rate $\bar{\alpha} \approx 1.6 \times 10^7$/day [37] represents the rate of the total TCR immigration. For each clone i, we can define its specific immigration rate α_i as

$$\alpha_i = \bar{\alpha} Q f_i = \frac{\bar{\alpha} Q}{H_{\theta,Q} i^\theta}. \tag{S9}$$

By definition, the immigration rates of all clones α_i $(1 \leq i \leq Q)$ are sampled from $\pi_\alpha(\alpha)$. Since the quantity Q is very large, one can assume that given bounds a and b,

$$\int_a^b \pi_\alpha(\alpha) \mathrm{d}\alpha \approx \frac{1}{Q} \sum_{i=1}^{Q} \mathbb{1}\left(\alpha_i \in [a, b[\right), \tag{S10}$$

where $\mathbb{1}\left(\alpha_i \in [a, b[\right)$ is the indicator function, of unitary value if α_i falls in the $[a, b[$ interval, and zero otherwise. From Eq. S8, we find the equivalent representation

$$\int_a^b \pi_\alpha(\alpha) \mathrm{d}\alpha \simeq \frac{1}{Q} \sum_{i=1}^{Q} \mathbb{1}\left(i \in \left[\left(\frac{\bar{\alpha} Q}{b H_{\theta,Q}}\right)^{1/\theta}, \left(\frac{\bar{\alpha} Q}{a H_{\theta,Q}}\right)^{1/\theta}\right[\right). \tag{S11}$$

After using the first order approximation in the polynomial approximation for $b \to a$, we find

$$\pi_\alpha(\alpha) = \frac{1}{\theta \alpha^{1+1/\theta}} \left(\frac{\bar{\alpha} Q^{1-\theta}}{H_{\theta,Q}}\right)^{1/\theta}, \tag{S12}$$

which is a Pareto distribution with parameters $\bar{\alpha} Q^{1-\theta}/H_{\theta,Q}$ and $1/\theta$. Note that since Q is large, the exact computation of $H_{\theta,Q}$ is numerically challenging but a very accurate approximation can be obtained by considering its integral representation

$$H_{\theta,Q} \simeq \int_0^Q \frac{1}{x^\theta} \mathrm{d}x = \frac{Q^{1-\theta}}{1-\theta}, \tag{S13}$$

so that finally

$$\pi_\alpha(\alpha) = \frac{(\bar{\alpha}(1-\theta))^{1/\theta}}{\theta \alpha^{1+1/\theta}}. \tag{S14}$$

B.2 Shape of c_k in the heterogeneous immigration model

In the model with immigration heterogeneity, two regimes appear in the clone abundance c_k (see Figs. 2(b) and S2(b)). Small-sized clones are distributed similarly as in the neutral model; larger sized clones tend to follow a power-law. The interpretation of this behavior, provided in the main text, is that small-sized clones (population 1) are characterized by a low immigration rate $\alpha < \bar{r}$, while larger-sized clones (population 2) are represented by a higher immigration rate $\alpha > \bar{r}$. For clones in population 1, new cells arise in the clone mainly due to proliferation, while in population 2, they increase mainly due to immigration.

To verify this interpretation, we consider the clone count fraction $p_k = c_k/C$ for general $\pi(\alpha, r)$:

$$p_k := \frac{Q}{C} \left(\frac{\bar{r}}{\mu(N^*)}\right)^k \frac{1}{k!} \int_0^\infty \left(1 - \frac{\bar{r}}{\mu(N^*)}\right)^{\frac{\alpha}{\bar{r}}} \prod_{\ell=0}^{k-1} \left(\frac{\alpha}{\bar{r}} + \ell\right) \pi_\alpha(\alpha) \, \mathrm{d}\alpha. \tag{S15}$$

We separate the clone abundance distribution p_k into $p_k^{\{1\}}$, where immigration rates $\alpha < \bar{r}$, and $p_k^{\{2\}}$ where immigration rates $\alpha > \bar{r}$ so that

$$p_k^{\{1\}} := \int_0^{\bar{r}} \frac{Q}{C} \left(\frac{\bar{r}}{\mu(N^*)}\right)^k \frac{1}{k!} \left(1 - \frac{\bar{r}}{\mu(N^*)}\right)^{\frac{\alpha}{\bar{r}}} \prod_{\ell=0}^{k-1} \left(\frac{\alpha}{\bar{r}} + \ell\right) \pi_\alpha(\alpha)\, \mathrm{d}\alpha,$$
$$p_k^{\{2\}} := \int_{\bar{r}}^{\infty} \frac{Q}{C} \left(\frac{\bar{r}}{\mu(N^*)}\right)^k \frac{1}{k!} \left(1 - \frac{\bar{r}}{\mu(N^*)}\right)^{\frac{\alpha}{\bar{r}}} \prod_{\ell=0}^{k-1} \left(\frac{\alpha}{\bar{r}} + \ell\right) \pi_\alpha(\alpha)\, \mathrm{d}\alpha$$

and $p_k = p_k^{\{1\}} + p_k^{\{2\}}$ (see Eq. S15). We approximate $p_k^{\{1\}}$ by focusing on the contribution to the integrand by terms with immigration rates $\alpha \ll \bar{r}$, and similarly for $p_k^{\{2\}}$ where we assume the greatest contribution arises from terms with $\alpha \gg \bar{r}$.

To evaluate $p_k^{\{1\}}$, since $\alpha \ll \bar{r}$ in the corresponding integrand, we can proceed in the same way as for the derivation of Eq. S7 in Section A to obtain a log-series distribution

$$p_k^{\{1\}} \simeq \frac{\pi_\alpha([0, \bar{r}[)}{\log\left(\frac{1}{1-\bar{r}/\mu(N^*)}\right)} \frac{1}{k} \left(\frac{\bar{r}}{\mu(N^*)}\right)^k. \tag{S16}$$

To evaluate $p_k^{\{2\}}$, we consider a random variable $X_{\alpha,\bar{r}}$ following a negative binomial distribution of parameters $\alpha/\bar{r}$ and $\bar{r}/\mu(N^*)$ and note that $p_k^{\{2\}}$ can be written as

$$p_k^{\{2\}} = \frac{Q}{C} \int_{\bar{r}}^{\infty} \mathbb{P}\left[X_{\alpha,\bar{r}} = k\right] \pi_\alpha(\alpha)\, \mathrm{d}\alpha. \tag{S17}$$

Since we have imposed $\alpha \gg \bar{r}$, we consider the $\bar{r} \to 0$ limit of $\mathbb{P}\left[X_{\alpha,\bar{r}} = k\right]$, or equivalently the behavior of its moment generating distribution $\mathbb{E}\left[e^{\xi X_{\alpha,\bar{r}}}\right]$ as $\bar{r} \to 0$. The latter is given by

$$\mathbb{E}\left[e^{\xi X_{\alpha,\bar{r}}}\right] = \left(\frac{1 - \bar{r}/\mu(N^*)}{1 - e^{\xi}\bar{r}/\mu(N^*)}\right)^{\alpha/\bar{r}}. \tag{S18}$$

Under the assumption $\alpha \gg \bar{r}$, we take the $\bar{r} \to 0$ limit of Eq. S18 to find

$$\begin{aligned}
\mathbb{E}\left[e^{\xi X_{\alpha,\bar{r}}}\right] &= \exp\left[\frac{\alpha}{\bar{r}} \log\left(\frac{1 - \bar{r}/\mu(N^*)}{1 - e^{\xi}\bar{r}/\mu(N^*)}\right)\right] \\
&= \exp\left[\frac{\alpha}{\bar{r}} \log\left(1 + \frac{\bar{r}}{\mu(N^*)}\left(e^{\xi} - 1\right) + o(\bar{r})\right)\right] \\
&= \exp\left[\frac{\alpha}{\bar{r}} \left(\frac{\bar{r}}{\mu(N^*)}\left(e^{\xi} - 1\right) + o(\bar{r})\right)\right] \\
&\sim \exp\left[\frac{\alpha}{\mu(N^*)}\left(e^{\xi} - 1\right)\right].
\end{aligned} \tag{S19}$$

We now recognize the moment generating function of a Poisson distribution of parameter $\alpha/\mu(N^*)$ so that

$$\mathbb{P}\left[X_{\alpha,\bar{r}} = k\right] \xrightarrow{\bar{r}\to 0} \frac{1}{k!} \left(\frac{\alpha}{\mu(N^*)}\right)^k e^{-\alpha/\mu(N^*)} \tag{S20}$$

and

$$p_k^{\{2\}} \simeq \frac{Q}{C} \int_{\bar{r}}^{\infty} \frac{1}{k!} \left(\frac{\alpha}{\mu(N^*)}\right)^k e^{-\alpha/\mu(N^*)} \pi_\alpha(\alpha)\, \mathrm{d}\alpha. \tag{S21}$$

Thus, in this regime, clone abundance ratios obey a Poisson distribution of parameter $\alpha/\mu(N^*)$. The approximations for $p_k^{\{1\}}$ and $p_k^{\{2\}}$ given by Eqs. S16 and S21 respectively are shown in Fig S2(b). Both approximations well represent the two sub-populations:

- Clones with a small cell number are mainly clones with an immigration rate $\alpha \ll \bar{r}$ (population 1). When considering a clone in this regime that has at least one individual in the body, its distribution globally follows a log-normal distribution of parameter $\bar{r}/\mu(N^*)$. The creation of new cells in these clones is mainly driven by proliferation.
- Clones with a high cell number are mainly clones with an immigration rate $\alpha \gg \bar{r}$ (population 2). When considering a clone in this regime with an immigration rate α, its distribution globally follows a Poisson distribution of parameter $\alpha/\mu(N^*)$. The creation of new cells in these clones is mainly driven by thymic output.

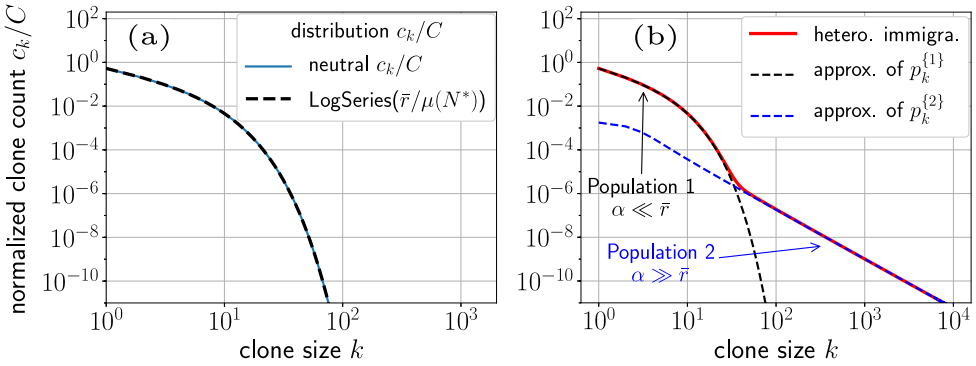

Fig S2. (a) Comparison of the clone abundance distribution (Eq. S7) with the log-series distribution in the neutral model. (b) Illustration of the two regimes in the clone abundance distribution of the heterogeneous immigration. The normalized clone abundance distribution $p_k = c_k/C$ is simply the normalization of c_k of Fig. 2(b) with $\theta = 0.8$. The population p_k is subdivided in two sub-populations $p_k^{\{1\}}$ and $p_k^{\{2\}}$ that represent clones with immigration rates α respectively lower or higher than $\bar{r}$ as described in Section B.2. The approximations of $p_k^{\{1\}}$ and $p_k^{\{2\}}$ (given by Eqs. S16 and S21), well represent the two regimes of p_k.

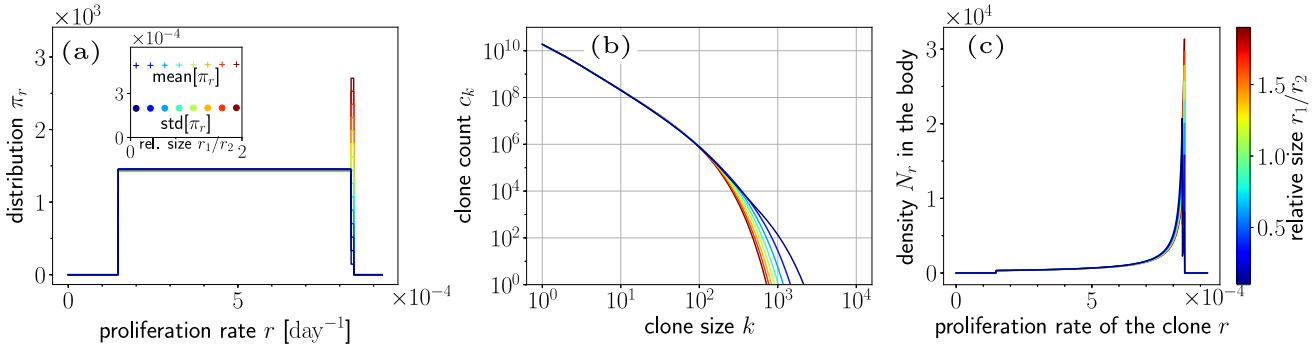

Fig S3. Effect of $\pi_r(r \approx R)$ on the predicted mean clone count c_k. (a) To study the effect of $\pi_r(r)$ when r is near its upper-bound R, we consider a series of almost-similar proliferation distributions $\pi_r(r)$ where only the local behavior for $r \approx R$ changes. Specifically, we allow $\pi_r(r)$ to take on two constant values: $\pi_r(r) = r_1$ for $r \leq 0.99R$ and $\pi_r(r) = r_2$ for $R \geq r \geq 0.99R$. By varying the ratio r_1/r_2, we change the local behavior at $r \approx R$ without significantly affecting the mean or the variance of $\pi_r(r)$ (see inset). (b) Decreasing r_1/r_2, increases the proportion of larger-size clones. (c) Spreading of proliferation rates r among the T cell population: N_r denotes the density of cells in the whole body with a proliferation rate between r and $r + dr$. Cells with higher r have a fitness advantage. Paradoxically, when r_1/r_2 decreases, the local competition between clones with proliferation rate $r \approx R$ decreases (the bluest curves are lower than the others around the upper bound R). Thus the fewer the number of clones with replication rate $r \approx R$ the less competition arises, leading to higher clone sizes k as observed in the (b).

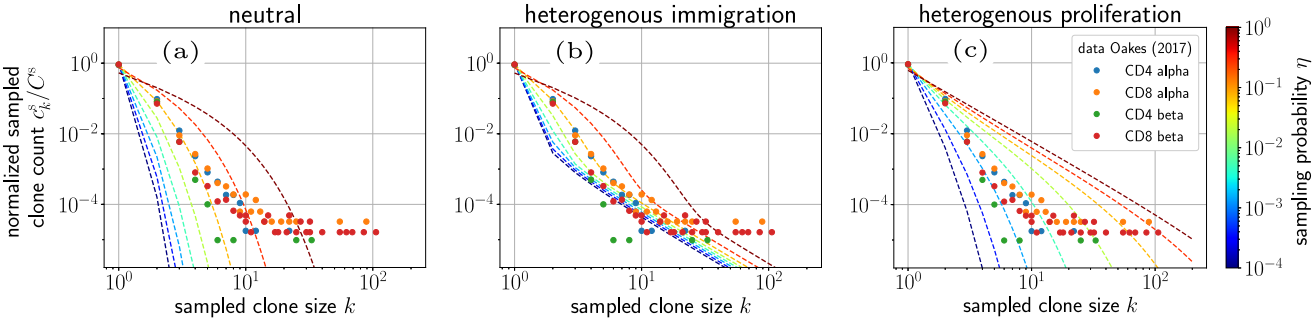

Fig S4. Effect of different sampling fractions η on predicted sampled clone abundances c_k^s. For all plots, the dots correspond to the experimental results of [13]. (a) Sampled clone abundances for the neutral model (without immigration or proliferation heterogeneity). (b) Sampled clone abundances for a model with high immigration rate heterogeneity corresponding to Fig. 2 for $\theta \approx 1$. The fit is poor even at this unrealistic value of θ. (c) Expected sampled clone counts for a high proliferation rate heterogeneity model (corresponding to the Fig. 3 for $w = 2\bar{r} = 9.8 \times 10^{-4}$/day).

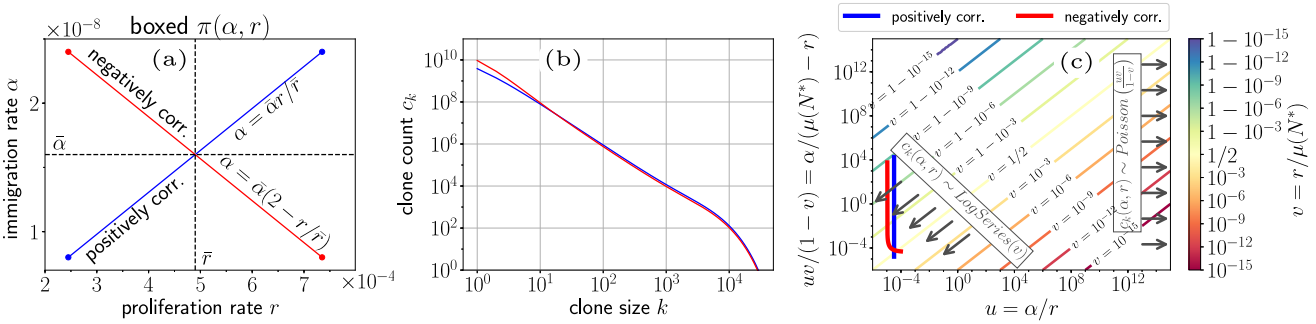

Fig S5. Positively and negatively correlated $\pi(\alpha, r)$. (a) For $\bar{r}/2 \leq r \leq 2\bar{r}$, we consider $\pi(\alpha, r)$ distributions with positively and negatively correlated α and r (Eqs. 17). (b) Mean sampled clone counts corresponding to positively and negatively correlated $\pi(\alpha, r)$ show negligible differences. (c) "Line integrals" of the positively and negatively correlated distributions $\pi(\alpha, r)$ in the $uv/(1-v)$-u diagram. Clones counts predicted by such $\pi(\alpha, r)$ follow log-series distributions, similar to those of a neutral model.