How Naive T-Cell Clone Counts Are Shaped By Heterogeneous Thymic Output and Homeostatic Proliferation

Renaud Dessalles, Yunbei Pan, Mingtao Xia, Davide Maestrini, Maria R. D’Orsogna and Tom Chou

1 Department of Computational Medicine, University of California at Los Angeles (UCLA), Los Angeles, CA, United States,
2 Department of Mathematics, California State University at Northridge, Los Angeles, CA, United States,
3 Department of Mathematics, University of California at Los Angeles (UCLA), Los Angeles, CA, United States

The specificity of T cells is that each T cell has only one T cell receptor (TCR). A T cell clone represents a collection of T cells with the same TCR sequence. Thus, the number of different T cell clones in an organism reflects the number of different T cell receptors (TCRs) that arise 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, are important factors in immune functions. 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. We describe the heterogeneous steady-state population of naive T cells (those that have not yet been antigenically triggered) by using a mean-field model of a regulated birth-death-immigration process. After accounting for random sampling, we investigate 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, or “clone counts”). By using reasonable physiological parameter values and fitting predicted clone counts to experimentally sampled clone abundances, we show that realistic levels of heterogeneity in immigration rates cause very little change to predicted clone-counts, but that modest heterogeneity in proliferation rates can generate the observed clone abundances. Our analysis provides constraints among physiological parameters that are necessary to yield predictions that qualitatively match the data. Assumptions of the model and potentially other important mechanistic factors are discussed.

Keywords: naive T cells, T-cell receptor, repertoire diversity, clone-count distributions, mathematical modeling, immigration-proliferation model, heterogeneity
INTRODUCTION

Naive T cells play a crucial role in the immune system’s response to pathogens, tumors, and other infectious agents. These cells are produced in the bone marrow, mature 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.

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 number of unknown pathogens, TCR clonal diversity is a key factor for mounting an effective immune response. Recent studies also reveal that human TCR clonal diversity is implicated in healthy aging, 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 hosts 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 incorporates heterogeneity in naive T cell generation and reproduction. Model predictions are compared with T cell clone data to estimate reasonable and realistic parameter values.

One way to describe the TCR repertoire is by tallying the population $n_i$ of T cells carrying receptor $i$. Another is to use the clone abundance distribution or “clone count” that measures the number of distinct clones composed of exactly $k$ T cells, $c_k := \sum_{n=k}^\infty f(n, k)$, where the indicator function $f(n, k) = 1$ if $n = k$ and 0 otherwise. Clone counts $c_k$ do not carry TCR identity information as $n_i$ does, however, they can be used to construct other summary indices for T cell diversity such as Shannon’s entropy, Simpson’s index, or the whole population richness $C := \sum_{k=1}^\infty k c_k$ (10).

Cloned counts $c_k$ and the total number of circulating naive T cells are difficult to measure in humans. Nonetheless, high-throughput DNA sequencing on samples of peripheral blood containing T cells (11–14) have provided some insight into TCR diversity. A commonly invoked model is that clone counts $c_k$ exhibit a power-law distribution (4, 12, 15–17) in the clone abundance $k$. Several models have been developed to explain observed features of clone counts (3, 4, 15, 18, 19), including the apparent power-law behavior. One proposal 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 (20) 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 (20).

Although power-law models have been motivated, this behavior has been observed across only about two decades of clone sizes $k$, as shown in Figure 1. Moreover, the above models have not systematically incorporated and compared heterogeneity in both immigration and replication rates, and/ or fitted models to measured TCR clone abundance distributions. Finally, some of them have not taken into account subsampling in measurements, which will affect the predicted clone counts, especially for small clone sizes $k$ which can be missed in small samples. In this paper, we analyze the effects of heterogeneity and sampling within a dynamic mean-field model based on a stochastic clone-dependent birth-death-immigration (BDI) process that includes (i) immigration representing the arrival of new clones from the thymus, (ii) birth during homeostatic proliferation of naive T cells that yield newborn naive T cells with the same TCR as their parent, and (iii) death representing cell apoptosis (10). We also include a regulating “carrying capacity” mechanism through a total population-dependent death rate which may represent the

![Figure 1](https://example.com/image1.png)
global competition for cytokines, such as Interleukin-7 (21–25), needed for naive T cell survival and homeostasis (26, 27). Since these cytokine signals are TCR-independent, the regulatory interaction, which ensures a finite homeostatic naive T cell population, is clone-independent (23).

We derive analytic expressions for the steady state clone counts in the entire organism and show that the predicted distributions are negative binomials. However, since T cell clone populations are measured in small blood subsamples extracted from an organism, we modify our predictions to include the effects of random subsampling and find that the negative binomial structure is preserved. Finally, the subsampled prediction will be averaged over distributions of TCR generation (thymic output) and homeostatic proliferation rates. The distribution of TCR generation rates are extracted from new computational tools: Inference and Generation of Repertoires (1GoR) (20) and Optimized Likelihood estimate of immunoGlobulin Amino-acid sequences (OLGA) (28). Since there are no equivalent tools that measure proliferation rates, we will assume simple functions for the distribution of homeostatic proliferation rates. These model-derived results depend on the rate parameters of the model and the hyperparameters defining the probability distributions over these T cell production and proliferation rates (see Table 1).

Our results are then compared to the data shown in Figure 1 and used to estimate hyperparameters associated with the heterogeneity in the TCR-specific immigration and proliferation rates. Specifically, we quantify how the width of a simple uniform proliferation rate distribution and the heterogeneity of immigration rates from a generative model affect the predicted clone counts. Our analysis explicitly shows that within reasonable physiological parameter ranges, heterogeneity in the thymic immigration rate cannot significantly change clone count distributions. However, clone counts are sensitive to heterogeneity in T cell proliferation rates. Thus, different levels of heterogeneity in proliferation rates can give rise to qualitatively different clone count distributions. This finding of the dominance of proliferation in shaping clone count distributions is consistent with the observation that in older humans with severely reduced thymic output a broad clone count distribution is still maintained (9, 29).

### MATERIALS AND METHODS

To understand the observed clone counts, we focus on the clone count distribution \( c_k \) 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 largely 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. Here, we first present the mathematical framework of the BDI process to provide an initial qualitative understanding for clone counts.

### Heterogeneous Birth-Death-Immigration Model

The multiclonal BDI process is depicted in Figure 2. We define \( Q \) to be the theoretical number of all possible functional naive T cell receptor clones that can be generated by V(D)J recombination in the thymus which is estimated to be \( Q \sim 10^{15} - 10^{18} \) (6, 28). As we will later show, results of our model will not depend on the explicit value of \( Q \) as long as \( Q \gg 1 \). Due to naive T cell death or removal from the sampling-accessible pool, not all possible clone types will be presented in the organism, so we denote the number of clones actually present in the body (or “richness”) by \( \hat{C} \ll Q \), where estimates of \( \hat{C} \) range from \( \sim 10^6 - 10^8 \) in mice and humans (1, 6, 32, 33, 35, 36).

Although naive T cells are difficult to distinguish from the entire T cell population, the total number of naive T cells (across all clones present) in humans has been estimated to be about \( N \sim 10^{11} \). Circulating naive T cells number approximately \( 10^9 \) (37) but can exchange, at different time scales, with those that reside in peripheral tissue, which may carry their own proliferation and death rates. The effective pool that is ultimately sampled is thus difficult to estimate, but measurements show that the theoretical number of different clones is much larger than the total number of naive T cells, which is in turn much greater than the total number of different T cell clones actually in the body \( (Q \gg N \gg \hat{C}) \). Regardless of the precise values of the discrete quantities \( Q, N, \hat{C} \), they are related to the discrete clone counts \( c_k \) via

\[
\hat{C} = \sum_{k \geq 1} \hat{c}_k \ll Q \quad \text{and} \quad \hat{N} = \sum_{k \geq 1} k \hat{c}_k.
\]

As depicted in Figure 2, each distinct clone \( i \) (with \( 1 \leq i \leq Q \)) is characterized by an immigration rate \( \alpha_i \) and a per cell replication rate \( r_i \). The immigration rate \( \alpha_i \) 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 both the numbers of theoretically possible \( (Q \gg 1) \) and observed \( (\hat{C} \gg 1) \) clones are extremely large, we can define a continuous, normalized probability density \( \pi(\alpha, r) \) from which immigration and proliferation rates \( \alpha \) and \( r \) of a randomly chosen clone are drawn. This means that the probability that a randomly chosen clone has an immigration rate between \( \alpha \) and \( \alpha + d\alpha \)...

\[
\text{TABLE 1} | \text{Model parameters} \theta \text{ and hyperparameters} \theta_\beta.
\]

| (Hyper) Parameters \( \theta \) | \text{definition} |
|-----------------------------|-----------------|
| \( \alpha \in \mathbb{R}^+ \) | naive T cell production rate |
| \( \alpha \in \mathbb{R}^+ \) | mean production rate across all possible Q TCRs |
| \( r \in [0, R] \) | maximum proliferation rate across all possible Q TCRs |
| \( R \in \mathbb{R}^+ \) | maximum proliferation rate of all possible Q TCRs |
| \( w \in [0, 1] \) | dimensionless width of box distribution of \( r \) |
| \( \mu^* > R \) | naive T cell death rate at steady state |
| \( \eta \in [0,1] \) | blood subsampling fraction |

The dimensional parameters associated with our mechanistic population model. Hyperparameters such as \( \alpha, r, R \), \( w \) define the probability distribution or heterogeneity in the underlying rate parameters \( \alpha \) and \( r \). In our analyses, we typically nondimensionalize by normalizing all rates by \( R \), the maximum proliferation rate across all clones.
In Dessalles et al. (10) exact results were derived for the steady-clones, \( fi \) models, where both rates are replication rates allows us to go beyond typical consistent solution. The heterogeneity in the immigration and unbounded numbers of naive T cells and is necessary for a self-process, the upper bound \( R \) independent but a function of the total population \( \hat{N} \). The 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 (24, 31, 38). The for a fully stochastic neutral BDI model \( \alpha_i \) form of the regulation will not qualitatively affect our findings since we will ultimately be interested in only its value for all clones, \( \alpha_i = \alpha \) and \( r_i = r \) for all \( i \).

Finally, we assume the per cell death rate \( \mu(\hat{N}) \) is clone-independent but a function of the total population \( \hat{N} \). 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 (24, 31, 38). The specific form of the regulation will not qualitatively affect our findings since we will ultimately be interested in only its value \( \mu(N^n) \equiv \mu^* \) at the mean steady state population \( N^* \).

**Mean-Field Approximation of the BDI Process**

The exact steady-state probabilities of configurations of the discrete abundances \( \hat{c}_k \) for a fully stochastic neutral BDI model with regulated death rate \( \mu(\hat{N}) \) were recently derived (10). In Dessalles et al. (10) exact results were derived for the steady-state probability \( P(\hat{c}_1, \hat{c}_2, ..., \hat{c}_k) \) under uniform immigration, proliferation, and death rates \( \alpha, r, \) and \( \mu \), respectively. The significant contribution of this paper is that we go beyond the neutral model (equal immigration, proliferation, and death rates for all clones) by allowing for heterogeneous distributions of these rates. To incorporate TCR-dependent immigration and replication rates in a non-neutral model, we must consider distinct values of \( \alpha_i \) 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 effective number of naive T cells \( (N \sim 10^8 \text{ to } 10^{11}) \) 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 (39).

**i) Deterministic Approximation for the Total Population and the Effective Death Rate**

To implement the mean-field approximation in the presence of a general regulated death rate \( \mu(\hat{N}) \), we start by writing the deterministic, “mass-action” ODE for the mean number of cells \( n_{\alpha r}(t) \) with a realized immigration rate \( \alpha \) and proliferation rate \( r \) in a BDI process

\[
\frac{dn_{\alpha r}(t)}{dt} = \alpha + rn_{\alpha r}(t) - \mu(\hat{N}(t))n_{\alpha r}(t).
\] (2)

Next, we define and exploit the density of realized values of \( \alpha \) and \( r \). Since \( Q \gg 1 \), the number of TCRs that are associated with an extremely low probability sequences may never be formed during the organism’s lifetime. A strict lower bound on \( Q \) is the actual number of distinct clones \( \hat{C} \) in an entire organism \([\hat{C} \sim 10^6 \text{ to } 10^8 \text{ for humans } (1, 6, 32-34)].\)
between $r$ and $r + dr$ is denoted $Q \pi(\alpha, r) dr$, where $\pi(\alpha, r)$ is a normalized density that describes how these realized values of $\alpha$ and $r$ are distributed. Our model for the total mean number $N(t)$ of naive T cells can then be estimated as a weighted integral over all $n_{\alpha r}(t)$

$$N(t) = \int_0^A d\alpha \int_0^R dr \, n_{\alpha r}(t) \pi(\alpha, r).$$  \hspace{0.5cm} (3)

Note that the limits of the integration above can equivalently be taken as $A, R \to \infty$ as long as $\pi(\alpha, r) = 0$ when $\alpha > A$ or $r > R$. At steady-state, the solution to Eq. 2 can be simply expressed as

$$n_{\alpha r}^* = \frac{\alpha}{\mu(N^*) - r}$$  \hspace{0.5cm} (4)

in which $N^*$ is the predicted steady-state value of $N(t)$ as $t \to \infty$. Thus, upon weighting Eq. 4 over all possible values of $\alpha$ and $r$, we find

$$N^* = \int_0^R dr \int_0^\infty d\alpha \frac{\alpha \pi(\alpha, r)}{\mu(N^*) - r}$$  \hspace{0.5cm} (5)

a self-consistent equation for $N^*$ which depends implicitly on the parameters that define the distribution $\pi(\alpha, r)$. Eq. 5 clearly shows why a finite cutoff $\pi(\alpha, r > R) = 0, R < \mu(N^*)$ is required since the integral diverges if $\pi(\alpha, r \geq \mu(N^*)) > 0$. However, as long as $\pi(\alpha, r)$ decays faster than $1/\alpha^2$, the $\alpha$-integration converges with an explicit cutoff $A$.

We will first assume that $\alpha$ and $r$ are uncorrelated and that the distribution factorises: $\pi(\alpha, r) = \pi_{\alpha}(\alpha) \pi_r(r)$. Then, the self-consistent effective steady state death rate $\mu^* = \mu(N^*)$ depends only on the combination

$$\frac{N^*}{\bar{\alpha} Q} = \int_0^R dr \pi_r(r) \left( \frac{\alpha}{\mu^* - r} \right)$$

where

$$\bar{\alpha} = \int_0^\lambda \alpha \pi_r(\alpha) d\alpha$$

is the mean immigration rate across all possible clones. To simplify subsequent notation, we normalize all rates by the maximum proliferation rate $R$. To avoid population blow-up, we impose that the maximum proliferation is smaller than the steady-state death rate $R < \mu^*$. By measuring time in units of $1/R$, we redefine $r/R \to r \leq 1$, $\alpha/R \to \alpha$, $\bar{\alpha}/R \to \bar{\alpha}$, $\mu^*/R \to \mu^*$, and $R^2 \pi(\alpha, r) \to \pi(\alpha, r)$ so that these quantities are now dimensionless, unless otherwise explicitly stated. The steady-state self-consistent condition becomes

$$\frac{N^*}{\bar{\alpha} Q} = \frac{\lambda}{\bar{\alpha}} = \int_0^1 dr \frac{\pi_r(r)}{\mu^* - r}.$$  \hspace{0.5cm} (6)

Since the effective $Q$ is a large, uncertain number, we parameterize our model in terms of $\bar{\lambda} \equiv N^*/Q$, the total steady state naive T cell population normalized by the total possible number of clones $Q$. It is sometimes deemed a measure of the “coverage” of the entire repertoire (6). Values of $N^*$ and $Q$ that are consistent with measurements and physiologic expectations give $\bar{\lambda} \ll 1$. Once $\bar{\lambda}/\bar{\alpha}$ and $\pi_r(r)$ are estimated, we can self-consistently determine $\mu^*$ from Eq. 6. Besides $\bar{\lambda}/\bar{\alpha}$, the self-consistent value of $\mu^*$ will also depend on the function $\pi_r(r)$. Note from the form of Eq. 6, the self-consistent $\mu^*$ is inversely related to $\bar{\lambda}$.

ii) Mean-Field Model of Clone Counts

Given a relationship such as Eq. 6 that determines $\mu^*$, we can explicitly develop a model that quantifies naive T cell subpopulations according to their immigration and proliferation rates $\alpha$ and $r$. For a given, realized value of $\alpha$ and $r$, we denote the expected number of clones of size $k$ with these immigration and proliferation rates by $c_k(\alpha, r)$. The mean-field equations for the dynamics of these mean clone counts in the neutral model were derived in (39, 40) and are reviewed in Section 1 of the Supplementary Material. In a neutral model, we assume that all clones $k$ carry the same rates $\alpha$ and $r$ so that the mean field evolution equation for $c_k(\alpha, r)$ is given by solving (38, 39)

$$\frac{dc_k(\alpha, r)}{dt} = \left[ \frac{c_{k-1}(\alpha, r) - c_k(\alpha, r)}{\lambda^*} \right] + r \left[ (k-1)c_{k-1}(\alpha, r) - kc_k(\alpha, r) \right] + \mu(k) \left[ (k+1)c_{k+1}(\alpha, r) - kc_k(\alpha, r) \right],$$  \hspace{0.5cm} (7)

along with the constraint $\sum_{k=0}^\infty c_k(\alpha, r) = c_0 + \sum_{k=1}^\infty c_k(\alpha, r) = Q$. Note that $c_k(\alpha, r)$ and $n_{\alpha r}$ are related via $\sum_{k=1}^\infty kc_k(\alpha, r) = n_{\alpha r}$. We use the notation $\bar{c}_k$ to denote the predicted clone counts derived from our mathematical model to distinguish them from measured clone counts $\tilde{c}_k$. Equation 7 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 $N = \sum_{k=1}^\infty kc_k(\alpha, r)$.

The steady-state solution of Eq. 7 follows a negative binomial distribution with parameters $\alpha/r$ and $r/\mu^* < 1$ (10, 39)

$$c_{k+1}(\alpha, r, \mu^*) = Q \left( 1 - \frac{r}{\mu^*} \right)^{\alpha/r} \left( \frac{r}{\mu^*} \right)^k \frac{1}{k!} \prod_{\ell=0}^{k-1} \left( \frac{\alpha}{r} + \ell \right).$$  \hspace{0.5cm} (8)

The predicted number of absent clones is $c_0 = Q - \sum_{k=1}^\infty c_k(\alpha, r, \mu^*)$. The solution 8 depends implicitly on the parameter $\bar{\lambda}/\bar{\alpha}$ through $\mu^*$ determined by Eq. 6. Although $c_k(\alpha, r, \mu^*)$ has not yet been averaged over $\alpha, r$, it also implicitly depends on $\bar{\lambda}$ and the parameters that define $\pi_r(r)$ through $\mu^*$ and Eq. 6. Specifically, larger $\bar{\lambda}$ leading to smaller $\mu^*$ results in a more slowly decaying $c_k(\alpha, r, \mu^*)$ as a function of $k$.

This behavior will be propagated through subsampling and averaging over $\alpha$ and $r$.

**Subsampling**

Unless an animal is sacked and its entire naive T cell population is 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 $\eta$ 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 $\eta$ 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 clone is represented by $\ell$ cells in an organism. If $N^* \eta \gg \ell$, the probability that $k$ cells are randomly sampled from the same clone approximately follows a binomial distribution with parameters $\ell$ and $\eta$ (40–44)

$$P[k|\ell] = \binom{\ell}{k} \eta^k (1-\eta)^{\ell-k}, \quad k \leq \ell. \quad (9)$$

The associated mean sampled clone count $c_k^*$ depends on the predicted whole-organism clone count and $P[k|\ell]$ via the formula

$$c_k^*(\alpha, r, \mu^*, \eta) = \sum_{l \geq k} c_l(\alpha, r, \mu^*, \eta) \cdot P[k|\ell]$$

$$= \sum_{l \geq k} c_l(\alpha, r, \mu^*) \binom{\ell}{k} \eta^k (1-\eta)^{\ell-k}. \quad (10)$$

where $c_l(\alpha, r, \mu^*)$ is determined by Eq. 8. Explicitly performing the sum in Eq. 10 yields the sampled clone count

$$c_k^*(\alpha, r, \mu^*, \eta) = \frac{Q}{k!} \left( \frac{\eta \mu^*}{(1-\eta)(r/\mu^*)} \right)^k \frac{1-r/\mu^*}{(1-(1-\eta)(r/\mu^*)} \prod_{j=0}^{k-1} \left( \frac{\alpha}{r} + j \right). \quad (11)$$

The total expected number of clones in the sample (the richness) can be found via direct summation:

$$C^*(\alpha, r, \mu^*, \eta) = \sum_{k=1}^{\infty} c_k^*(\alpha, r, \mu^*, \eta)$$

$$= Q \left[ 1 - \left( 1 - \frac{r/\mu^*}{1-(1-\eta)(r/\mu^*)} \right)^{\alpha/r} \right]. \quad (12)$$

As shown in Figure 3, random subsampling greatly affects the observed clone counts, with small sampling fractions $\eta$ leading to fast decay in $k$ of $c_k^*(\alpha, r, \mu^*, \eta)$ and shifting $c_k^*$ at large $k$ to much smaller values of $k$ while reducing the values of $c_k$ for small $k$ (42). Note that setting $\eta = 1$ in Eq. 11 leads to Eq. 8, the whole-body clone count. In Figures 3A, B we plot results from our model using two very different dimensionless parameter sets, $\alpha = 10^{-5}$, $r = 1/2$, $\lambda = 0.01$, and $\alpha = \lambda = 10$, $r = 1/2$, to generate two qualitatively different patterns of neutral model clone counts $c_k$. If the subsampling $\eta \ll 1$ is sufficiently small, the resulting $c_k^*$ corresponding to the two qualitatively different $c_k$ can appear similar. This implies that small sampling fractions make the estimation of whole-body clone counts from sampled data somewhat ill-conditioned, i.e., different whole-body clone counts, upon sampling, may yield similar sampled clone counts. Although sampling can strongly affect the inference of $c_k$, immigration and proliferation rate distributions may also affect the observed clone count as we investigate below.

**Heterogeneity and Determination of $\pi(\alpha, r | \Theta_0)$**

The fundamental result given in Eq. 11 applies only to the clone count density in a neutral model in which the immigration and
proliferation rates are $\alpha$ and $r$ for all clones. We now average the sampled clone counts $c_k(\alpha, r, \mu^*, \eta)$ (Eq. 11) and the richness $C'(\alpha, r, \mu^*, \eta)$ (Eq. 12) over a distribution of immigration and proliferation rates $\pi(\alpha, r)$ to capture the heterogeneity across TCR clones. This final result can then be compared with experimentally measured clone counts. Recall that $\pi(\alpha, r)$ can depend on hyperparameters $\theta_0$ that define the shape of $\pi$. We then explicitly denote the distribution by $\pi(\alpha, r|\theta_0)$.

Once $\pi(\alpha, r|\theta_0)$ is defined, we can weight sampled clone counts accordingly. For example, one may assume $\theta_0 = \{\bar{\alpha}, w\}$, with each of the two hyperparameters defining $\pi(\alpha, r|\theta_0) = \pi_{\alpha}(\alpha|\bar{\alpha})\pi_{r}(r|\bar{\alpha})$, leading to

$$c_k^*(\mu^*, \eta, \theta_0) = \{\bar{\alpha}, w\} = \int_0^\infty d\alpha \int_0^1 dr \pi(\alpha, r|\theta_0) c_k(\alpha, r, \mu^*, \eta).$$

### i) Proliferation Rate Heterogeneity

First, we consider a distribution of TCR sequence-dependent proliferation rates. Since TCR-antigen affinity depends on the receptor amino-acid sequence, the rate of T cell activation and subsequent proliferation can be clone-specific (31, 45). Thus, the specific interactions between TCRs and low-affinity MHC/self-peptide complexes maps to a distribution of proliferation rates among all the Q possible clones. Since there are no data (known to us) that can be used to infer this mapping or the specific shape of $\pi_{r}(r|\bar{\alpha})$, we assume, for simplicity, a simple uniform “box” distribution centered about a mean value $\bar{r} = 1/2$:

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

where $0 \leq w \leq 1$ represents the relative width of the uniform box distribution. The minimum and maximum dimensionless proliferation rates in this distribution are then $1/2 - w/2$ and $1/2 + w/2$, respectively. The dimensionless self-consistency condition (Eq. 6) thus yields

$$\mu^* = \frac{\left(\frac{1}{2} + \frac{w}{2}\right) e^{\frac{\lambda - \bar{\alpha}}{\alpha}} - \left(\frac{1}{2} - \frac{w}{2}\right)}{e^{\frac{\lambda - \bar{\alpha}}{\alpha}} - 1}$$

(14)

To understand the effects of proliferation rate heterogeneity we begin by considering it effects on whole-organism ($\eta = 1$) clone counts. Since the function $c_k(\alpha, r, \mu^*)$ defined by Eq. 8 contains the exponentially decaying term $(r/\mu^*)^k$, a fixed dimensionless value of $\mu^*$ and $r = 1/2$ leads to an exponential decay in $c_k$ in $k$. However, if $w > 0$, different values of $r$ and $\mu^*$ contribute to this decay term, yielding nontrivial behavior and a much slower decay as seen in Figure 4 for $\lambda/\bar{\alpha} = 8, 80$ and different values of $w$.

### ii) Immigration Rate Heterogeneity

Next, we use previous studies that predict V(D)J recombination frequencies associated with each TCR sequence to construct a distribution $\pi_{\alpha}(\alpha)$ for the TCR sequence-dependent thymic output. A statistical model for differential V(D)J recombination in humans is implemented in the Optimized Likelihood estimate of immunoGlobulin Amino-acid sequences (OLGA) software (28), which is an updated version of the Inference and Generation of Repertoires (IGoR) software (20). Below, we estimate $\pi_{\alpha}(\alpha|\bar{\alpha})$ by sampling a large number of TCRs from OLGA that draws sequences according to their generation probability. Our working assumption is 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.

Both IGoR and OLGA can be used to generate the probabilities corresponding to each drawn sequence but this

![FIGURE 4](image-url)
requires significant computational time and memory. Equivalently, since the sequence draws are proportional to the underlying probabilities, we simply drew $N_*$ sequences and counted the frequencies of each amino acid sequence. Out of $N_*$ sequence draws from IGoR or OLGA, there will be $C_*$ distinct amino sequences (the richness of the drawn sequences). Since some sequences are drawn $j > 1$ times, $C_* < N_*$. If $b_j$ distinct sequences are drawn $j$ times, and the maximum observed frequency $\max[j] = J = 52,294$ for the alpha chain and $C_* = 875,920,705 \approx 8.76 \times 10^8$ and $J = 6430$ for the beta chain.

We model the effective immigration rate of a TCR sequence drawn $j$ times to be proportional to $j$ so that $\alpha_j \equiv \alpha_j / j$. To fix the proportionality $\alpha_j$, we identify the mean immigration rate averaged across the $C_*$ observed sequences with the mean physiological rate $\bar{\alpha}$

$$\alpha = \frac{\bar{\alpha}}{C_*} = \bar{\alpha}$$

(15)

to find $\alpha_j = \bar{\alpha} C_* / N_*$ and thus

$$\alpha_j = \frac{\bar{\alpha} C_* / N_*}{C_*} = \frac{\bar{\alpha} j}{N_* / C_*}.$$  

(16)

The frequencies $j$ of the drawn realization of clones are plotted in decreasing order against the $C_*$ distinct sequences in Figures 5A, B. From these frequencies $j$ and the number of sequences $b_j$ exhibiting them, we approximate averages of any function $\gamma(\alpha)$ over $\pi_\alpha(\alpha|x)$ by taking a sum over the values $\alpha_j$:

$$\int \pi_\alpha(\alpha|x) \gamma(\alpha) = \sum_{j=1}^J \frac{b_j}{C_*} \gamma(\alpha_j),$$  

(17)

Alternatively, when drawing sequences IGoR and OLGA (using the Pgen feature) one can also directly output their probabilities $p$, whose values would be proportional to the frequency $j$ if large numbers of sequences are drawn as described above. We can use these countable sequences and probabilities to construct $\alpha$ and $\pi_\alpha(\alpha)$ by defining $\alpha_\alpha = \alpha Q C_*, p^T = p_T$ where $p_T = \sum_{i=1}^J b_i$. By plotting the values of $p_T$, we arrive at a distribution similar to that shown in Figure 5. In this case too, we find that a large number of low-probability sequences dominates the averaging of clone counts using the distribution of immigration rates constructed using IGoR/OLGA.

Now that we have specified the distributions for $\pi_\alpha(\alpha | \bar{\alpha})$ and $\pi_\alpha(r | w)$, we can compute the mean, sampled, immigration- and proliferation-averaged clone counts and compare them with measurements. The full formula for the immigration and proliferation rate-averaged clone counts under subsampling is

$$c_k(\bar{\alpha}, \mu^*, w, \eta, \gamma) = \int_0^\infty \int_0^1 dr \int_0^1 d\alpha \pi_\alpha(\alpha | \bar{\alpha}) \pi_\alpha(r | w) c_k(\alpha, r, \mu^*, \eta, \gamma)$$

$$= \frac{Q}{k!} \sum_{j=1}^{\max[j]} b_j \int_{(1-w)/2}^{(1+w)/2} dr \left( \frac{\eta r / \mu^*}{1 - (1 - \eta) r / \mu^*} \right)^k \left( \frac{1 - r / \mu^*}{1 - (1 - \eta) r / \mu^*} \right) \prod_{i=0}^{n-1} \left( \frac{\alpha_i + i}{r} \right),$$

(18)

where $\alpha_i$ is given by Eq. 16 and $\mu^*$ is given by Eq. 14. Eq. 18 is our “full model” from which we make predictions of clones count-related quantities and compare them with data. Using this
expression, we can mathematically study the importance of the heterogeneities in $\alpha$ and $r$ by comparing predictions from simple forms of $\pi_1(\alpha/\bar{\alpha})$ and $\pi_2(r/\bar{w})$, as presented in Section 2 of the Supplementary Material to those derived from $\pi(\alpha, r) = \delta(\alpha - \bar{\alpha})\delta(r - 1/2)$ of the neutral model.

From Figure 5, observe that $b_1 \gg b_{j>1}$. In fact, a majority of the naive T cell population is comprised of clones that are produced only once. The linear-scale insets also show a long tail indicating a large number of clones that are generated few times. Thus, for sufficiently small $\bar{\alpha}$, our formulae for $\bar{c}_k$ and all subsequent quantities can be approximated by taking the $\bar{\alpha}/r \ll 1$ limit. As we show in Section 3 of the Supplementary Material, such a simpler expression remains highly accurate, provided the dimensionless $\bar{\alpha} < 10^{-2}$, and allows efficient computation. This implies that the full result arising from averaging $c^*_k(\alpha, r, \mu^*, \eta)$ over $\pi_2(\alpha/\bar{\alpha})$ can also be approximated by using a single effective value $c^*_k(\bar{\alpha}, r, \mu^*, \eta)$, supporting our overall conclusion that predicted heterogeneity in human T cell immigration rates do not appreciably influence clone count distributions. While physiological distributions $\pi_2(\alpha/\bar{\alpha})$ do not yield clone counts appreciably different from those of a neutral immigration model, small changes in proliferation rate heterogeneity $w$ can significantly affect the clone count structure $c^*_k$. Nonetheless, for completeness, we will perform the full summation over $\alpha$ (Eq. 18). All parameters, hyperparameters, and variables used in our modeling and data analysis are listed in Tables 1, 2.

### RESULTS AND ANALYSIS

Before performing a quantitative comparison with measured clone counts from Oakes et al. (12), we discuss the qualitative features of our model and typical physiological parameter ranges. While even the basic model parameters are difficult to measure, our nondimensionalized model unifies the mechanisms and concepts common to the maintenance of diversity in the T cell repertoire across different organisms.

When considering the data, we observe that even after significant subsampling, there are appreciable clone counts at reasonably large clone sizes $k$, whereas the unsampled clone counts decay exponentially in $k$ with rate $\log(\mu^*/r)$. Even though $r$ may take on a range of values, as determined by $\pi_2(r)$, the slowest decay of $c_k$ arises from the largest possible values of $r$. Thus, a larger proliferation rate heterogeneity $w$ will generally yield a long-tailed $c_k$, as illustrated in Figure 4. Since the data we analyze are derived from human samples, we will use the following arguments as a rough guide to the relevant range of parameters:

- The average total number of naive T cells is not completely known but is estimated to be about $N^\ast \sim 10^{11}$ (35). However, the circulating population in the peripheral blood is approximately two orders of magnitude smaller. These circulating naive T cells nonetheless exchange with those in the much larger population in the lymph and other tissues. The timescale of this exchange (relative to the age of the organism being sampled or the intersample times) will determine the effective statistically accessible $N^\ast$ relevant for sampling clone counts $c_k$. We will use an order-of-magnitude estimate on the lower range of measurements and estimate $N^\ast \sim 10^{10}$.

- The theoretical total possible number $Q$ of TCRs of either alpha or beta chains may be in the range $10^{13} - 10^{18}$ (46), but the actual number of clones with immigration rate $\bar{\alpha}$ that allows it to be produced even once in a lifetime is much more relevant and probably much smaller. Thus, the effective value of $Q$ may reside at the lower range, leading to $\bar{\alpha} \equiv N^\ast/Q \sim 10^{-4} - 10^{-2}$.

- The average (dimensional) immigration rate per clone $\bar{\alpha}$ can be deduced from the total thymic output of all clones $\bar{\alpha}Q$, which has been estimated across a wide range of values $\bar{\alpha}Q \sim 10^{7} - 10^{8}$/day (29, 47-50). If we use an effective repertoire size of $Q \sim 10^{13} - 10^{14}$, the average per clone immigration rate becomes $\bar{\alpha} \sim 10^{-2} - 10^{-5}$/day.

- The mean proliferation rate $r$ is difficult to measure but has been estimated to be on the order of $r \sim 10^{-4} - 10^{-2}$/day (29). If we nondimensionalize using $R = 2r$, the dimensionless $\bar{\alpha} \sim 10^{-4} - 10^{-1}$.

- The sampling fraction $\eta$, although in principle determined experimentally, is also hard to quantify due to the uncertainty in $N^\ast$. Blood sampling volume fractions from humans are typically $\sim 10^{-3}$; however, in recent experiments (12) the number of enumerated sequences $\sim 10^7$, which, given rough estimates of effective $N^\ast \sim 10^{10}, 10^{11}$, yield $\bar{\eta} \sim 10^{-6} - 10^{-4}$. Due to this uncertainty in $\bar{\eta}$, we will explore different fixed values of $\eta$ around $10^{-3}$.

### TABLE 2 | Model variables and their definitions.

| variables | definition |
|----------|------------|
| $Q \equiv Z^+$ | theoretical number of possible TCRs $\sim 10^{13} - 10^{18}$ (36) |
| $\hat{N} \equiv Z^+$ | number of naive T cells in organism $\sim 10^{10} - 10^{11}$ (5) |
| $N(\theta) \equiv R^+$ | number of naive T cells from model |
| $N^\ast \equiv \eta N(\theta) \equiv R^+$ | subsampled number of naive cells from model |
| $N_k \equiv Z^+$ | number of draws from IGoR/OLGA |
| $\mathcal{C} \equiv Z^+$ | total number of clones in organism (richness) $\sim 10^3 - 10^{6}$ (36) |
| $\mathcal{C}_4 \equiv Z^+$ | total number of sampled clones (sampled richness) |
| $\mathcal{C}(\theta, \eta) \equiv R^+$ total number of clones in organism from model |
| $\mathcal{C}_4(\theta, \eta) \equiv R^+$ total number of sampled clones from model |
| $\mathcal{C}_0 \equiv Z^+$ | number of different sequences drawn from IGoR/OLGA |
| $\hat{c}_k \equiv Z^*$ | discrete number of clones of size $k$ |
| $c_k(\theta) \equiv R^+$ | model of number of clones containing $k$ cells |
| $c_k^* \equiv Z^*$ | discrete number of clones of size $k$ in sample |
| $c_k^{\ast}(\theta, \eta) \equiv R^+$ modeled number of sampled clones containing $k$ cells |
| $\hat{q}_k = \frac{k^2 c_k^*}{\mathcal{C}_4(\theta, \eta)} \equiv [0, 1]$ | fraction of all sampled cells in clones of size $k$ |
| $q_k(\theta, \eta) = \frac{k^2 c_k^{\ast}(\theta, \eta)}{\mathcal{C}_4(\theta, \eta)} \equiv [0, 1]$ | modeled fraction of all sampled cells in clones of size $k$ |

The variables with ‘*’ denote measured numbers, while populations written as functions of parameters $\theta$ are those predicted from our model [the dimensionless parameters used in our model are $\theta = (\alpha, r)$]. The probability distributions $\pi(\alpha, r/\theta)$ are defined by hyperparameters $\theta_0$ (the dimensionless hyperparameters used in this study are $\theta_0 = (\alpha, w)$). Upon averaging predicted quantities such as $c_k^*_j(\alpha, r)$ over $\pi(\alpha, r/\theta)$ we find $c_k^*_j(\theta_0)$.
Using the above guide for reasonable parameter ranges, we now consider fitting our results in Eqs. 18, S9-S14 to some of the available data (12). Before doing so, note that although the log-log plots shown in Figures 1A, B provide a simple visual for the unknown repertoire size $Q$ since $c^*_k$ is proportional to $Q$. The implicit factor of $Q$ in $c^*_k$ from Eq. 11 cancels the explicit $Q$ in the denominator of Eq. 19 so that $f^*_k$ as well as $c^*_k/C^*$ depend on $Q$ only through the determination of $\mu^*$ through $\lambda \equiv N^*/Q$ in Eq. 6.

Our mathematical framework provides only mean sampled clone counts while each sample of the data represents one realization. Large sample-to-sample variations in the clone counts would render the fitting less informative, but these large variations were not seen in the triplicate samples in Oakes et al. (12). Mechanistically, we expect that for large $k$ the number of cells contributing to $f^*_k$ 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 $f^*_k$. Large clones are also likely to include memory T cells that have been produced after antigen stimulation of specific clones. Memory T cells are difficult to accurately distinguish from naive T cells (12) but we will see that large $k$ components of $f^*_k$ negligibly influence the fitting. We can now compare our model $f^*_k(\alpha, \lambda, w, \eta)$ with the data $f^*_k$ (data) by constructing the error

$$H(\alpha, \lambda, w, \eta) = \sum_{k=1}^{\infty} \left| f^*_k(\text{data}) - f^*_k(\alpha, \lambda, w, \eta) \right|^2$$

and exploring how it depends on the parameters $\alpha$, $\lambda$, $w$, and $\eta$. Our goal is to find relationships among the parameters $\lambda$, $\alpha$, and $w$ that minimize $H(\alpha, \lambda, w, \eta)$.

In Figures 6A–C the data $f^*_k$ (data) were derived from the average of three samples of beta chain CD4 sequences from one

**FIGURE 6** The error $H(\alpha, \lambda, w, \eta)$ plotted as a function of $\alpha$ (on a log10 scale) and $\lambda$. Darker colors represent smaller values of error as shown by the scale bar on the right. The data used are the clone counts of beta chain sequences of naive CD4 cells from one patient, averaged over three samples. Panels (A–C) use the simple neutral model (Eqs. S9 and S10) and sampling fractions $\eta = 10^{-4}$, $10^{-5}$, and $10^{-6}$, respectively. Since $\alpha$ is on a log scale, the error is minimal along a line $\lambda_{min} \sim \alpha$; the error does not change appreciably along this path and only slightly decreases as $\lambda$ and $\alpha$ become smaller. For the neutral model ($w = 0$), the error is very sensitive to the sampling fraction $\eta$. Here, a fixed, physiologically reasonable value of $\alpha$ results in a minimizing $\lambda_{min}$ that is unreasonably large, in excess of one and that does not agree well with our expectations of $\lambda = N^*/Q < 1$. Panels (D–F) show results for the distributed proliferation rate model at full width ($w = 1$). In this case, the errors are insensitive to the specific choice of $\eta$ and the minimizing $\lambda_{min}$ values are much smaller, consistent with our estimates of $N^*$ and repertoire size. For $w = 1$, the values of the errors $H$ are also smaller along the $\lambda_{min} \sim \alpha$ minimum valley.
These data, were used to compute and plot the error $H(\bar{\alpha}, \lambda, w = 0, \eta)$ as a function of $\lambda$ for various values of $\bar{\alpha}$ using the neutral model ($w = 0$, Eq. S9 in Section 2 of the Supplementary Material). For reasonable values of dimensionless $\bar{\alpha} = 10^{-5} - 0.01$ and sampling fractions $\eta = 10^{-4}, 10^{-5}, 10^{-6}$, we find that the value of $\lambda$ that minimizes $H(\bar{\alpha}, \lambda, w = 0, \eta)$, $\lambda_{\text{min}}$, is typically $O(1)$ or larger. In Figures 6D–F we use the full-width distribution $\pi(\eta|w = 1)$ to show the error for the same data using the same sampling fractions $\eta = 10^{-4}, 10^{-5}, 10^{-6}$. Note that the values of $\lambda_{\text{min}}$ are significantly larger than those in found using $w = 0$ in Figures 6A–C and that the results are rather insensitive to the sampling fraction $\eta$. These smaller values of $\lambda_{\text{min}}$ are more consistent with known physiological understanding. Thus, the distributed proliferation rate model provides a much more self-consistent fit to the data than the fixed proliferation rate neutral model. Figure 6 also reveals that the values of $H$ along the minimum valley are nearly constant, only slightly decreasing as $\bar{\alpha} \to 0$. For each value of $\alpha$ we can identify the corresponding $\lambda_{\text{min}}$ that minimizes $H$. However since the values of $H(\bar{\alpha}, \lambda_{\text{min}}, w = 0, \eta)$ for each $(\bar{\alpha}, \lambda_{\text{min}})$ pair do not change appreciably, we cannot independently determine both.

An alternate representation is shown in Figure 7 where the relationship between $\alpha$ and $\lambda_{\text{min}}$ is seen to be approximately linear for both the neutral model ($w = 0$) and the heterogeneous, full-width model ($w = 1$). The color shading represents the corresponding value of $H(\bar{\alpha}, \lambda_{\text{min}}, w, \eta)$. One major observation is that the full-width case yields values of $(\bar{\alpha}, \lambda_{\text{min}})$ that are closer to measured and expected physiological values and that these results are also less sensitive to $\eta$ compared to those of the neutral case. On the other hand, although the variation in $H$ is negligible across $\bar{\alpha}$ in both cases, the fully heterogeneous model ($w = 1$) carries a slightly larger error than the neutral one ($w = 0$). This is solely a consequence of our use of $f_2^\ast$ which weights the small $k$ values significantly more in the fitting.

Since experimentally we expect small $\lambda$, we also investigate whether small errors $H$ emerge for values of $(\bar{\alpha}, \lambda_{\text{min}} \ll 1)$ at intermediate $0 < w < 1$. In Figure 7C, we plot $\lambda_{\text{min}}$ as a function of $w$ for various values of $\bar{\alpha}$. Note that even small $w$ significantly reduces, relative to the neutral case, the corresponding $\lambda_{\text{min}}$. However, if our target is $\lambda_{\text{min}} \sim 10^{-4}-10^{-3}$, the required $w$ can become quite large. These results indicate that more heterogeneity is associated with more realistic values of the experimentally observed values of $N/Q$.

Finally, to explore the dependence of the error on the proliferation rate heterogeneity $w$, we fix $\bar{\alpha}, \lambda, \eta$, and plot $H(\bar{\alpha}, \lambda, w, \eta)$ as a function of $w$. Figure 8 shows that the $H$-minimizing $w$ is very sensitive to $\lambda/\bar{\alpha}$: for fixed $\eta$, as $\lambda/\bar{\alpha}$ is decreased the error is lowest for larger proliferation heterogeneity $w$. The minimum value of $H(\bar{\alpha}, \lambda, w, \eta)$, however, is rather insensitive to $\lambda/\bar{\alpha}$ for all chosen $\eta$. Hence, near-optimal solutions with $\lambda \ll 1$ can be found when the proliferation rate heterogeneity $w$ is appreciable. Using the parameters associated with the minima in Figure 8A ($\eta = 10^{-5}$), we plot our predicted $f_2^\ast$ against the data $f_2^\ast$(data) in Figure 9. As can be seen, when proliferation rate heterogeneity is allowed, the best-fits have small error and are found using realistic values, $\lambda \ll 1$. Note that most of the information in the data lies in how $f_2^\ast$(data) decreases over the first few values of $k$. The neutral model ($w = 0$) fits best for small values of $k$, but the corresponding values of $\lambda$ and $\alpha$ are too large and small, respectively. The goodness of fit of our model to the data depends mostly on the predicted initial decreases in $f_2^\ast(\alpha, \lambda, w, \eta)$. The constraints among the parameters $\lambda, \bar{\alpha}, w, \eta$ derived from our model and can be applied to different clone counts such as the data shown in Figure 1. However, due to the ill-conditioning when $\eta \ll 1$, the differences in these constraints across different data sets do not vary appreciably are only quantitatively different. Generally, the more rapidly decaying a clone count, the smaller the $w$, the smaller the $\eta$, of the larger the $\lambda$, all else being equal.
DISCUSSION

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.

Distinct T Cell Components

It is known that naive T cells can change in time, with recent thymic emigrants evolving into mature naive T cells that carry different proliferation and death rates (51). For simplicity, we have assumed a single naive T cell compartment. To incorporate naive T cell evolution, we can allow the distribution \( p(r) \) to evolve in time to reflect the relative abundances of T cell subpopulations, or, one can explicitly include multiple compartments, with cells from a recent emigrant compartment transitioning into a mature compartment. Each compartment would be described by its own steady-state death rates, clone counts, and distributions of proliferation rates. An analysis of a related sequential cell state transition model has been developed for clonal tracking in hematopoiesis (41).

Factorization of \( \pi (\alpha, r) \)

For mathematical tractability, we have assumed \( \pi(\alpha, r|\theta_0) = \pi_{\alpha}(\alpha|\theta) \pi_r(r|w) \). Given the typical physiological values of \( \alpha \), the clone count formulae derived from our model can be accurately approximated by a single value of \( \alpha \). Thus, we expect that the immigration rate distribution can be...
approximated by $\pi_q(\alpha/\bar{\alpha}) = \delta(\alpha - \bar{\alpha})$. This allows further approximation of our formulae as shown in Section 3 of the Supplementary Material. In Section 4 of the Supplementary Material, we explicitly show that factorisation is an accurate approximation.

We have also assumed that selection is uncorrelated with the generation probabilities of the TCR nucleotide sequences encoded in IGoR/OLGA. The assumption is that the recombination statistics are uncorrelated with the statistics of thymic selection, a process that is based on TCR amino acid sequences. However, we note that it has been suggested that selection pressure may induce a correlation between TCRs generated and selected (52). The corresponding statistics of the frequencies of selected TCRs would be modified from those of the generated TCRs shown in Figure 5. Nonetheless, we assume that the resulting distribution can still be approximated by a single-$\alpha$ model which will not qualitatively alter our conclusions.

Mean-Field Approximation

Our mean-field approximation for the mean clone count $c_k$ is embodied in Eq. 7, where correlations between fluctuations in the total population $N = \sum_k k c_k$ 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 $aQ^2 > \bar{\mu}(N^*)$ (39). The mean-field results overestimate the clone counts for $k \lesssim N^*$. Moreover, when the total steady-state T cell immigration rate is extremely small, the effects of competitive exclusion dominate and a single large clone arises (39, 53, 54). Nonetheless, an accurate approximation for the steady-state clone abundance $c_k$ can be obtained using a variation of the two-species Moran model as shown in (39). For the naive T cell system, because $Q$ is so large, the mean immigration rate $\bar{\alpha}$ is such that competitive exclusion is not a dominant feature. Moreover, since $N^* > 10^{10}$, clones counts at comparable sizes are not observed and predicted to be negligible in all models. Since the values of $f_k^*(data)$ become exponentially smaller for large $k$, our inference is most sensitive to the values of $f_k^*(data)$ for small to modest $k$. The information in the data is primarily manifested by how the $f_k^*(data)$ decays in $k$, we before the mean-field approximation deviates from the exact solution. Thus, the parameters associated with the human adaptive immune system satisfy the conditions for the mean-field approximation to be accurate, justifying its use in the BDI model.

Steady State Assumption

In this study, we only considered the steady state of our birth-death-immigration model in Eq. 8 because this limit allowed relatively easy derivations of analytical results. This was also the strategy for previous modeling work (4, 6, 7, 38, 39). However, the per-clone immigration and proliferation times may be on the order of months or years, a time scale over which thymic output diminishes as an individual ages (29, 55–57). Indeed, clone abundance distributions have been shown to show specific patterns as a function of age (58–60). Although $N(t)$, with fixed $\bar{\alpha}$ and $r$, relaxes to steady-state quickly, on a timescale of months, the different subpopulations of specific sizes described by their number $c_k$ relax to quasi-steady-state across a spectrum of time scales depending on the clone sizes $k$ (39, 61). The timescales of relaxation of the largest clones can be estimated from the eigenvalues of the linearized system (Eqs. 7) and are found to be $\sim 10$ years. Thymic involution could be modeled by using a time-dependent $\alpha(t)$ that slowly decreases with age (57). Although T cells are thought to be primarily maintained through proliferation, thymic regeneration has also been shown to affect the naive T cell pool many years after thymectomy in infants. Here, a time dependent increase in $\alpha(t)$ after early thymectomy could be used. Indeed, the clone counts may be determined in early life (17) suggesting the dynamics of certain clones may be very slow, precluding a strict steady-state analysis for the entire repertoire.

In addition to time-dependent changes in $\alpha$, more subtle time-inhomogeneities such as changes in proliferation and death rates have been demonstrated (55, 56, 62). Thus, our steady-state assumption could be relaxed by incorporation of time-dependent perturbations to the model parameters $\mu^*$ 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. The timescales required to reach steady state fall between $1/(\bar{\alpha}Q)$ and $1/\bar{\alpha}$. Thus, it is possible that some components of $c_k$ does not reach steady state in an organism’s lifetime and our steady state model might not be be valid for all values of $c_k$ (57, 61) and a dynamic approach must be taken.

Clustered Immigration

Our mean field model assumed that each immigration event introduced a single naive T cell in the immune system. However, T cells can divide before leaving the thymus and reach a homeostatic state in the periphery. This process can be described by the simultaneous immigration of more than one naive T cell with the same TCR. Clustered immigration of $q$ cells can be implemented in the core model for $c_k$ (Eq. 7) via an immigration term of the form $\alpha_q(c_k, q)$, where $c_k, q = 0$ for $k-q < 0$ (see Section 5 of the Supplementary Material). For $q > 1$, an informative analytic expression for $c_k$ is not available. In Figure S2 of the Section 5 of the Supplementary Material, we show the predicted clone abundance $c_k$ for a neutral model in which $q = 5$. When compared to the case where there is only one cell per immigration, the clone abundance $c_k$ will have a larger slope for $k \lesssim q$, making it kink more downward near $k = q$. Thus, from Figures S2 and 9A, we can see that paired immigration ($q = 2$) would increase $f_k^*$ for $k = 2$, providing an improved fitting to data over single copy immigration ($q = 1$).

Thus, in addition to appreciable sensitivity of the predicted clone counts to $\pi_q$, we also expect clustered immigration defined through the immigration rates $\alpha_q$ $q > 1$ to control the goodness of fit to data. Indeed, Figure S2 suggests that the distribution of immigration cluster sizes $q$. In addition to the proliferation rate heterogeneity $w$, is an important determinant of measured clone counts and that $\alpha_q$ may be constrained by data. We leave this for future investigation.

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 \( (N^r \sim 10^{10} - 10^{11}) \), without the need for simulation. The properties of the BDI model and the overall shape of the sampled clone count data renders the first few \( k \)-values of \( c_k \) or \( f_k \) the most important for determining the constraints among the model parameters. In other words, only the initial rate of the decrease in \( f_k \) (data) for small \( k \) governs the quality of fitting to the model, and one should not expect to be able to explicitly infer more than one or two free parameters.

Our heterogeneous BDI model produced mean sampled clone count distributions that we could directly compare with measured clone counts. The unsampled clone counts \( c_k \) of the neutral model (homogeneous \( \alpha \) and \( r \)) follow a negative binomial distribution which is further modified upon sampling and distribution over the heterogeneous immigration and proliferation rates. Although we determined \( \pi_{k}(\alpha|\bar{\alpha}) \) through a code that implemented recombination statistics inferred from cDNA and gDNA sequences (20, 28), we found that the behavior of the model is rather insensitive to distributions \( \pi_{k}(\alpha|\bar{\alpha}) \) with mean values \( \bar{\alpha} \) much smaller than the largest proliferation rates \( r \). The model results are dominated by many low immigration-rate clones and a model that replaces \( \alpha \) with its mean value \( \bar{\alpha} \) is sufficient.

Conversely, we find that the shape of the clone count profiles \( c_k \) are quite sensitive to the proliferation rate heterogeneity \( w \). A small amount of heterogeneity quickly reduces the best-fit values of \( \lambda \) to reasonable values. For estimated values \( \eta \sim 10^{-6} - 10^{-4} \), \( \bar{\alpha} \sim 10^{-4} \), and small values of \( \lambda = N^r/Q \leq 10^{-3} \), requires a best-fit width \( w = 1 \). Heterogeneity is needed to generate clones of sufficiently large size that persist after sampling. 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 shape of expected clone abundance is sensitive to the behavior of the proliferation rate distribution near the maximum dimensional proliferation rate \( R, \pi_{k}(r = R) \). The predicted clone counts are also modestly sensitive to the distribution of immigration cluster sizes \( q \) (representing transient proliferation just before thymic output). When \( q > 1 \) cells of a clone are simultaneously exported by the thymus, the predicted mean clone counts decay much more slowly for small \( k \leq q \) (see Figure S2). This modification will allow for better fitting since clustered immigration increases the predicted clone counts for larger \( k, c_k, c_j \), etc., and eventually \( f_2 \), \( f_j \), etc. Thus, we expect that a model containing multiple clustered immigration rates \( \alpha_{q \geq 1} \) will lower the error and provide better fitting, particularly at larger \( w \). Additional analysis using a distribution of immigration cluster sizes may allow this type of clone count data to reveal more information about the physiological mechanism of naive T cell maintenance.

Even assuming modest heterogeneity, our work leads to the conclusion that the typical immigration heterogeneity is not enough to influence measured clone counts and that varying levels of proliferation heterogeneity is needed to shape \( c_k \) and \( f_k \) (12). These results are consistent with the finding that naive T cells in humans are maintained by proliferation rather than thymic output (9). Since we have only investigated the effects of a uniform distribution for \( \pi_{k}(r|\bar{\alpha}) \), further studies using more complex shapes of \( \pi_{k}(\alpha, r|\bar{\alpha}) \) can be easily explored numerically 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. Finally, it will be important to extend our steady-state model to allow \( \alpha(t), \pi_{k}(r|t), \) and \( \mu^*(t) \) to be functions of time in order to predict clone abundances in the presence of thymic involution and reduced proliferation with age (62, 63), which can even arise differentially in different compartments (64).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.frontiersin.org/articles/10.3389/fimmu.2017.01267/full.

AUTHOR CONTRIBUTIONS

RD, TC, and MRD developed and analyzed the model and wrote the manuscript. YP organized published data, and DM assisted in sorting and organizing generated data. TC, YP, and MX performed numerical analyses and data fitting. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by grants from the NIH through grant R01HL146552 (TC), the Army Research Office through grant W911NF-18-1-0345 (MRD), the NSF through grants DMS-1814364 (TC) and DMS-1814090 (MRD). The authors also thank the Collaboratory in Institute for Quantitative and Computational Biosciences at UCLA for support to RD.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2021.735135/full#supplementary-material

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. Proc Natl Acad Sci (USA) (2014) 111:13139–44. doi: 10.1073/pnas.1409155111

2. van den Broek T, Borghans JAM, van Wijk F. The Full Spectrum of Human Naive T Cells. Nat Rev Immunol (2018) 18:363–73. doi: 10.1038/s41577-018-0001-y

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) 370:20140291. doi: 10.1098/rstb.2014.0291
4. Desponts J, Mora T, Walczak AM. Fluctuating Fitness Shapes the Clone-Size Distribution of Immune Repertoires. Proc Natl Acad Sci USA (2016) 113:274–9. doi: 10.1073/pnas.1512977112
5. Desponts J, Mayer A, Mora T, Walczak AM. Population Dynamics of Immune Repertoires. In: Molina-París C, Lythe G, editors. Mathematical, Computational and Experimental T Cell Immunology. Cham, Switzerland: Springer (2021). p. 67–79.
6. Lythe G, Callard RE, Hoare RL, Molina-París C. How Many TCR Clonotypes Does a Body Maintain? J Theor Biol (2016) 389:214–24. doi: 10.1016/j.jtbi.2015.10.016
7. de Greef PC, Oakes T, Gerritsen B, Ismail M, Heather JM, Hermsen R, et al. Maintenance of Peripheral Naïve T Cells Is Sustained by Thymus Output in Mice But Not Humans. Immunity (2012) 36:288–97. doi: 10.1016/j.immuni.2012.02.006
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 Comput Biol (2018) 14:e1006571. doi: 10.1371/journal.pcbi.1006571
9. Oakes T, Heather JM, Best K, Byng-Maddick R, Husovsky C, Ismail M, et al. Maximum Entropy Models for Complete and Accurate Recovery of T Cell Repertoires From High Throughput Sequencing Data. Bioinformatics (2016) 32:3098–106. doi: 10.1093/bioinformatics/btw124
10. Lythe G, Molina-París C. T Cell Clone Size Distributions. Frontiers in Immunology (2014) 4:243. doi: 10.3389/fimmu.2014.00434
11. Marcou Q, Mora T, Walczak AM. High-Throughput Immune Repertoire Sequencing Data. Bioinformatics (2017) 33:3530–9. doi: 10.1093/bioinformatics/btx217
12. Reyes LF, Weigmann A, Katre KG, de Souza MM, Silva SL, Sousa AE. Establishment and Maintenance of the Human Naïve CD4+ T-Cell Compartment. Front Pediatr (2016) 4:119. doi: 10.3389/fped.2016.00119
13. Selin LK, Gorski J, Callard RE, Lythe G. Neutral Models for the T Cell Receptor repertoire. Front Pediatr (2018) 6:129. doi: 10.3389/fped.2018.00129
14. Antipa T, Gheyas U, Ofner KJ, Rajan V, Kalra A, Teschler C, et al. A Filtered Nearest-Neighbor Approach to Comparative Analysis of the Clone Size Distribution of the T Cell Receptor Repertoire. PloS Comput Biol (2018) 14:e1006571. doi: 10.1371/journal.pcbi.1006571
15. Reynolds J, Coles M, Lythe G, Molina-París C. Mathematical Model of Naïve T Cell Division and Survival IL-7 Thresholds. Front Immunol (2013) 4:434. doi: 10.3389/fimmu.2013.00434
16. Gallego-Sandoval CR, OYang O, Jojic N, Lovato P, Chen DY, Boechat MI, et al. Interleukin-7 Mediates the Critical for Homeostatic Proliferation and Survival of Naïve T Cells. Proc Natl Acad Sci USA (2001) 98:7832–7. doi: 10.1073/pnas.161126098
17. Schluns KS, Kieper WC, Jameson SC, Franchini L, Interleukin-7 Mediates the Homeostasis of Naïve and Memory CD8 T Cells In Vivo. Nat Immunol (2000) 1:426–32. doi: 10.1038/35088868
18. Cupe SM, Devlin BH, Markert ML, Kepler TB. The Dynamics of T-Cell Receptor Repertoire Diversity Following Thymus Transplantation for DiGeorge Anomaly. PloS Comput Biol (2009) 5:e1000396. doi: 10.1371/journal.pcbi.1000396
19. Farber DL, Yudanin NA, Restifo NP. Human Memory T Cells: Generation, Compartmentalization and Homeostasis. Nat Rev Immunol (2014) 14:24–35. doi: 10.1038/nri3567
20. Senthia Z, Elhanati Y, Callan CG, Walczak AM, Mora T, OLGA: Fast Computation of Generation Probabilities of B- and T-Cell Receptor Amino Acid Sequences and Motifs. Bioinformatics (2019) 35:2974–81. doi: 10.1093/bioinformatics/btz335
21. Westera L, van Hoeven V, Deweylic J, Spiereburg G, van Velzen JJ, de Boer RJ, et al. Lymphocyte Maintenance During Healthy Aging Requires No Substantial Alterations in Cellular Turnover. Aging Cell (2014) 19:229–37. doi: 10.1111/acel.12311
46. Davis MM, Bjorkman PJ. T-Cell Antigen Receptor Genes and T-Cell Recognition. *Nature* (1988) 334:395. doi: 10.1038/334395a0

47. Clark DR, de Boer RJ, Wolthers KC, Miedema F. T Cell Dynamics in HIV-1 Infection. *Adv Immunol* (1999) 73:301–27. doi: 10.1016/S0065-2776(08)60789-0

48. Ye P, Kirschner DE. Reevaluation of T Cell Receptor Excision Circles as a Measure of Human Recent Thymic Emigrants. *J Immunol* (2002) 168:4968–79. doi: 10.4049/jimmunol.168.10.4968

49. Hazenben MD, Borghans JAM, de Boer RJ, Miedema F. Thymic Output: A Bad TREC Record. *Nat Immunol* (2003) 4:497–9. doi: 10.1038/nit0203-97

50. Bains I, Thiebaut R, Yates AJ, Callard R. Quantifying Thymic Export: Combining Models of Naive T Cell Proliferation and TCR Excision Circle Dynamics Gives an Explicit Measure of Thymic Output. *J Immunol* (2009) 183:4329–36. doi: 10.4049/jimmunol.0900743

51. Cunningham CA, Helm EY, Fink PJ. Reinterpreting Recent Thymic Emigrant Function: Defective or Adaptive? *Curr Opin Immunol* (2018) 51:1–6. doi: 10.1016/j.coi.2017.12.006

52. Elhanati Y, Murugan A, Callan CG Jr, Mora T, Walczak AM. Quantifying Selection in Immune Receptor Repertoires. *Proc Natl Acad Sci USA* (2014) 111:9875–80. doi: 10.1073/pnas.1409572111

53. Hardin G. The Competitive Exclusion Principle. *Science* (1960) 131:1292–7. doi: 10.1126/science.131.3409.1292

54. Hutchinson GE. The Paradox of the Plankton. *Am Nat* (1961) 95:137–45. doi: 10.1086/282171

55. 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. *Proc Natl Acad Sci USA* (2015) 112:E6917–26. doi: 10.1073/pnas.1517261112

56. 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 Comput Biol* (2018) 16:e2003949. doi: 10.1371/journal.pcbi.2003949

57. Lewkiewicz S, Chuang YL, Chou T. A Mathematical Model of the Effects of Aging on Naive T Cell Populations and Diversity. *Bull Math Biol* (2019) 81:2783–817. doi: 10.1007/s11538-019-00630-z

58. 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. *Proc Natl Acad Sci USA* (2012) 109:21432–7. doi: 10.1073/pnas.1209283110

59. 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. *J Immunol* (2016) 196:5005–13. doi: 10.4049/jimmunol.1600005

60. Egorov ES, Kasatksaya SA, Zubov VN, Izraelson M, Nakonechnaya TO, Staroverov DB, et al. The Changing Landscape of Naive T Cell Receptor Repertoire With Human Aging. *Front Immunol* (2018) 9:1618. doi: 10.3389/fimmu.2018.01618

61. Lewkiewicz SM, Chuang YL, Chou T. Dynamics of T Cell Receptor Distributions Following Acute Thymic Atrophy and Resumption. *Math Biosci Eng* (2020) 17:28–55. doi: 10.3934/mbe.2020002

62. Mold JE, Reu P, Olin A, Bernard S, Michaelsson J, Rane S, et al. Cell Generation Dynamics Underlying Naive T-Cell Homeostasis in Adult Humans. *PloS Biol* (2019) 17:e3000383. doi: 10.1371/journal.pbio.3000383

63. van den Broek T, Delemarre EM, Janssen WJ, Nievelstein RA, Broen JC, Tesselaar K, et al. Neonatal Thymectomy Reveals Differentiation and Plasticity Within Human Naive T Cells. *J Clin Invest* (2016) 126:1126–36. doi: 10.1172/JCI84997

64. Thome JJC, Grinshpun B, Kumar BV, Kubota M, Ohmura Y, Lerner H, et al. Long-Term Maintenance of Human Naive T Cells Through In Situ Homeostasis in Lymphoid Tissue Sites. *Sci Immunol* (2016) 1:eaa6506. doi: 10.1126/sciimmunol.aah6506

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher’s Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Dessalles, Pan, Xia, Maestrini, D’Orosogna and Chou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.