Title

Mathematical Characterization of Private and Public Immune Receptor Sequences.

Journal

Bulletin of Mathematical Biology, 85(10)

Authors

Böttcher, Lucas
Wald, Sascha
Chou, Tom

Publication Date

2023-09-14

DOI

10.1007/s11538-023-01190-z

Peer reviewed

Bulletin of Mathematical Biology (2023) 85:102
https://doi.org/10.1007/s11538-023-01190-z

Society for Mathematical Biology

ORIGINAL ARTICLE

Mathematical Characterization of Private and Public Immune Receptor Sequences

Lucas Böttcher[1,2,3] · **Sascha Wald**[4] · **Tom Chou**[2,5]

Received: 11 April 2023 / Accepted: 26 July 2023 / Published online: 14 September 2023

Abstract

Diverse T and B cell repertoires play an important role in mounting effective immune responses against a wide range of pathogens and malignant cells. The number of unique T and B cell clones is characterized by T and B cell receptors (TCRs and BCRs), respectively. Although receptor sequences are generated probabilistically by recombination processes, clinical studies found a high degree of sharing of TCRs and BCRs among different individuals. In this work, we use a general probabilistic model for T/B cell receptor clone abundances to define "publicness" or "privateness" and information-theoretic measures for comparing the frequency of sampled sequences observed across different individuals. We derive mathematical formulae to quantify the mean *and the variances* of clone richness and overlap. Our results can be used to evaluate the effect of different sampling protocols on abundances of clones within an individual as well as the commonality of clones across individuals. Using synthetic and empirical TCR amino acid sequence data, we perform simulations to study expected clonal commonalities across multiple individuals. Based on our formulae, we compare these simulated results with the analytically predicted mean and variances of the

✉ Lucas Böttcher
l.boettcher@fs.de

Sascha Wald
sascha.wald@coventry.ac.uk

Tom Chou
tomchou@ucla.edu

1 Department of Computational Science and Philosophy, Frankfurt School of Finance and Management, 60322 Frankfurt am Main, Germany

2 Department of Computational Medicine, University of California, Los Angeles, 621 Charles E. Young Dr. S., Los Angeles 90095-1766, CA, USA

3 Department of Medicine, University of Florida, Gainesville 32610, FL, USA

4 Statistical Physics Group, Centre for Fluid and Complex Systems, Coventry University, Priory Street, Coventry CV1 5FB, UK

5 Department of Mathematics, University of California, Los Angeles, 520 Portola Plaza, Los Angeles 90095-1555, CA, USA

repertoire overlap. Complementing the results on simulated repertoires, we derive explicit expressions for the richness and its uncertainty for specific, single-parameter truncated power-law probability distributions. Finally, the information loss associated with grouping together certain receptor sequences, as is done in spectratyping, is also evaluated. Our approach can be, in principle, applied under more general and mechanistically realistic clone generation models.

Keywords T cell repertoire · Diversity · Public/private clones · Overlap · Sampling

1 Introduction

A major component of the adaptive immune system in most jawed vertebrates is the repertoire of B and T lymphocytes. A diverse immune repertoire allows the adaptive immune system to recognize a wide range of pathogens (Xu et al. 2020). B and T cells develop from common lymphoid progenitors (CLPs) that originate from hematopoietic stem cells (HSCs) in the bone marrow. B cells mature in the bone marrow and spleen while developing T cells migrate to the thymus where they undergo their maturation process. After encountering an antigen, naive B cells may get activated and differentiate into antibody-producing plasma cells, which are essential for humoral (or antibody-mediated) immunity. In recognizing and eliminating infected and malignant cells, T cells contribute to cell-mediated immunity of adaptive immune response.

T-cell receptors bind to antigenic peptides (or epitopes) that are presented by major histocompatibility complex (MHC) molecules on the surface of antigen-presenting cells (APCs). T cells that each carry a type of TCR mature in the thymus and undergo V(D)J recombination, where variable (V), diversity (D), and joining (J) gene segments are randomly recombined (Alt et al. 1992; Travers et al. 1997). The receptors are heterodimeric molecules and mainly consist of an α and a β chain while only a minority, about 1–10% (Girardi 2006), of TCRs consists of a δ and a γ chain. The TCR α and γ chains are made up of VJ and constant (C) regions. Additional D regions are present in β and γ chains. During the recombination process, V(D)J segments of each chain are randomly recombined with additional insertions and deletions. After recombination, only about 5% or even less (Yates 2014) of all generated TCR sequences are selected based on their ability to bind to certain MHC molecules ("positive selection") and to not trigger autoimmune responses ("negative selection"). These naive T cells are then exported from the thymus into peripheral tissue where they may interact with foreign peptides that are presented by APCs. The selection process as well as subsequent interactions are specific to an individual.

The most variable parts of TCR sequences are the complementary determining regions (CDRs) 1, 2, and 3, located within the V region, among which the CDR3β is the most diverse (Abbas et al. 2021). Therefore, the number of distinct receptor sequences, the richness R, of TCR repertoires is typically characterized in terms of the richness of CDR3β sequences. Only about 1% of T cells express two different TCRβ chains (Davodeau et al. 1995; Padovan et al. 1995; Schuldt and Binstadt 2019), whereas the proportion of T cells that express two different TCRα chains may be as high as 30% (Rybakin et al. 2014; Schuldt and Binstadt 2019).

B cells can also respond to different antigens via different B cell receptors (BCRs) that are comprised of heavy and light chains. As with TCRs, the mechanism underlying the generation of a diverse pool of BCRs is VDJ recombination in heavy chains and VJ recombination in light chains. Positive and negative selection processes sort out about 90% of all BCRs that react too weakly or strongly with certain molecules (Tussiwand et al. 2009). As a result of the various recombination and joining processes and gene insertions and deletions, the practical theoretical maximum repertoire size Ω_0 of the variable region of BCR and TCR receptors can be $\sim 10^{14}-10^{20}$ (Davis and Bjorkman 1988; Venturi et al. 2008; Zarnitsyna et al. 2013; Lythe et al. 2016). This value is comparable to the possible number of amino acid sequences of typical length $\sim 11-12$. However, many of these sequences are not viable, are removed through thymic selection, or are have such low probability occuring that they are never expected to be produced in an organism's lifetime. Thus, the effective number of TCR variable regions that are produced and that can contribute to the organism's repertoire size, Ω, should be much less than Ω_0. Estimating the true size of BCR and TCR repertoires realized in an organism is challenging since the majority of such analyses are based on small blood samples, leading to problems similar to the "unseen species" problem in ecology (Laydon et al. 2015). Nonetheless, the number of unique TCRs realized in organisms has been estimated to be about 10^6 for mice (Casrouge et al. 2000) and about 10^8 for humans (Soto et al. 2020). B-cell repertoire size for humans is estimated to be 10^8-10^9 (DeWitt et al. 2016). These values are significantly smaller than Ω_0 and might be used as an effective Ω.

Each pool of BCR and TCR sequences realized in one organism i can be seen as a subset $\mathcal{U}_i$ of the set of all possible species-specific sequences $\mathcal{S}$. Sequences that occur in at least two different organisms i and j (i.e., sequences that are elements of $\mathcal{U}_i \cap \mathcal{U}_j$) are commonly referred to as "public" sequences (Laydon et al. 2015) while "private" sequences occur only in one of the individuals tested. The existence of public TCRβ sequences has been established in several previous works (Putintseva et al. 2013; Robins et al. 2010; Shugay et al. 2013; Soto et al. 2020). More recently, a high degree of shared sequences has been also observed in human BCR repertoires (Briney et al. 2019; Soto et al. 2019).

The notions of public and private clonotypes have been loosely defined. Some references use the term "public sequence" to refer to those sequences that "are *often* shared between individuals" (Shugay et al. 2013) or "shared across individuals" (Greiff et al. 2017). Recently, Elhanati et al. (2018) and Ruiz Ortega et al. (2023) have formulated a mathematical and statistical framework to quantify "publicness" and "privateness." Building on these works, we derive a set of measures that enable us to quantify immune repertoire properties, including the expected total richness, the expected numbers of public and private clones, *and their variances* (confidence levels), all expressed in terms of the general set of clone generation probabilities or clone populations. One of our metrics is the expected "M-overlap" or "M-publicness," defined as the expected number of clones that appear in samples drawn from M separate individuals. This quantity is a clinically interpretable limit of the expected "sharing number" defined in Elhanati et al. (2018). Similarly, we define M-private clones as clones that are not shared by all M individuals, i.e., occurring in at most $M-1$ individuals.

In the next section, we first give an overview of the mathematical concepts that are relevant to characterize TCR and BCR distributions. We then formulate a statistical model of receptor distributions in Sect. 3. In Sects. 3.1 and 3.2, we derive quantities associated with receptor distributions in single organisms and across individuals, respectively. We will primarily focus on the overlap of repertoires across individuals and on the corresponding confidence intervals that can be used to characterize "public" and "private" sequences of immune repertoires. Formulae we derived are listed in Table 1. In Sect. 4, we use synthetic and empirical TCR amino acid sequence data and perform simulations to compare theoretical predictions of repertoire overlaps between different individuals with corresponding observations. Finally, when analyzing empirical sequence data, one may use continuous approximations (Elhanati et al. 2018; Ruiz Ortega et al. 2023) and averaging (i.e., coarse-graining) methods that change the information content in the underlying dataset. Coarse-graining of TCR and BCR data may also be a result of the employed sequencing techniques (Gorski et al. 1994; Fozza et al. 2017). In Sect. 6, we therefore briefly discuss the information loss associated with analyzing processed cell data. We discuss our results and conclude our paper in Sect. 7. Our source codes are publicly available at GitLab (2022).

2 Mathematical Concepts

Although receptor sequences and cell counts are discrete quantities, using continuous functions to describe their distribution may facilitate the mathematical analysis of the quantities that we derive in the subsequent sections. For example, a continuous approximation (i.e., a "density-of-states approximation") has been used to characterize the number of T cell receptor sequences possible within a continuous range of generation probabilities (Murugan et al. 2012; Elhanati et al. 2018; Ruiz Ortega et al. 2023). Another instance of a continuous cell statistics approximation involves employing power laws to describe the rank-abundance curves associated with immune repertoires (see, e.g., Gaimann et al. 2020). We therefore briefly review some elementary concepts associated with continuous distributions and their discretization.

Let $p(x)$ be the *probability density* associated with the distribution of traits, as depicted in Fig. 1a. The probability that a certain trait occurs in $[x, x+\mathrm{d}x)$ is $p(x)\,\mathrm{d}x$. The corresponding discretized distribution elements are

$$p_i := \int_{i\Delta}^{(i+1)\Delta} p(x)\,\mathrm{d}x, \tag{1}$$

where Δ is the discretization step size of the support of $p(x)$. If we discretize the values of probabilities, the number of clones with a certain relative frequency p_i is given by the *clone count*

$$c_k := \sum_i \mathbb{1}\left(k\delta \le p_i < (k+1)\delta\right), \tag{2}$$

where the indicator function $\mathbb{1} = 1$ if its argument is satisfied and 0 otherwise. As shown in Fig. 1b, the parameter δ defines an interval of frequency values and modulates the clone-count binning. Figures 1b, c show how p_i and c_k are constructed from a continuous distribution $p(x)$. If $p(x)$ is not available from data or a model, an alternative

Table 1 Table of mathematical results

Measure	Description	**n**-representation	**p**-representation
$\mathbb{E}[R]$	Individual richness	$\sum_{k=1}^{N}\sum_{i=1}^{\Omega}\mathbb{1}(n_i,k)$	$\sum_{i=1}^{\Omega}\rho_i$
$\mathbb{E}[R^2]$	2nd moment of richness	$\left[\sum_{k=1}^{N}\sum_{i=1}^{\Omega}\mathbb{1}(n_i,k)\right]^2$	$\sum_{i=1}^{\Omega}\rho_i+\sum_{j\neq i}^{\Omega}\rho_{ij}$
$\mathbb{E}[R_{\mathrm{s}}]$	Mean sampled richness	$\sum_{i=1}^{\Omega}\sigma_i=\Omega-\frac{1}{\binom{N}{S}}\sum_{i=1}^{\Omega}\binom{N-n_i}{S}$	$\sum_{i=1}^{\Omega}\rho_i(S)$
$\mathbb{E}[R_{\mathrm{s}}^2]$	2nd moment, sampled richness	$\sum_{i=1}^{\Omega}\sigma_i+\sum_{j\neq i}^{\Omega}\sigma_{ij}$	$\sum_{i=1}^{\Omega}\rho_i(S)+\sum_{j\neq i}^{\Omega}\rho_{ij}(S)$
$\mathbb{E}[R^{(M)}]$	Mean group richness	$\sum_{k\geq 1}\sum_{i=1}^{\Omega}\mathbb{1}\left(\sum_{m=1}^{M}n_i^{(m)},k\right)$	$\sum_{i=1}^{\Omega}\tilde{\rho}_i$
$\mathbb{E}\left[(R^{(M)})^2\right]$	2nd mom., grp richness	$\left[\sum_{k\geq 1}\sum_{i=1}^{\Omega}\mathbb{1}\left(\sum_{m=1}^{M}n_i^{(m)},k\right)\right]^2$	$\sum_{i}^{\Omega}\tilde{\rho}_i+\sum_{i\neq j}^{\Omega}\tilde{\rho}_{ij}$
$\mathrm{var}[R^{(M)}]$	Variance of grp richness	0	$\sum_{i\neq j}^{\Omega}\tilde{\rho}_{ij}+(1-\sum_{i=1}^{\Omega}\tilde{\rho}_i)\sum_{i=1}^{\Omega}\tilde{\rho}_i$
$\mathbb{E}[R_{\mathrm{s}}^{(M)}]$	Mean sampled grp richness	$\sum_{i=1}^{\Omega}\tilde{\sigma}_i$	$\sum_{i=1}^{\Omega}\tilde{\rho}_i(S)$
$\mathbb{E}\left[(R_{\mathrm{s}}^{(M)})^2\right]$	2nd mom., sampled grp richness	$\sum_{i=1}^{\Omega}\tilde{\sigma}_i+\sum_{i\neq j}^{\Omega}\tilde{\sigma}_{ij}$	$\sum_{i=1}^{\Omega}\tilde{\rho}_i(S)+\sum_{i\neq j}^{\Omega}\tilde{\rho}_{ij}(S)$
$\mathbb{E}[K^{(M)}]$	Expected M-overlap	$\sum_{i=1}^{\Omega}\prod_{m=1}^{M}\sum_{k^{(m)}\geq 1}\mathbb{1}(n_i^{(m)},k^{(m)})$	$\sum_{i=1}^{\Omega}\prod_{m=1}^{M}\rho_i^{(m)}$
$\mathbb{E}\left[(K^{(M)})^2\right]$	2nd moment, M-overlap	$\left[\sum_{i=1}^{\Omega}\prod_{m=1}^{M}\sum_{k^{(m)}\geq 1}\mathbb{1}(n_i^{(m)},k^{(m)})\right]^2$	$\sum_{i=1}^{\Omega}\prod_{m=1}^{M}\rho_i^{(m)}+\sum_{i\neq j}^{\Omega}\prod_{m=1}^{M}\rho_{ij}^{(m)}$
$\mathbb{E}[K_{\mathrm{s}}^{(M)}]$	Sampled M-overlap	$\sum_{i=1}^{\Omega}\prod_{m=1}^{M}\sigma_i^{(m)}$	$\sum_{i=1}^{\Omega}\prod_{m=1}^{M}\rho_i^{(m)}(S)$
$\mathbb{E}[(K_{\mathrm{s}}^{(M)})^2]$	2nd mom., sampled M-overlap	$\sum_{i=1}^{\Omega}\prod_{m=1}^{M}\sigma_i^{(m)}(S)+\sum_{i\neq j}^{\Omega}\prod_{m=1}^{M}\sigma_{ij}^{(m)}(S)$	$\sum_{i=1}^{\Omega}\prod_{m=1}^{M}\rho_i^{(m)}(S)+\sum_{i\neq j}^{\Omega}\prod_{m=1}^{M}\rho_{ij}^{(m)}(S)$

We list our main mathematical derivations and expressions for richness and overlap, unsampled and sampled, in both the **n**-representation and the **p**-representation. The component probabilities $\rho_i^{(m)}$, $\rho_{ij}^{(m)}$, $\tilde{\rho}_i$, $\tilde{\rho}_{ij}$, $\sigma_i^{(m)}$, $\sigma_{ij}^{(m)}$, $\tilde{\sigma}_i$, and $\tilde{\sigma}_{ij}$ are given in Eqs. (6), (8), (13), (16), (24), (25), (26), and (27), respectively. The component probabilities $\rho_i^{(m)}(S)$, $\rho_{ij}^{(m)}(S)$, $\tilde{\rho}_i(S)$, $\tilde{\rho}_{ij}(S)$ associated with sampled quantities in the **p**-representation are given by Eqs. (37), (38), (39), and (40), respectively

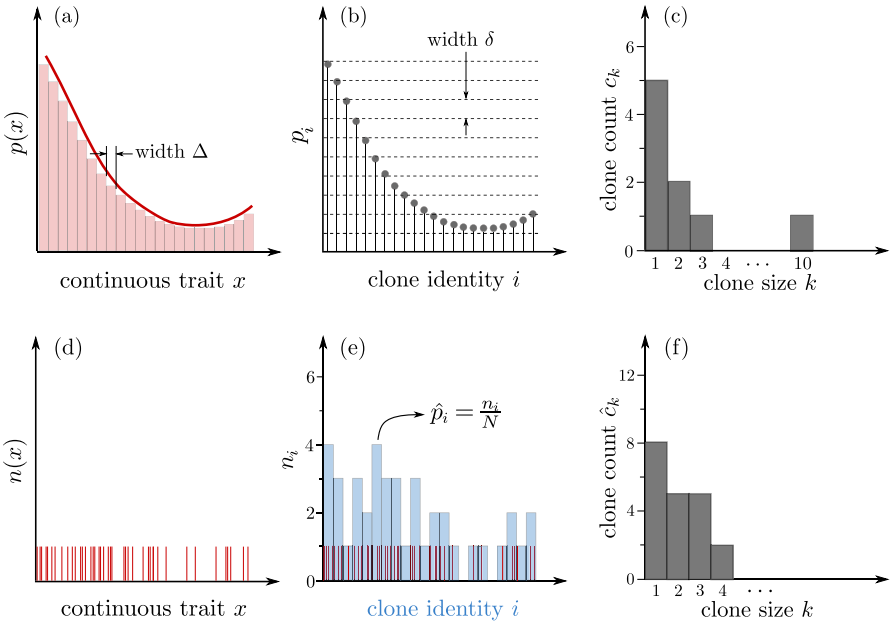

Fig. 1 Sampling from a continuous distribution, described in terms of an underlying probability density $p(x)$ and number density $n(x)$. The probability density $p(x)$ (solid red line) and the Riemann sum approximation to the probability (red bars of width Δ) are shown in panel (**a**). The probability that a trait in the interval $[x, x + \mathrm{d}x)$ arises is $p(x)\mathrm{d}x$. As shown in (**b**), this distribution can be discretized directly by the intervals $[i\Delta, (i + 1)\Delta)$ (red bars) defining discrete traits and their associated probabilities p_i (see Eq. (1)). The probabilities p_i can be transformed into clone counts c_k (the number of identities i that are represented by k individuals) using Eq. (2), which are shown in (**c**). A finite sample of a population described by $p(x)$ yields the binary outcome shown in (**d**). In this example, the total number of samples is $N = 41$ and since the trait space x is continuous, the probability that the exact same trait arises in more than one sample is almost surely zero. Light blue bars in panel (**e**) represent number counts n_i binned according to Δ. The probabilities $\hat{p}_i = n_i/N$ provide an approximation of p_i. Clone counts for the empirical $\hat{p}_i$ are calculated according to Eq. (3) and shown in (**f**) (Color figure online)

representative starts with the number density $n(x)$, which can be estimated by sampling a process which follows $p(x)$. The probability that a continuous trait x is drawn twice from a continuous distribution $p(x)$ is almost surely zero. Hence, the corresponding number counts $n(x)$ are either 1 if $X \in [x, x + \mathrm{d}x)$ (i.e., if trait X is sampled) or 0 otherwise, as shown in Fig. 1d, e. We say that X is of *clonotype* i if $X \in [i\Delta, (i+1)\Delta)$ $(1 \leq i \leq \Omega)$ and we use n_i to denote the number of cells of clonotype i. Then, if Ω denotes the effective number of different clonotypes, the total T-cell (or B-cell) population is $N \equiv \sum_{i=1}^{\Omega} n_i$. The relative empirical abundance of clonotype i is thus $\hat{p}_i = n_i/N$ (see Fig. 1e), satisfying the normalization condition $\sum_i \hat{p}_i = 1$. Besides the simple discrete estimate $\hat{p}_i = n_i/N$, one can also reconstruct $p(x)$ from $\mathbf{n} = \{n_i\}$ using methods such as kernel density estimation. The corresponding empirical clone count derived from the number representation n_i is defined as

$$\hat{c}_k := \sum_{i=1}^{\Omega} \mathbb{1}(n_i, k) \tag{3}$$

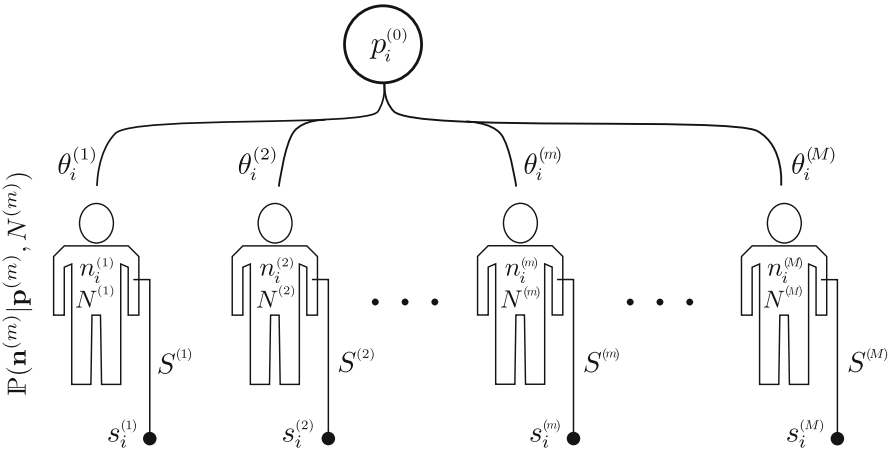

Fig. 2 Schematic of sampling of multiple species from multiple individuals. A central process produces (through V(D)J recombination) TCRs. Individuals select for certain TCRs resulting in population $n_i^{(m)}$ of T cells with receptor i in individual m, for a total T-cell count $N^{(m)} = \sum_i^{\Omega} n_i^{(m)}$. The selection of TCR i by individual m (in their individual thymuses) is defined by the parameter $\theta_i^{(m)}$ which gives an effective probability $p_i^{(m)} \equiv \theta_i^{(m)} p_i^{(0)}$. A sample with cell numbers $S^{(m)} \ll N^{(m)}$ is drawn from individual m and sequenced to determine $s_i^{(m)}$, the number of cells of type i in the subsample drawn from individual m

and shown in Fig. 1f. The indicator function $\mathbb{1}(a, b)$ with arguments $a, b \in \mathbb{Z}_{\geq 0}$ is equal to 1 if $a = b$ and 0 otherwise. Clone counts can be used to describe T cell repertoires, especially if clone identities are not important. Simple birth-death-immigration models can also be cast in terms of, e.g., expected clone counts $\mathbb{E}[\hat{c}_k(t)]$ (Goyal et al. 2015; Lewkiewicz et al. 2019).

3 Whole Organism Statistical Model

Using the mathematical quantities defined above, we develop a simple statistical model for BCR and TCR sequences distributed among individuals. Although our model is applicable to both BCR and TCR sequences, we will primarily focus on the characterization of TCRs for simplicity. B cells undergo an additional process of somatic hypermutation and class switching leading to a more dynamic evolution of the more diverse B cell repertoire (Elhanati et al. 2015). By focusing on naive T cells, we can assume their populations are generated by the thymus via a single, simple effective process.

Assume a common universal recombination process (see Fig. 2) in T-cell development that generates a cell carrying TCR of type $1 \leq i \leq \Omega_0$ with probability $p_i^{(0)}$. Here, $\Omega_0 \gg \Omega$ is the theoretical number of ways the full TCR sequence can be constructed which is itself much larger than the effective number Ω that appears in an individual after thymic selection. Although each new T cell produced carries TCR i with probability $p_i^{(0)}$, many sequences i are not realized given thymic selection (that eliminates $\sim$ 98% of them), the finite number of T cells produced over a lifetime (Travers et al. 1997; Yates 2014; Lythe et al. 2016), or the extremely low generation

probability of some clones. These effects are invoked to truncate Ω_0 to $\Omega \ll \Omega_0$. However, we will see in Sect. 5, explicit scaling relastionships for the limit $\Omega \gg 1$ can be found for general power-law ordered probabilities p_i.

Besides VDJ recombination (Slabodkin et al. 2021), thymic selection and subsequent death, activation, and proliferation occur differently across individuals $1 \leq m \leq M$ and may be described by model parameters $\theta_i^{(m)}$. Such a model translates the fundamental underlying recombination process into a population of $n_i^{(m)}$ T cells with TCR i and total population $N^{(m)} = \sum_{i=1}^{\Omega} n_i^{(m)}$ in individual m. The connection between $p_i^{(0)}, \theta_i^{(m)}$ and $n_i^{(m)}, N^{(m)}$ might be described by dynamical models, deterministic or stochastic, such as those presented in Dessalles et al. (2022).

At any specific time, individual m will have a cell population configuration $\mathbf{n}^{(m)} \equiv (n_1^{(m)}, n_2^{(m)}, \ldots, n_\Omega^{(m)})$ with probability $\mathbb{P}(\mathbf{n}^{(m)}|\boldsymbol{\theta}^{(m)}, N^{(m)})$. Each individual can be thought of as a biased sample from all cells produced via the universal probabilities $p_i^{(0)}$. We can approximate individual probabilities $p_i^{(m)} \equiv \theta_i^{(m)} p_i^{(0)}$, $1 \leq i \leq \Omega$, where the number of effective TCRs Ω for individual m might have as upper bound $\Omega \sim 10^{14}$, if, for example, we are considering just the CDR3 region of the β chain. Assuming a time-independent model (*e.g*, a model in steady-state), we can describe the probability of a T-cell population $\mathbf{n}^{(m)}$ in individual m by a multinomial distribution over individual probabilities $\mathbf{p}^{(m)} \equiv \{p_i^{(m)}\}$:

$$\mathbb{P}(\{\mathbf{n}^{(m)}\}|\{\mathbf{p}^{(m)}, N^{(m)}\}) = N^{(m)}! \prod_{i=1}^{\Omega} \frac{[p_i^{(m)}]^{n_i^{(m)}}}{n_i^{(m)}!}, \tag{4}$$

where $\sum_{i=1}^{\Omega} n_i^{(m)} \equiv N^{(m)}$ and $\sum_{i=1}^{\Omega} p_i^{(m)} \equiv 1$. Thus, each individual can be thought of as a "sample" of the "universal" thymus. Neglecting genetic relationships amongst individuals, we can assume them to be independent with individual probabilities $p_i^{(m)}$. These are the probabilities that a randomly drawn cell from individual m is a cell of clone i. Repeated draws (with replacement) would provide the samples for the estimator $\hat{p}_i^{(m)} = n_i^{(m)}/N^{(m)}$, assuming $n_i^{(m)}$ are counted and $N^{(m)}$ is also known or estimated. This representation allows us to easily express the probabilities of any configuration $\mathbf{n}^{(m)}$ analytically. A dynamical model for $n_i^{(m)}$ cannot be directly described by our simple probabilities $p_i^{(m)}$. A mechanistically more direct model could incorporate the production rate of clone i T cells from the thymus, the proliferation and apoptosis rates of clone i cells, and interactions manifested as, e.g., carrying capacity as model parameters. Probability distributions for $\mathbf{n}^{(m)}$, as a function of birth, death, and immigration rates, have been found in Dessalles et al. (2018) and can also be used, instead of Eq. (4), to construct probabilities.

3.1 Single Individual Quantities

First, we focus on quantities intrinsic to a single individual organism; thus, we can suppress the "$m = 1$" label. Within an individual, we can use clone counts to define measures such as the richness

$$R(\mathbf{n}) := \sum_{i=1}^{\Omega} \mathbb{1}(n_i \geq 1) = \sum_{k\geq 1} \sum_{i=1}^{\Omega} \mathbb{1}(n_i, k) \equiv \sum_{k\geq 1} \hat{c}_k, \tag{5}$$

where $\hat{c}_k \equiv \sum_{i=1}^{\Omega} \mathbb{1}(n_i, k)$ is defined in Eq. (3) (the number of clones that are of size k). Other diversity/entropy measures such as Simpson's indices, Gini indices, etc. Rempala and Seweryn (2013) and Xu et al. (2020) can also be straightforwardly defined. Given the clone populations $\mathbf{n}$, the individual richness can be found by direct enumeration of Eq. (5).

We can also express the richness in terms of the underlying probabilities $\mathbf{p}$ associated with the individual by first finding the probability ρ_i that a type-i cell appears at all among the N cells within said individual. This probability is

$$\rho_i \equiv \mathbb{P}(n_i \geq 1|\mathbf{p}, N) = 1 - (1 - p_i)^N \tag{6}$$

and corresponds to that of a binary outcome, either appearing or not appearing. Higher order probabilities like ρ_{ij} (both i- and j-type cells appearing in a specific individual) can be computed using the marginalized probability

$$\mathbb{P}(n_i, n_j|\mathbf{p}, N) = \frac{N!\, p_i^{n_i} p_j^{n_j} (1 - p_i - p_j)^{N-n_i-n_j}}{n_i!\, n_j!\, (N - n_i - n_j)!} \tag{7}$$

to construct

$$\begin{aligned} \rho_{ij} &\equiv \mathbb{P}(n_i, n_j \geq 1|\mathbf{p}, N) \\ &= 1 + (1 - p_i - p_j)^N - (1 - p_i)^N - (1 - p_j)^N. \end{aligned} \tag{8}$$

Higher moments can be straightforwardly computed using quantities such as

$$\begin{aligned} \rho_{ijk} &\equiv \mathbb{P}(n_i, n_j, n_k \geq 1|\mathbf{p}, N) \\ &= 1 - (1 - p_i - p_j - p_k)^N - \sum_{\ell=i,j,k} (1 - p_\ell)^N + \sum_{q\neq\ell=i,j,k} (1 - p_q - p_\ell)^N. \end{aligned} \tag{9}$$

These expressions arise when we compute the moments of R [defined by Eq. (5)] in terms of the probabilities $\mathbf{p}$. Using the single-individual multinomial probability $\mathbb{P}(\mathbf{n}|\mathbf{p}, N)$ (Eq. (4)) allows us to express moments of the richness in a single individual in terms of the underlying system probabilities $\mathbf{p}$. The first two are given by

$$\begin{aligned} \mathbb{E}[R(\mathbf{p})] &= \sum_{\mathbf{n}} \sum_{i=1}^{\Omega} \mathbb{1}(n_i \geq 1) \mathbb{P}(\mathbf{n}|\mathbf{p}, N) = \sum_{i=1}^{\Omega} \mathbb{P}(n_i \geq 1|\mathbf{p}, N) = \sum_{i=1}^{\Omega} \rho_i, \\ \mathbb{E}[R^2(\mathbf{p})] &= \sum_{\mathbf{n}} \left[\sum_{i=1}^{\Omega} \mathbb{1}(n_i \geq 1) \right]^2 \mathbb{P}(\mathbf{n}|\mathbf{p}, N) \\ &= \sum_{i,j=1}^{\Omega} \mathbb{P}(n_i, n_j \geq 1|\mathbf{p}, N) \equiv \mathbb{E}[R] + \sum_{j\neq i}^{\Omega} \rho_{ij}. \end{aligned} \tag{10}$$

3.2 Multi-individual Quantities

Here, we consider the distribution $\mathbf{n}^{(m)}$ across different individuals and construct quantities describing group properties. For example, the combined richness of all TCR clones of M individuals is defined as

$$R^{(M)}(\{\mathbf{n}^{(m)}\}) := \sum_{k\geq 1}\sum_{i=1}^{\Omega} \mathbb{1}\Big(\sum_{m=1}^{M} n_i^{(m)}, k\Big). \tag{11}$$

To express the expected multi-individual richness in terms of the underlying individual systems probabilities $\mathbf{p}^{(m)}$, we weight Eq. (11) over the M-individual probability

$$\mathbb{P}_M(\{\mathbf{n}^{(m)}\}|\{\mathbf{p}^{(m)}, N^{(m)}\}) \equiv \prod_{m=1}^{M} \mathbb{P}(\{\mathbf{n}^{(m)}\}|\{\mathbf{p}^{(m)}, N^{(m)}\}), \tag{12}$$

and sum over all allowable $\mathbf{n}^{(m)}$. For computing the first two moments of the total-population richness in terms of $\mathbf{p}^{(m)}$, we will make use of the marginalized probability $\tilde{\rho}_i$ that clone i appears in at least one of the M individuals

$$\begin{aligned}\tilde{\rho}_i &\equiv \mathbb{P}\Big(\sum_{m=1}^{M} n_i^{(m)} \geq 1 \,\Big|\, \{\mathbf{p}^{(m)}, N^{(m)}\}\Big) = 1 - \mathbb{P}\left(n_i^{(m)} = 0 \;\forall\, m\right) \\ &= 1 - \prod_{m=1}^{M} \left(1 - p_i^{(m)}\right)^{N^{(m)}}.\end{aligned} \tag{13}$$

Note that $\tilde{\rho}_i > \prod_{m=1}^{M} \rho_i^{(m)}$ describes the probability that a type i cell occurs at all in the total population, while $\prod_{m=1}^{M} \rho_i^{(m)}$ describes the probability that a type i cell appears in each of the M individuals.

We will also need the joint probability that clones i and j both appear in at least one of the M individuals $\mathbb{P}\big(\sum_{m=1}^{M} n_i^{(m)} \geq 1, \sum_{\ell=1}^{M} n_j^{(\ell)} \geq 1|\{\mathbf{p}^{(m)}, N^{(m)}\}\big)$, which we can decompose as

$$\begin{aligned}&\mathbb{P}\Big(\sum_{m=1}^{M} n_i^{(m)} \geq 1, \sum_{\ell=1}^{M} n_j^{(\ell)} \geq 1 \,\Big|\, \{\mathbf{p}^{(m)}, N^{(m)}\}\Big) \\ &\qquad = 1 - \mathbb{P}\Big(\sum_{m=1}^{M} n_i^{(m)} \geq 1, \sum_{\ell=1}^{M} n_j^{(\ell)} = 0 \,\Big|\, \{\mathbf{p}^{(m)}, N^{(m)}\}\Big) \\ &\qquad\quad - \mathbb{P}\Big(\sum_{m=1}^{M} n_i^{(m)} = 0, \sum_{\ell=1}^{M} n_j^{(\ell)} \geq 1 \,\Big|\, \{\mathbf{p}^{(m)}, N^{(m)}\}\Big) \\ &\qquad\quad - \mathbb{P}\Big(\sum_{m=1}^{M} n_i^{(m)} = \sum_{\ell=1}^{M} n_j^{(\ell)} = 0 \,\Big|\, \{\mathbf{p}^{(m)}, N^{(m)}\}\Big).\end{aligned} \tag{14}$$

Upon using Eqs. (4) and (7), we find

$$\begin{aligned}\mathbb{P}\Big(\sum_{m=1}^{M} n_i^{(m)} \geq 1, \sum_{\ell=1}^{M} n_j^{(\ell)} = 0 \,\Big|\{\mathbf{p}^{(m)}, N^{(m)}\}\Big) \\ = \prod_{m=1}^{M}\big(1-p_i^{(m)}\big)^{N^{(m)}} - \prod_{m=1}^{M}\big(1-p_i^{(m)}-p_j^{(m)}\big)^{N^{(m)}},\end{aligned} \tag{15}$$

allowing us to rewrite Eq. (14) as

$$\begin{aligned}\tilde{\rho}_{ij} &\equiv \mathbb{P}\Big(\sum_{m=1}^{M} n_i^{(m)} \geq 1, \sum_{\ell=1}^{M} n_j^{(\ell)} \geq 1 \,\Big|\{\mathbf{p}^{(m)}, N^{(m)}\}\Big) \\ &= 1 - \prod_{m=1}^{M}\big(1-p_i^{(m)}\big)^{N^{(m)}} - \prod_{m=1}^{M}\big(1-p_j^{(m)}\big)^{N^{(m)}} \\ &\quad + \prod_{m=1}^{M}\big(1-p_i^{(m)}-p_j^{(m)}\big)^{N^{(m)}}.\end{aligned} \tag{16}$$

Note also that $\tilde{\rho}_{ij} > \prod_{m=1}^{M} \rho_{ij}^{(m)}$.

Using the above definitions, we can express the mean total-population richness as

$$\begin{aligned}\mathbb{E}[R^{(M)}(\{\mathbf{p}^{(m)}\})] &= \sum_{\mathbf{n}^{(m)}}\sum_{i=1}^{\Omega}\sum_{k\geq 1} \mathbb{1}\Big(\sum_{\ell=1}^{M} n_i^{(\ell)}, k\Big) \prod_{m=1}^{M} \mathbb{P}(\{\mathbf{n}^{(m)}\}|\{\mathbf{p}^{(m)}, N^{(m)}\}) \\ &= \sum_{i=1}^{\Omega} \mathbb{P}\Big(\sum_{m=1}^{M} n_i^{(m)} \geq 1 \,\Big|\{\mathbf{p}^{(m)}, N^{(m)}\}\Big) \\ &= \sum_{i=1}^{\Omega}\left[1 - \prod_{m=1}^{M}(1-p_i^{(m)})^{N^{(m)}}\right] \equiv \sum_{i=1}^{\Omega}\tilde{\rho}_i \\ &= \Omega - \sum_{i=1}^{\Omega}\prod_{m=1}^{M}(1-p_i^{(m)})^{N^{(m)}} \\ &\approx \Omega - \sum_{i=1}^{\Omega} e^{-\sum_{m=1}^{M} p_i^{(m)} N^{(m)}},\end{aligned} \tag{17}$$

where the last approximation holds for $p_i^{(m)} \ll 1, N^{(m)} \gg 1$. The second moment of the total M-population richness can also be found in terms of $\mathbb{E}[R^{(M)}]$ and Eq. (16),

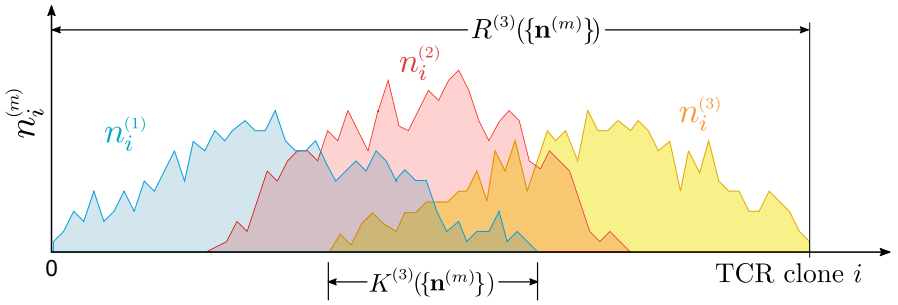

Fig. 3 Three individuals with overlapping cell number distributions $n_i^{(m)}$, $m = 1, 2, 3$. The richness $R^{(3)}$ is the total number of distinct TCRs found across all individuals, and is defined in Eq. (11). The overlap $K^{(3)}$ is the number of TCR clones found in all three individuals, as defined in Eq. (19). For visual simplicity, the set of clones i present in each individual are drawn to be contiguous. When considering subsampling of cells from each individual, $K^{(M)}$ will be reduced since some cell types i will be lost. The corresponding values, $s_i^{(m)}$, $K_{\mathrm{s}}^{(M)}$, and $R_{\mathrm{s}}^{(M)}$ can be constructed from Eqs. (23) and (24) reflecting the losses from subsampling

$$\begin{aligned}
\mathbb{E}\Big[\Big(R^{(M)}(\{\mathbf{p}^{(m)}\})\Big)^2\Big] &= \sum_{\{\mathbf{n}^{(M)}\}}\Bigg[\sum_{i=1}^{\Omega}\mathbb{1}\Big(\sum_{m=1}^{M}n_i^{(m)}\geq 1\Big)\Bigg]^2\prod_{m=1}^{M}\mathbb{P}\left(\{\mathbf{n}^{(m)}\}|\{\mathbf{p}^{(m)},N^{(m)}\}\right)\\
&= \sum_{i,j=1}^{\Omega}\mathbb{P}\Big(\sum_{m=1}^{M}n_i^{(m)}\geq 1,\sum_{\ell=1}^{M}n_j^{(\ell)}\geq 1\,\Big|\,\{\mathbf{p}^{(m)},N^{(m)}\}\Big)\\
&= \sum_{i=1}^{\Omega}\mathbb{P}\Big(\sum_{m=1}^{M}n_i^{(m)}\geq 1\,\Big|\,\{\mathbf{p}^{(m)},N^{(m)}\}\Big)\\
&\quad+\sum_{i\neq j}^{\Omega}\mathbb{P}\Big(\sum_{m=1}^{M}n_i^{(m)}\geq 1,\sum_{\ell=1}^{M}n_j^{(\ell)}\geq 1\,\Big|\,\{\mathbf{p}^{(m)},N^{(m)}\}\Big)\\
&= \mathbb{E}[R^{(M)}(\{\mathbf{p}^{(m)}\})]+\sum_{i\neq j}^{\Omega}\tilde{\rho}_{ij}.
\end{aligned}\tag{18}$$

Given $\mathbf{n}^{(m)}$ of all individuals, we can also easily define the number of distinct TCR clones that appear in all of M randomly selected individuals, the "M-overlap" or "M-publicness"

$$K^{(M)}(\{\mathbf{n}^{(m)}\}):=\sum_{i=1}^{\Omega}\prod_{m=1}^{M}\sum_{k^{(m)}\geq 1}\mathbb{1}(n_i^{(m)},k^{(m)}).\tag{19}$$

Figure 3 provides a simple example of three individuals each with a contiguous distribution of cell numbers $n_i^{(m)}$ that overlap.

As with Eqs. (5) and (10), we can express the overlap in terms of the underlying individual probabilities $\mathbf{p}^{(m)}$ by weighting Eq. (19) by the M-population probability

$\prod_{m=1}^{M} \mathbb{P}(\mathbf{n}^{(m)}|\mathbf{p}^{(m)}, N^{(m)})$ (see Eq. (4)) to find

$$\begin{aligned}
\mathbb{E}[K^{(M)}(\{\mathbf{p}^{(m)}\})] &= \sum_{\{\mathbf{n}^{(m)}\}} \left[\sum_{i=1}^{\Omega}\prod_{m=1}^{M} \mathbb{1}(n_i^{(m)} \geq 1)\right] \mathbb{P}_M(\{\mathbf{n}^{(m)}\}|\{\mathbf{p}^{(m)}, N^{(m)}\}) \\
&= \sum_{i=1}^{\Omega}\prod_{m=1}^{M} \mathbb{P}(n_i^{(m)} \geq 1|\{\mathbf{p}^{(m)}, N^{(m)}\}) \\
&= \sum_{i=1}^{\Omega}\prod_{m=1}^{M} \left[1 - \left(1 - p_i^{(m)}\right)^{N^{(m)}}\right] \equiv \sum_{i=1}^{\Omega}\prod_{m=1}^{M} \rho_i^{(m)}
\end{aligned} \tag{20}$$

$$\begin{aligned}
&\mathbb{E}\left[\left(K^{(M)}(\{\mathbf{p}^{(m)}\})\right)^2\right] \\
&= \sum_{\{\mathbf{n}^{(m)}\}} \left[\sum_{i=1}^{\Omega}\prod_{m=1}^{M} \mathbb{1}(n_i^{(m)} \geq 1)\right]^2 \mathbb{P}_M(\{\mathbf{n}^{(m)}\}|\{\mathbf{p}^{(m)}, N^{(m)}\}) \\
&= \sum_{i,j=1}^{\Omega}\prod_{m=1}^{M} \mathbb{P}(n_i^{(m)}, n_j^{(m)} \geq 1|\{\mathbf{p}^{(m)}, N^{(m)}\}) \\
&= \sum_{i=1}^{\Omega}\prod_{m=1}^{M} \left[1 - (1 - p_i^{(m)})^{N^{(m)}}\right] \\
&\quad + \sum_{j\neq i}^{\Omega}\prod_{m=1}^{M} \left[1 + (1 - p_i^{(m)} - p_j^{(m)})^{N^{(m)}} - (1 - p_i^{(m)})^{N^{(m)}} - (1 - p_j^{(m)})^{N^{(m)}}\right] \\
&\equiv \mathbb{E}[K^{(M)}(\{\mathbf{p}^{(m)}\})] + \sum_{j\neq i}^{\Omega}\prod_{m=1}^{M} \rho_{ij}^{(m)}.
\end{aligned} \tag{21}$$

The expected number of clones shared among all M individuals, $\mathbb{E}[K^{(M)}]$, provides a natural measure of M-overlap. Clearly, $\mathbb{E}[K^{(M)}] < \mathbb{E}[K^{(M')}]$ if $M > M'$. As with M-publicness, we can identify the expected M-privateness as $\Omega - \mathbb{E}\left[K^{(M)}\right]$, the expected number of clones that are not shared by all M individuals, i.e., that occur in at most $M-1$ individuals. This "privateness" is related to a multi-distribution generalization of the "dissimilarity probability" of samples from two different discrete distributions (Hampton and Lladser 2012). Variations in M-overlap associated with a certain cell-type distribution are captured by the variance $\mathrm{var}\left[K^{(M)}\right] = \mathbb{E}\left[(K^{(M)})^2\right] - \mathbb{E}\left[K^{(M)}\right]^2$. If the total number of sequences Ω is very large, parallelization techniques (see Sect. 4) should be employed to evaluate the term $\sum_{j\neq i}^{\Omega}\prod_{m=1}^{M} \rho_{ij}^{(m)}$ in $\mathbb{E}\left[(K^{(M)})^2\right]$.

A more specific definition of overlap or privateness may be that a clone must appear in at least some specified fraction of M tested individuals. To find the probability that $M_i \leq M$ individuals share at least one cell of a single type i, we use the Poisson binomial distribution describing independent Bernoulli trials on individuals with different

success probabilities $\rho_i^{(m)} \equiv \rho(n_i^{(m)} \geq 1)$:

$$\mathbb{P}(M_i|\{p_i^{(m)}\}) = \sum_{A \in F_{M_i}} \prod_{m \in A} \rho_i^{(m)} \prod_{\ell \in A^c} (1 - \rho_i^{(\ell)}), \tag{22}$$

where F_{M_i} is the set of all subsets of M_i integers that can be selected from the set $(1, 2, 3, \ldots, M)$ and A^c is the complement of A. Equation (22) gives a probabilistic measure of the prevalence of TCR i across M individuals. For example, one can use it to define a mean frequency $\mathbb{E}[M_i]/M$. One can evaluate Eq. (22) recursively or using Fourier transforms, particularly for $M < 20$ (Chen and Liu 1997; Hong 2013).

3.3 Subsampling

The results above are described in terms of the entire cell populations $\mathbf{n}^{(m)}$ or their intrinsic generation probabilities $\mathbf{p}^{(m)}$. In practice, one cannot measure $n_i^{(m)}$ or even $N^{(m)}$ in any individual m. Rather, we can only sample a much smaller number of cells $S^{(m)} \ll N^{(m)}$ from individual m, as shown in Fig. 2. Within this subsample from individual m, we can count the number $s_i^{(m)}$ of type-i cells. Since only subsamples are available, we wish to define quantities such as probability of occurrence, richness, and overlap in terms of the cell counts $\mathbf{s}^{(m)} \equiv \{s_i^{(m)}\}$ in the sample extracted from an individual. Quantities such as *sampled* richness and overlap can be defined in the same way except with $\mathbf{s}^{(m)}$ as the underlying population configuration. To start, first assume that the cell count $\mathbf{n}$ in a specific individual is given. If that individual has N cells of which S are sampled, the probability of observing the population $\mathbf{s} = \{s_1, s_2, \ldots, s_\Omega\}$ in the sample is given by (assuming all cells are uniformly distributed and randomly subsampled at once, without replacement) (Chao and Lin 2012)

$$\mathbb{P}(\mathbf{s}|\mathbf{n}, S, N) = \frac{1}{\binom{N}{S}} \prod_{i=1}^{\Omega} \binom{n_i}{s_i}, \quad \sum_{i=1}^{\Omega} s_i = S. \tag{23}$$

The probability that cell type j appears in the sample from an individual with population $\mathbf{n}$ can be found by marginalizing over all $s_{j \neq i}$, giving

$$\sigma_i \equiv \mathbb{P}(s_i \geq 1|\mathbf{n}, S, N) = 1 - \frac{\binom{N-n_i}{S}}{\binom{N}{S}}. \tag{24}$$

This result can be generalized to more than one TCR clone present. For example, the probability that both clones i and j are found in a sample is

$$\sigma_{ij} \equiv \mathbb{P}(s_i, s_j \geq 1|\mathbf{n}, S, N) = 1 + \frac{\binom{N-n_i-n_j}{S}}{\binom{N}{S}} - \frac{\binom{N-n_i}{S}}{\binom{N}{S}} - \frac{\binom{N-n_j}{S}}{\binom{N}{S}}. \tag{25}$$

Using Eq. (23) as the probability distribution, we can also find the probability that clone i appears in any of the M $S^{(m)}$-sized samples

$$\tilde{\sigma}_i \equiv \mathbb{P}\Big(\sum_{m=1}^{M} s_i^{(m)} \geq 1 \,\Big|\, \{\mathbf{n}^{(m)}, S^{(m)}, N^{(m)}\}\Big) = 1 - \prod_{m=1}^{M} \frac{\binom{N^{(m)}-n_i^{(m)}}{S^{(m)}}}{\binom{N^{(m)}}{S^{(m)}}}, \tag{26}$$

and the joint probabilities that clones i and j appear in any sample

$$\begin{aligned}\tilde{\sigma}_{ij} &\equiv \mathbb{P}\Big(\sum_{m=1}^{M} s_i^{(m)} \geq 1, \sum_{\ell=1}^{M} s_j^{(\ell)} \geq 1 \,\Big|\, \{\mathbf{n}^{(m)}, S^{(m)}, N^{(m)}\}\Big) \\ &= 1 + \prod_{m=1}^{M} \frac{\binom{N^{(m)}-n_i^{(m)}-n_j^{(m)}}{S^{(m)}}}{\binom{N^{(m)}}{S^{(m)}}} - \prod_{m=1}^{M} \frac{\binom{N^{(m)}-n_i^{(m)}}{S^{(m)}}}{\binom{N^{(m)}}{S^{(m)}}} - \prod_{m=1}^{M} \frac{\binom{N^{(m)}-n_j^{(m)}}{S^{(m)}}}{\binom{N^{(m)}}{S^{(m)}}}.\end{aligned} \tag{27}$$

Quantities such as richness and publicness *measured within samples* from the group can be analogously defined in terms of clonal populations $\mathbf{s}^{(m)}$:

$$R_{\mathrm{s}}(\mathbf{s}) := \sum_{k\geq 1} \sum_{i=1}^{\Omega} \mathbb{1}(s_i, k), \tag{28}$$

$$R_{\mathrm{s}}^{(M)}(\{\mathbf{s}^{(m)}\}) := \sum_{k\geq 1} \sum_{i=1}^{\Omega} \mathbb{1}\Big(\sum_{m=1}^{M} s_i^{(m)}, k\Big), \tag{29}$$

and

$$K_{\mathrm{s}}^{(M)}(\{\mathbf{s}^{(m)}\}) := \sum_{i=1}^{\Omega} \prod_{m=1}^{M} \sum_{k^{(m)}\geq 1} \mathbb{1}(s_i^{(m)}, k^{(m)}). \tag{30}$$

For a given $\mathbf{n}^{(m)}$, these quantities can be first averaged over the sampling distribution Eq. (23) to express them in terms of $\mathbf{n}^{(m)}$ and to explicitly reveal the effects of random sampling. The first two moments of R_{s}, $R_{\mathrm{s}}^{(M)}$, and $K_{\mathrm{s}}^{(M)}$ expressed in terms of $\mathbf{n}^{(m)}$ can be easily found by weighting Eqs. (28), (29), and (30) by $\mathbb{P}(\mathbf{s}|\mathbf{n}, S, N)$ and $P^{(M)} = \prod_{m=1}^{M} \mathbb{P}(\mathbf{s}^{(m)}|\mathbf{n}^{(m)}, S^{(m)}, N^{(m)})$:

$$\begin{aligned}\mathbb{E}[R_{\mathrm{s}}(\mathbf{n})] &= \Omega - \frac{1}{\binom{N}{S}} \sum_{i=1}^{\Omega} \binom{N-n_i}{S} \equiv \sum_{i=1}^{\Omega} \sigma_i, \\ \mathbb{E}\big[\big(R_{\mathrm{s}}(\mathbf{n})\big)^2\big] &= \mathbb{E}[R_{\mathrm{s}}(\mathbf{n})] + \sum_{i\neq j}^{\Omega} \sigma_{ij}\end{aligned} \tag{31}$$

$$
\begin{aligned}
&\mathbb{E}\big[R_{\mathrm{s}}^{(M)}(\{\mathbf{n}^{(m)}\})\big] \\
&\quad= \sum_{\{\mathbf{s}^{(m)}\}} \sum_{i=1}^{\Omega} \mathbb{1}\Big(\sum_{m=1}^{M} s_i^{(m)} \geq 1\Big) \prod_{m=1}^{M} \mathbb{P}(\{\mathbf{s}^{(m)}\}|\{\mathbf{n}^{(m)}, S^{(m)}, N^{(m)}\}) \\
&\quad= \sum_{i=1}^{\Omega} \mathbb{P}\Big(\sum_{m=1}^{M} s_i^{(m)} \geq 1 \,\Big|\, \{\mathbf{n}^{(m)}, S^{(m)}, N^{(m)}\}\Big) \\
&\quad= \Omega - \sum_{i=1}^{\Omega} \prod_{m=1}^{M} \frac{\binom{N-n_i^{(m)}}{S}}{\binom{N}{S}} \equiv \sum_{i=1}^{\Omega} \tilde{\sigma}_i, \\
&\mathbb{E}\Big[\big(R_{\mathrm{s}}^{(M)}(\{\mathbf{n}^{(m)}\})\big)^2\Big] \\
&\quad= \sum_{\{\mathbf{n}^{(m)}\}} \left[\sum_{i=1}^{\Omega} \mathbb{1}\Big(\sum_{m=1}^{M} s_i^{(m)} \geq 1\Big)\right]^2 \prod_{m=1}^{M} \mathbb{P}(\{\mathbf{s}^{(m)}\}|\{\mathbf{n}^{(m)}, S^{(m)}, N^{(m)}\}) \\
&\quad= \sum_{i,j=1}^{\Omega} \prod_{m=1}^{M} \mathbb{P}\Big(\sum_{m=1}^{M} s_i^{(m)}, \sum_{\ell=1}^{M} s_j^{(m)} \geq 1 \,\Big|\, \{\mathbf{n}^{(m)}, S^{(m)}, N^{(m)}\}\Big) \\
&\quad= \mathbb{E}\big[R_{\mathrm{s}}^{(M)}(\{\mathbf{n}^{(m)}\})\big] + \sum_{i\neq j=1}^{\Omega} \tilde{\sigma}_{ij},
\end{aligned}
\tag{32}
$$

$$
\begin{aligned}
&\mathbb{E}\big[K_{\mathrm{s}}^{(M)}(\{\mathbf{n}^{(m)}\})\big] \\
&\quad= \sum_{\{\mathbf{s}^{(m)}\}} \sum_{i=1}^{\Omega} \prod_{m=1}^{M} \mathbb{1}(s_i^{(m)} \geq 1) \mathbb{P}(\mathbf{s}^{(m)}|\{\mathbf{n}^{(m)}, S^{(m)}, N^{(m)}\}) \\
&\quad= \sum_{i=1}^{\Omega} \prod_{m=1}^{M} \mathbb{P}(s_i^{(m)} \geq 1|\{\mathbf{n}^{(m)}, S^{(m)}, N^{(m)}\}) \\
&\quad= \sum_{i=1}^{\Omega} \prod_{m=1}^{M} \left[1 - \frac{\binom{N^{(m)}-n_i^{(m)}}{S^{(m)}}}{\binom{N^{(m)}}{S^{(m)}}}\right] \equiv \sum_{i=1}^{\Omega} \prod_{m=1}^{M} \sigma_i^{(m)}, \\
&\mathbb{E}\Big[\big(K_{\mathrm{s}}^{(M)}(\{\mathbf{n}^{(m)}\})\big)^2\Big] \\
&\quad= \sum_{\{\mathbf{s}^{(m)}\}} \left[\sum_{i=1}^{\Omega} \prod_{m=1}^{M} \mathbb{1}(s_i^{(m)} \geq 1)\right]^2 \prod_{m=1}^{M} \mathbb{P}(\{\mathbf{s}^{(m)}\}|\{\mathbf{n}^{(m)}, S^{(m)}, N^{(m)}\}) \\
&\quad= \sum_{i,j=1}^{\Omega} \prod_{m=1}^{M} \mathbb{P}(s_i^{(m)}, s_j^{(m)} \geq 1|\{\mathbf{n}^{(m)}, S^{(m)}, N^{(m)}\}) \\
&\quad\equiv \mathbb{E}\big[K_{\mathrm{s}}^{(M)}(\{\mathbf{n}^{(m)}\})\big] + \sum_{i\neq j}^{\Omega} \prod_{m=1}^{M} \sigma_{ij}^{(m)}.
\end{aligned}
\tag{33}
$$

All of the above quantities can also be expressed in terms of the underlying probabilities $\mathbf{p}^{(m)}$ rather than the population configurations $\mathbf{n}^{(m)}$. To do so, we can further weight Eqs. (31), (32), and (33) over the probability Eq. (4) to render these quantities in terms of the underlying probabilities $\mathbf{p}^{(m)}$. However, we can also first convolve Eq. (23) with the multinomial distribution in Eq. (4) (suppressing the individual index m)

$$\mathbb{P}(\mathbf{s}|\mathbf{p}, S, N) = \sum_{\mathbf{n}} \mathbb{P}(\mathbf{s}|\mathbf{n}, S, N)\mathbb{P}(\mathbf{n}|\mathbf{p}, N), \tag{34}$$

along with the implicit constraints $\sum_{i=1}^{\Omega} n_i \equiv N$ and $\sum_{i=1}^{\Omega} s_i = S$ to find

$$\mathbb{P}(\mathbf{s}|\mathbf{p}, S) = S! \prod_{i=1}^{\Omega} \frac{p_i^{s_i}}{s_i!}, \quad \sum_{i=1}^{\Omega} s_i = S, \tag{35}$$

which is a multinomial distribution identical in form to $\mathbb{P}(\mathbf{n}|\mathbf{p}, N)$ in Eq. (4), except with $\mathbf{n}$ replaced by $\mathbf{s}$ and N replaced by S. Uniform random sampling from a multinomial results in another multinomial. Thus, if we use the full multi-individual probability

$$\mathbb{P}_M(\{\mathbf{s}^{(m)}\}|\{\mathbf{p}^{(m)}, S^{(m)}\}) \equiv \prod_{m=1}^{M} \mathbb{P}(\{\mathbf{s}^{(m)}\}|\{\mathbf{p}^{(m)}, S^{(m)}\}) \tag{36}$$

to compute moments of the sampled richness and publicness, they take on the same forms as the expressions associated with the whole-organism quantities. For example, in the $\mathbf{p}$ representation, the probability that clone i appears in the sample from individual m is

$$\rho_i^{(m)}(S) \equiv \mathbb{P}(s_i^{(m)} \geq 1|\{\mathbf{p}^{(m)}, S^{(m)}\}) = 1 - (1 - p_i^{(m)})^{S^{(m)}}, \tag{37}$$

in analogy with Eq. (6), while the two-clone joint probability in the sampled from individual m becomes

$$\begin{aligned} \rho_{ij}^{(m)}(S) &\equiv \mathbb{P}(s_i^{(m)}, s_j^{(m)} \geq 1|\{\mathbf{p}^{(m)}, S^{(m)}\}) \\ &= 1 + (1 - p_i^{(m)} - p_j^{(m)})^{S^{(m)}} - (1 - p_i^{(m)})^{S^{(m)}} - (1 - p_j^{(m)})^{S^{(m)}}, \end{aligned} \tag{38}$$

in analogy with Eq. (8). Similarly, for the overlap quantities, in analogy with Eqs. (13) and (16), we have

$$\begin{aligned} \tilde{\rho}_i(S) &\equiv \mathbb{P}\Big(\sum_{m=1}^{M} s_i^{(m)} \geq 1 \Big| \{\mathbf{p}^{(m)}, S^{(m)}\}\Big) = 1 - \mathbb{P}\left(s_i^{(m)} = 0 \ \forall m\right) \\ &= 1 - \prod_{m=1}^{M} \left(1 - p_i^{(m)}\right)^{S^{(m)}}. \end{aligned} \tag{39}$$

$$\begin{aligned}\tilde{\rho}_{ij}(S) &\equiv \mathbb{P}\Big(\sum_{m=1}^{M} s_i^{(m)} \geq 1, \sum_{\ell=1}^{M} s_j^{(\ell)} \geq 1 \,\Big|\, \{\mathbf{p}^{(m)}, S^{(m)}\}\Big) \\ &= 1 - \prod_{m=1}^{M} \Big(1 - p_i^{(m)}\Big)^{S^{(m)}} - \prod_{m=1}^{M} \Big(1 - p_j^{(m)}\Big)^{S^{(m)}} \\ &\quad + \prod_{m=1}^{M} \Big(1 - p_i^{(m)} - p_j^{(m)}\Big)^{S^{(m)}}. \end{aligned} \tag{40}$$

The expressions for the sampled moments $\mathbb{E}[R_{\mathrm{s}}(\mathbf{p})]$, $\mathbb{E}[R_{\mathrm{s}}^2(\mathbf{p})]$, $\mathbb{E}\big[R_{\mathrm{s}}^{(M)}(\{\mathbf{p}^{(m)}\})\big]$, $\mathbb{E}\big[\big(R_{\mathrm{s}}^{(M)}(\{\mathbf{p}^{(m)}\})\big)^2\big]$, $\mathbb{E}[K_{\mathrm{s}}^{(M)}(\{\mathbf{p}^{(m)}\})]$, and $\mathbb{E}\big[\big(K_{\mathrm{s}}^{(M)}(\{\mathbf{p}^{(m)}\})\big)^2\big]$ follow the same form as their unsampled counterparts given in Eqs. (10), (17), (18), (20), and (21), except with $\rho_i^{(m)}$, $\rho_{ij}^{(m)}$, $\tilde{\rho}_i$, and $\tilde{\rho}_{ij}$ replaced by their $\rho_i^{(m)}(S)$, $\rho_{ij}^{(m)}(S)$, $\tilde{\rho}_i(S)$, and $\tilde{\rho}_{ij}(S)$ counterparts. This simplifying property arises because of the conjugate nature of the multinomial distributions (4), (35), and (34).

In addition to simple expressions for the moments of $K_{\mathrm{s}}^{(M)}$, we can also find expressions for the probability distribution over the values of $K_{\mathrm{s}}^{(M)}$. In terms of $\mathbf{n}^{(m)}$, since the probability that $s_i^{(m)} \geq 1$ in the samples from all $1 \leq m \leq M$ individuals is $\sigma_i \equiv \prod_{m=1}^{M} \mathbb{P}(s_i^{(m)} \geq 1 | \{\mathbf{n}^{(m)}, S^{(m)}, N^{(m)}\}) = \prod_{m=1}^{M} \sigma_i^{(m)}$, the probability that exactly k clones are shared by all M samples is

$$\mathbb{P}(K_{\mathrm{s}}^{(M)} = k | \{\sigma_i^{(m)}\}) = \sum_{A \in F_k} \prod_{i \in A} \Big(\prod_{m=1}^{M} \sigma_i^{(m)}\Big) \prod_{j \in A^{\mathrm{c}}} \Big[1 - \prod_{m=1}^{M} \sigma_j^{(m)}\Big], \tag{41}$$

where F_k is the set of all subsets of k integers that can be selected from the set $\{1, 2, 3, \ldots, K^{(M)}\}$ and A^{c} is the complement of A. Equation (41) is the Poisson binomial distribution, but this time the underlying success probabilities $\prod_{m=1}^{M} \sigma_i^{(m)}$ across all M individuals vary with TCR clone identity i.

Finally, inference of individual measures from subsamples can be formulated. One can use the sampling likelihood function $\mathbb{P}(\mathbf{s}|\mathbf{n}, S, N)$, Bayes' rule, and the multinomial (conjugate) prior $\mathbb{P}(\mathbf{n}|\mathbf{p}, N)$ to construct the posterior probability of $\mathbf{n}$ *given* a sampled configuration $\mathbf{s}$:

$$\mathbb{P}(\mathbf{n}|\mathbf{s}, S, N, \mathbf{p}) = \frac{\mathbb{P}(\mathbf{s}|\mathbf{n}, S, N)\mathbb{P}(\mathbf{n}|\mathbf{p}, N)}{\sum_{\mathbf{n}} \mathbb{P}(\mathbf{s}|\mathbf{n}, S, N)\mathbb{P}(\mathbf{n}|\mathbf{p}, N)}. \tag{42}$$

The normalization in Eq. (42) has already been found in Eqs. (34) and (35). Thus, we find the posterior

$$\mathbb{P}(\mathbf{n}|\mathbf{s}, S, N, \mathbf{p}) = (N - S)! \prod_{i=1}^{\Omega} \frac{p_i^{n_i - s_i}}{(n_i - s_i)!}, \quad \sum_{i=1}^{\Omega} (n_i - s_i) = N - S \tag{43}$$

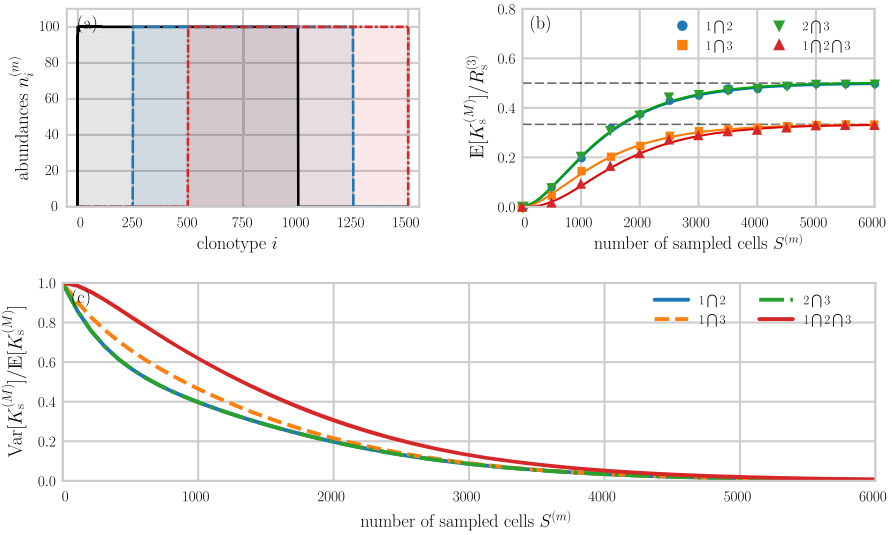

Fig. 4 Sampling from shifted uniform distributions. **a** Synthetic TCR or BCR distributions of $M = 3$ individuals. The distributions in individuals 1, 2, and 3 are indicated by solid black, dashed blue, and dash-dotted red lines, respectively. Each individual has 10^5 cells uniformly distributed across 1000 clones (100 cells per clone). The sampled group richness $R_s^{(3)}$ is 1500. **b** Samples of size $S^{(m)}$ have been generated to compute the relative overlaps between individuals 1 and 2 (blue disks), 2 and 3 (green inverted triangles), 1 and 3 (orange squares), and 1–3 (red triangles). The solid lines show the corresponding analytical solutions $\mathbb{E}\big[K_s^{(M)}\big]/R_s^{(3)}$ (see Eq. (20)). Dashed grey lines show the maximum possible relative overlaps $500/1500 \approx 0.33$ and $750/1500 = 0.5$. **c** The Fano factor $\mathrm{var}\big[K_s^{(M)}\big]/\mathbb{E}\big[K_s^{(M)}\big]$ associated with relative overlaps between individuals 1 and 2 (solid blue line), 2 and 3 (dash-dotted green line), 1 and 3 (dashed orange line), and 1–3 (solid red line) as a function of the number of sampled cells $S^{(m)}$ (Color figure online)

in terms of the hyperparameters **p**. Using this posterior, we can calculate the expectation of the whole organism richness $R = \sum_{k\geq 1} \sum_{i=1}^{\Omega} \mathbb{1}(n_i, k)$,

$$\mathbb{E}[R(\mathbf{s}, \mathbf{p})] = \Omega - \sum_{j|s_j=0} \big(1 - p_j\big)^{N-S}, \tag{44}$$

which depends on the sampled configuration only through the sample-absent clones j. Bayesian methods for estimating overlap between two populations from samples have also been recently explored (Larremore 2019).

4 Simulations

The sampling theory derived in the previous sections is useful for understanding the effect of different sampling distributions on measurable quantities such as the proportion of shared TCRs and BCRs among different individuals. Figures 4 and 5 show two examples of receptor distributions, along with the respective relative overlaps and Fano factors, for three individuals. To illustrate our methodology clearly and concisely, we utilize three shifted uniform distributions as models of synthetic sequence

distributions in Fig. 4. In this example, the number of TCR or BCR sequences per individual is $N^{(m)} = 10^5$, $(m = 1, 2, 3)$, and the sampled group richness $R_{\mathrm{s}}^{(3)} = 1500$. Based on the abundance curves shown in Fig. 4a, we can readily obtain the overlaps between individuals 1–3 (solid black, dashed blue, and dash-dotted red lines), as well as between all pairs of individuals. The maximum possible overlap, normalized by $R_{\mathrm{s}}^{(3)}$, between all three individuals and between individuals 1 and 3 is $500/1500 \approx 0.33$. For the two remaining pairs, the corresponding maximum relative overlap, normalized by the richness associated with all three sampled individuals, is $750/1500 = 0.5$.

Using $S^{(m)} < N^{(m)}$ sampled cells from each individual, we observe in Fig. 4b that the increase of $\mathbb{E}[K_{\mathrm{s}}^{(M)}]/R_{\mathrm{s}}^{(3)}$ with $S^{(m)}$ is well-described by Eq. (20). In Fig. 4c, we plot the Fano factor $\mathrm{var}[K_{\mathrm{s}}^{(M)}]/\mathbb{E}[K_{\mathrm{s}}^{(M)}]$ as a function of the number of sampled cells $S^{(m)}$ from each individual. For sample sizes of about 1000 (i.e., about 1% of the total sequence population), the Fano factor is between 0.4 (for overlaps between individuals 1 and 2 and between individuals 2 and 3) and 0.6 (for the overlap between individuals 1–3). As sample sizes reach about 5% of the total number of sequences (10^5), the variance $\mathrm{var}[K_{\mathrm{s}}^{(M)}]$ becomes negligible with respect to the expected overlap $\mathbb{E}[K_{\mathrm{s}}^{(M)}]$.

As an example of an application to empirical TRB CDR3 data, we used the SONIA package (Elhanati et al. 2014) to generate amino acid sequence data for three individuals, each with $N^{(m)} = 10^5$ cells. The combined richness across all individuals is $R_{\mathrm{s}}^{(3)} = 284,598$. We show the abundances of all sequences in Fig. 5a. The majority of sequences has an abundance of 1 while only very few sequences have abundances that exceed 5. Figure 5b shows the expected number of shared sequences as a function of the sampled number of cells $S^{(m)}$. To evaluate Eq. (33), we compute the binomial terms in $\tilde{\sigma}_i$ and $\tilde{\sigma}_{ij}$ by expanding them according to, e.g.,

$$\frac{\binom{N^{(m)}-n_i^{(m)}}{S^{(m)}}}{\binom{N^{(m)}}{S^{(m)}}} = \prod_{\ell=1}^{n_i^{(m)}} \left(1 - \frac{S^{(m)}}{N^{(m)} - n_i^{(m)} + \ell}\right), \tag{45}$$

where $S^{(m)}/N^{(m)}$ is the sample fraction drawn from the m^{th} individual. For large n_i, other approximations, including variants of Stirling's approximations can be employed for fast and accurate evaluation of binomial terms.

We compare these number-representation results with the **p**-representation results by using the estimates $\hat{p}_i^{(m)} = n_i^{(m)}/N^{(m)}$ in $\rho_i^{(m)}(S)$ and $\rho_{ij}^{(m)}(S)$ to compute the quantities in Eqs. (20) and (21). If the number of sampled cells $S^{(m)}$ is not too large, the analytic approximation of using $\hat{p}_i^{(m)}$ in $\rho_i^{(m)}(S)$ to calculate $\mathbb{E}[K_{\mathrm{s}}^{(M)}]/R_{\mathrm{s}}^{(3)}$ is fairly accurate, as shown by the dashed curves in Fig. 5b. Since the abundances of the majority of sequences are very small, finite-size effects lead to deviations from the naive approximation (37) as the numbers of sampled cells $S^{(m)}$ grows large. Of course, we can also extract generation probabilities from SONIA and directly use Eq. (20) and $\rho_i^{(m)}(S)$ from Eq. (37) to find the **p**-representation M-overlap $\mathbb{E}[K_{\mathrm{s}}^{(M)}(\{\mathbf{p}^{(m)}\})]/R_{\mathrm{s}}^{(3)}$.

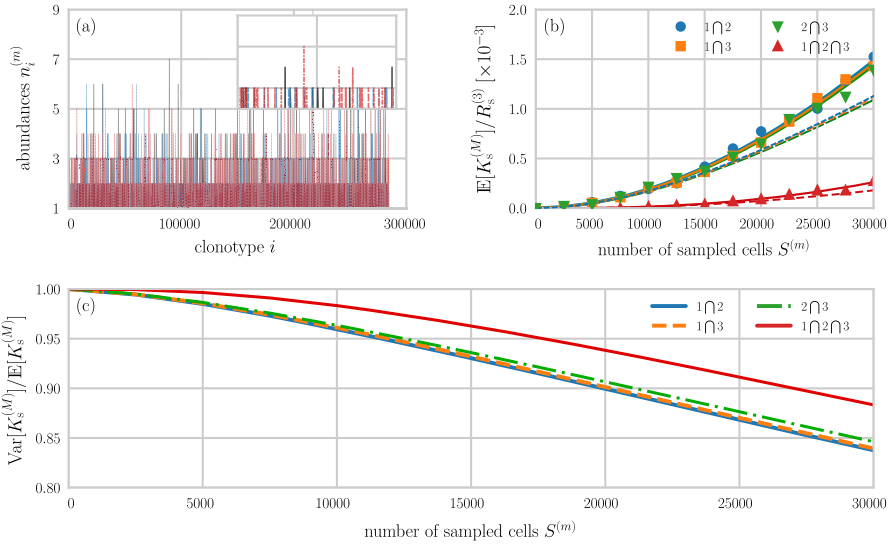

Fig. 5 Sampling from empirical TRB CDR3 distributions and overlap measures in the number representation. **a** Distributions of TRB CDR3 cells in $M = 3$ individuals. We used the SONIA package (Elhanati et al. 2014) to generate 10^5 TRB CDR3 sequences for each individual. The sampled group richness $R_s^{(3)}$ was found to be 284, 598. Equal sample sizes $S^{(m)}$ were then drawn. **b** Relative overlaps between individuals 1 and 2 (blue disks), 2 and 3 (green inverted triangles), 1 and 3 (orange squares), and 1–3 (red triangles). The solid lines plot the corresponding analytical solutions $\mathbb{E}\big[K_s^{(M)}(\{\mathbf{n}^{(m)}\})\big]/R_s^{(3)}$ found in Eqs. (33). The dashed curves correspond to using using the estimator $\hat{p}_i^{(m)} = n_i^{(m)}/N^{(m)}$ in the expression $\mathbb{E}\big[K_s^{(M)}(\{\mathbf{p}^{(m)}\})\big]/R_s^{(3)}$ (Eq. (20) evaluated using $\rho_i^{(m)}(S)$ from Eq. (37)). **c** The Fano factor $\mathrm{var}\big[K_s^{(M)}\big]/\mathbb{E}\big[K_s^{(M)}\big]$ associated with relative overlaps between individuals 1 and 2 (solid blue line), 2 and 3 (dash-dotted green line), 1 and 3 (dashed orange line), and 1–3 (solid red line) as a function of the number of sampled cells $S^{(m)}$ (Color figure online)

To examine the variance associated with a given expected number of shared empirical TRB CDR3 sequences, we show the Fano factor $\mathrm{var}[K_s^{(M)}]/\mathbb{E}[K_s^{(M)}]$ as a function of the sample size $S^{(m)}$ in Fig. 5c. For the shown sample sizes up to $S^{(m)} = 3 \times 10^4$, the Fano factor is larger than about 0.85, indicating a relatively large variance $\mathrm{var}[K_s^{(M)}]$. In addition to reporting mean values of measures of sequence sharing (i.e., "overlap" or "publicness") when analyzing empirical receptor sequence data (Elhanati et al. 2018; Ruiz Ortega et al. 2023), we thus recommend to compute $\mathrm{var}[K_s^{(M)}]$ to determine corresponding confidence intervals.

Calculations were performed on an AMD® Ryzen Threadripper 3970 using `Numba` to parallelize the calculation of Eqs. (33) and $\sum_{j\neq i}^{\Omega} \prod_{m=1}^{M} \rho_{ij}^{(m)}$ used in $\mathrm{var}[K^{(M)}]$.

5 Explicit Forms for Power-Law Probabilities

All of our results thus far assume a model or estimate of p_i or n_i, as well as knowledge of at least Ω. For our formulae to be useful, the theoretical maximum richness Ω also needs to be estimated or modeled. Numerous parametric and nonparametric approaches

have been developed in the statistical ecology literature (Chao and Lee 1992; Wang and Lindsay 2005; Gotelli and Colwell 2011; Colwell et al. 2012; Gotelli and Chao 2013; Chiu et al. 2014; Chao and Lin 2012; Chao et al. 2020), as well as expectation maximization methods to self-consistently estimate richness and most likely clone population $\mathbf{n}$ (Kaplinsky and Arnaout 2016).

To explore how our results depend on parameters such as Ω, we derive approximate analytic expressions when the identical individual probabilities $p_i^{(m)} = p_i$ obey truncated power-law distributions:

$$p_i = \frac{i^{-\nu}}{H_\nu(\Omega)}, \quad p_j \le p_i \text{ if } \nu \ge 0, \ \ j \le i, \ \ i, j = 1, 2, \ldots, \Omega, \tag{46}$$

where $H_\nu(\Omega) \equiv \sum_{j=1}^{\Omega} j^{-\nu}$. Such power laws are good approximations to measured ranked T cell clone abundances (Gaimann et al. 2020). If Ω is sufficiently large, we would like to show under what conditions the expectations of our diversity measures converge quickly to Ω-independent values. By approximating $\sum_{i=1}^{\Omega}(1-p_i)^N \approx \sum_{i=1}^{\Omega} e^{-Np_i} \approx \int_1^{\Omega} e^{-N/(H_\nu(\Omega)z^\nu)}\mathrm{d}z$ in Eq. (6) we find in the large Ω limit

$$\begin{aligned}
\frac{\mathbb{E}[R(\nu=0)]}{N} &\approx x(1-e^{-1/x}), && x \equiv \Omega/N \\
\frac{\mathbb{E}[R(\nu=\frac{1}{2})]}{N} &\approx x\left[1-2\mathrm{E}\left(3,\tfrac{1}{2x}\right)\right], && x \equiv \Omega/N \\
\frac{\mathbb{E}[R(\nu=1)]}{N} &\approx 1-\frac{\log N}{\log\Omega}+\frac{\log(\log\Omega)}{\log\Omega}, \\
\frac{\mathbb{E}[R(\nu>1)]}{N^{1/\nu}} &\approx x\left[1-\tfrac{1}{\nu}\mathrm{E}\left(1+\tfrac{1}{\nu},\tfrac{x^{-\nu}}{\zeta(\nu)}\right)\right], && x \equiv \Omega/N^{1/\nu}
\end{aligned} \tag{47}$$

where the exponential integral is defined by $\mathrm{E}(x,y) \equiv \int_1^{\infty} t^{-x}e^{-yt}\mathrm{d}t$ and $\zeta(\nu)$ is the Riemann zeta function. Consistent with known biology and previous estimates (Zarnitsyna et al. 2013; Lythe et al. 2016), we take the large-Ω limit where $x > 1$. From Eqs. (47), we see that the expected richnesses converge to fixed values for large enough Ω and all values of $\nu \not\approx 1$. The rescaled expected richnesses are plotted as functions of $x = \Omega/N$ or $x = \Omega/N^{1/\nu}$ in Fig. 6a.

Analogous cutoff-insensitive results can be found for the variance $\mathrm{var}[R^2(\nu)]$ as well as other quantities. A good approximation for the variance is

$$\begin{aligned}
\mathrm{var}[R(\nu)] &= \mathbb{E}[R^2(\nu)] - \Big(\mathbb{E}[R(\nu)]\Big)^2 \\
&\approx \sum_{i=1}^{\Omega} e^{-p_i N}\left(1-e^{-p_i N}\right) \\
&\approx \tfrac{\Omega}{\nu}\left(\mathrm{E}\left(1+\tfrac{1}{\nu},\tfrac{N}{H_\nu(\Omega)\Omega^\nu}\right)-\mathrm{E}\left(1+\tfrac{1}{\nu},\tfrac{2}{H_\nu(\Omega)\Omega^\nu}\right)\right),
\end{aligned} \tag{48}$$

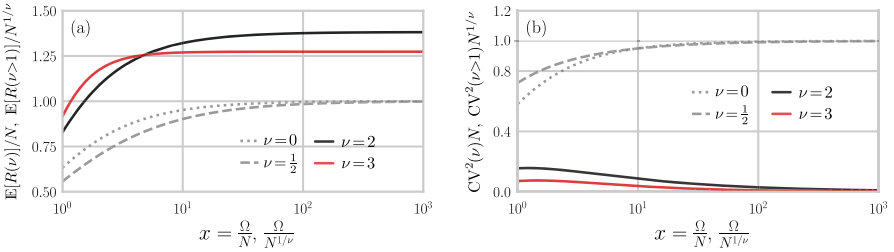

Fig. 6 Expected richness and uncertainty under power law-distributed probabilities p_i following Eq. (46). **a** Expected richness for different values of ν that lead to simple scaling and dependence only on $x = \Omega/N, \Omega/N^{1/\nu}$. For large x, the expected rescaled richnesses $\mathbb{E}[R(\nu)]/N$ and $\mathbb{E}[R(\nu > 0)]/N^{1/\nu}$ converge. Since the normalization of the expected richness (by $N^{1/\nu}$) for $\nu > 1$ is different than for $\nu = 0, \frac{1}{2}$, (normalized by N), $\mathbb{E}[R(\nu > 0)]/N^{1/\nu}$ converges to greater values, but $\lim_{x\to\infty} \mathbb{E}[R(\nu > 0)]$ remains < 1. **b** From the variances $\mathbb{E}[R^2(\nu)]$, we construct the squared coefficient of variation and plot $\mathrm{CV}^2 N \equiv N\mathrm{var}[R(\nu)]/(\mathbb{E}[R(\nu)])^2$ as a function of $x = \Omega/N$ (or $\mathrm{CV}^2(\nu > 1)N^{1/\nu}$ for as a function of $x = \Omega/N^{1/\nu}$). For large x, $\mathrm{CV}^2 \approx 1/N$ for $\nu = 0, 1/2$ but $\mathrm{CV}^2 N^{1/\nu} \sim 0$ for $\nu = 2, 3$.

where the power-law assumption in Eq. (46) is used in the last approximation. The normalized squared coefficients of variation $\mathrm{CV}^2 \equiv \mathrm{var}[R(\nu)]/(\mathbb{E}[R(\nu)])^2$ for representative ν are found to be

$$
\begin{aligned}
\mathrm{CV}^2_{\nu=0} N &\approx \frac{e^{-1/x}}{x(1-e^{-1/x})}, && x \equiv \Omega/N \\
\mathrm{CV}^2_{\nu=1/2} N &\approx \frac{2}{x}\frac{\mathrm{E}(3,\frac{1}{2x}) - \mathrm{E}(3,\frac{1}{x})}{\left(1-2\mathrm{E}(3,\frac{1}{2x})\right)^2}, && x \equiv \Omega/N \\
\mathrm{CV}^2_{\nu=1} N &\approx \frac{\log\Omega\left[\log(\Omega\log\Omega) - \log 4N\right]}{\left[\log(\Omega\log\Omega) - \log N\right]^2}, && \\
\mathrm{CV}^2_{\nu>1} N^{1/\nu} &\approx \frac{1}{\nu x}\frac{\mathrm{E}\left(1+\frac{1}{\nu},\frac{x^{-\nu}}{\zeta(\nu)}\right) - \mathrm{E}\left(1+\frac{1}{\nu},\frac{2x^{-\nu}}{\zeta(\nu)}\right)}{\left(1-\frac{1}{\nu}\mathrm{E}\left(1+\frac{1}{\nu},\frac{x^{-\nu}}{\zeta(\nu)}\right)\right)^2}, && x \equiv \Omega/N^{1/\nu}
\end{aligned}
\tag{49}
$$

Plots of the CV of the richness under power-law system probabilities are shown in Fig. 6b. We see that the squared CVs converge in the large x limit to N^{-1} for $\nu = 0, 1/2$ and vanish for $\nu > 1$.

The behavior of the sampled M-overlap, $\mathbb{E}[K_s^{(M)}(\{\mathbf{p}^{(m)}\})]$, can also be quantified under the power-law probability distribution. By using Eq. (20) and $\rho_i(S)$ (Eq. (37), assuming equal probabilities $p_i^{(m)} = p_i$ and sample sizes $S^{(m)} = S$ across individuals), we find

$$
\mathbb{E}\left[K_s^{(M)}(\{\mathbf{p}^{(m)}\})\right] = \sum_{i=1}^{\Omega}\prod_{m=1}^{M}\rho_i^{(m)}(S) \approx \sum_{i=1}^{\Omega}\left(1-e^{-p_i S}\right)^M. \tag{50}
$$

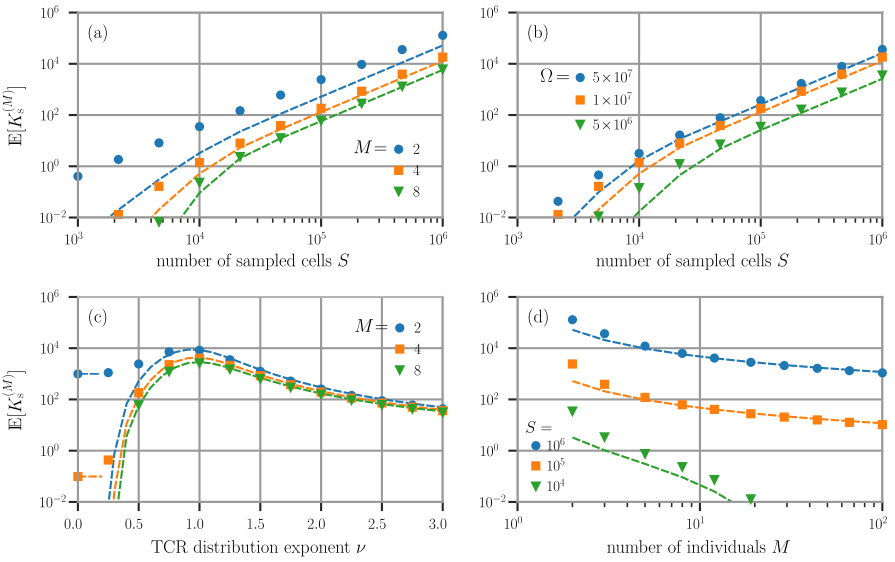

Fig. 7 The expected M-overlap $\mathbb{E}[K_s^{(M)}(\{\mathbf{p}^{(m)}\})]$. **a** Log-log plot of M-overlap as a function of individual sample size S using $p_i = i^{-1/2}/H_{1/2}(\Omega)$ ($\nu = 1/2$) and $\Omega = 10^7$. $M = 2, 4, 8$ are shown, with exponentially decreasing M-overlap as M is increased. **b** Fixing $M = 4$, a log-log plot of $\mathbb{E}[K_s^{(4)}(\nu = 1/2)]$ against S for different values of $\Omega = 5 \times 10^6, 10^7$ and 5×10^7. **c** $\mathbb{E}[K_s^{(M)}(\nu)]$ plotted against ν for fixed $\Omega = 10^7$ and different M. (d) With $\Omega = 10^7$, a log-log plot of $\mathbb{E}[K_s^{(M)}(\nu = 1/2)]$ as a function of the number M of individuals sampled with $S = 10^4, 10^5, 10^6$. In all panels, the dashed curves plot the analytic approximation for $\nu \gtrsim 0.7$ given in the second line of (51). In (c), the $\nu = 0$ limit matches the expression given by the first line in (51). The approximations given in (51) are especially accurate for large S and larger M and ν

This expression can be further simplified in the large Ω limit for specific ν,

$$\begin{aligned}
\mathbb{E}[K_s^{(M)}(\nu = 0)] &= \Omega\Big(1 - (1 - 1/\Omega)^S\Big)^M \approx \Omega\left(1 - e^{-S/\Omega}\right)^M, \\
\mathbb{E}[K_s^{(M)}(\nu \gtrsim 0.7)] &= \sum_{i=1}^{\Omega}\Big(1 - (1 - p_i)^S\Big)^M \approx \sum_{i=1}^{\Omega}\exp\Big[-Me^{-\frac{Si^{-\nu}}{H_\nu(\Omega)}}\Big] \\
&\sim \Big[1 - e^{-\frac{S}{H_\nu(\Omega)}}\Big]^M \left(\frac{S}{H_\nu(\Omega)\log M}\right)^{1/\nu}, \quad S, \frac{S}{H_\nu(\Omega)} \gg 1.
\end{aligned} \tag{51}$$

The last approximation is most accurate for $\nu > 1$ where $H_\nu(\Omega \to \infty)$ converges and the prefactor in brackets is ≈ 1. For sufficiently large S, it still provides a rough estimate of M-overlap for smaller values of ν. Asymptotic expressions for even smaller values of ν can be found in the $S/H_\nu(\Omega) \ll 1$ limit, but this limit yields very low expected M-overlap and is typically less informative.

Figure 7 plots the M-overlap $\mathbb{E}[K_s^{(M)}(\nu)]$ as a function of sample size, power-law ν, and M. For comparison, the analytic approximation for $\nu \gtrsim 0.7$ (51) is also plotted by the dashed curves. Equation (51) and plots such as those in Fig. 7a, b could

be useful for estimating the sample size S required in order to observe a specific overlap between the immune repertoires of M selected individuals. For instance, with $M = 4$ individuals, a repertoire size of $\Omega = 10^7$, and a sequence distribution exponent $\nu = 0.5$, an expected M-overlap of approximately 1 can be achieved with a sample size of $S = 10^4$.

Since the $(1/H_\nu(\Omega))^{1/\nu}$ term in Eq. (51) increases with ν, we expect that an effectively smaller repertoire size (recall $p_i \sim i^{-\nu}$ and larger ν leades to fewer larger-population clones), that the expected M-overlap increases with ν. However, the $(S/\log M)^{1/\nu}$ factor decreases with ν since larger S give rise to a larger number of ways clones sampled from different individuals can "avoid" each other. These features give rise to a maximum in $\mathbb{E}[K_s^{(M)}(\nu)]$, as shown in Fig. 7c.

Using Eqs. (33) and (38), we can also straightforwardly evaluate the variance of the M-overlap. For $\nu = 0$ and uniform $p_i = 1/\Omega$, $\text{var}[K_s^{(M)}(\nu = 0)] \approx \Omega(1-e^{-S/\Omega})^M \left(1 - (1 - e^{-S/\Omega})^M\right)$. We find the variances of the M-overlap, as with our other metrics, are well approximated by that of a binomial process in the $S \to \infty$ limit and when values of ν are modest: $\text{var}[K_s^{(M)}] \approx \Omega \frac{\mathbb{E}[K_s^{(M)}]}{\Omega}\left(1 - \frac{\mathbb{E}[K_s^{(M)}]}{\Omega}\right)$. Modest deviations from this approximation arise for finite S and large values of ν.

6 Sampling Resolution and Information Loss

We end with a brief discussion of information loss upon coarse-graining which arises when analyzing lower-dimensional experimental/biochemical classifications of clones that are commonly used. Such lower-dimensional representations can be obtained through spectratyping (Gorski et al. 1994; Fozza et al. 2017). For TCRs, spectratyping groups sequences together and produces compressed receptor representations describing CDR3 length, frequency, and associated beta variable (TRBV) genes (Gkazi et al. 2018). In addition to coarse-grained representations of sequencing data, some studies (Elhanati et al. 2018; Ruiz Ortega et al. 2023) use continuous approximations to describe the distribution of receptor sequences. Estimators of entropy and their errors have been developed for subsampling from discrete distributions (Schürmann 2004; Grassberger 2022). Therefore, in this section, we focus on quantifying differences in information content that are associated with (i) using continuous approximations of discrete sequencing data, and (ii) coarse-graining already-discretized (i.e., spectratyping) distributions.

Given a discrete random variable X describing Ω "traits" and taking on possible values $\{x_1, x_2, \ldots, x_\Omega\}$, let $p_i = \mathbb{P}(X = x_i)$. The entropy of this probability distribution is given by $H_p = -\sum_{i=1}^{\Omega} p_i \log p_i$. Similarly, one might define the differential entropy for a continuous random variable taking on values in the interval $[a, b]$ as $S_p = -\int_a^b p(x) \log p(x)\, \mathrm{d}x$. It is well-known that the differential entropy is not a suitable generalization of the entropy concept to continuous variables (Jaynes 1963) since it is not invariant under change of variables and can be negative. These issues can be circumvented by introducing the limiting density of discrete points. Here, we present a more direct approach that will be sufficient for our application. For a probability density function $p : [a, b] \to \mathbb{R}_0^+$ we introduce a discretizing morphism $\mathcal{D}_\Delta$

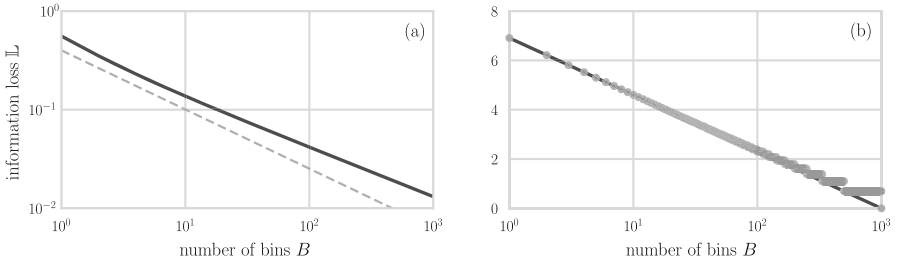

Fig. 8 The information loss $\mathbb{L}$ as a function of the number of discretization bins B. The loss is least as the number of integration bins $B \to \infty$. **a** The solid black line shows the information loss associated with discretizing a truncated power law (see Eq. (54)), and the dashed grey line is a guide-to-the-eye (power law) with slope -0.6. **b** Grey dots show the information loss associated with coarse graining a discrete and uniform random variable with initially $\Omega = 1000$ traits. The solid curve shows the corresponding analytical result for the difference in information $\mathbb{L} = -\log(B/\Omega)$ between discretizing a continuous uniform distribution of $\Omega = 1000$ traits using B bins

so that

$$q_i = \int_{a+(i-1)\Delta}^{a+i\Delta} p(x)\mathrm{d}x, \quad i = 1, 2, \ldots, B. \tag{52}$$

describes a random variable taking on values in each of the $(b-a)/\Delta = B$ bins.

To quantify the amount of information lost in this discretization step, consider the entropy $H_q = -\sum_{i=1}^{B} q_i \log q_i \sim \log \Delta$ in the $\Delta \to 0$ limit.

If we want to evaluate any information loss as a difference between the (finite) differential entropy S_p and the (diverging) entropy H_q we need to account for this logarithmic contribution by defining the corresponding information loss as

$$\mathbb{L}(\Delta) = |(S_p - \log \Delta) - H_q|. \tag{53}$$

By absorbing the logarithmic contribution into the differential entropy, we find the correct continuous entropy according to Jaynes (1963) using the limiting density of discrete points.

As an example, we compute the information loss associated with discretizing the truncated power law

$$p(x) = \begin{cases} \frac{1}{\gamma(\frac{1}{2},1)} \frac{e^{-x}}{\sqrt{x}}, & \text{if} \quad 0 \le x \le 1 \\ 0, & \text{else} \end{cases}, \tag{54}$$

where $\gamma(s, x) = \int_0^x t^{s-1} e^{-t}\, \mathrm{d}t$ is the lower incomplete gamma function. The distribution Eq. (54) gives rise to few high-abundance clones and many low-abundance clones, as typical for TCR receptor sequences (Xu et al. 2020). Analytic expressions for the discretized probabilities q_i are lengthy, so we numerically compute q_i to evaluate the information loss $\mathbb{L}(\Delta)$. Equation (53) is plotted as a function of the number of bins $B = 1/\Delta$ in Fig. 8a. The information loss decreases with the number of bins, as this

results in the discrete distribution gathering more information about its continuous counterpart.

While the connection between continuous probability distributions and their discretized counterparts has important consequences for sampling, information loss also occurs in spectratyping when an already discrete random variable is coarse grained. In this scenario, the information loss can be quantified uniquely (up to a global multiplicative constant) by the entropy difference of the distributions (Baez et al. 2011). The difference between the full $H_p = -\sum_{i=1}^{\Omega} p_i \log p_i$ and the coarse-grained $H_q = -\sum_{i=1}^{B} q_i \log q_i$ can be explicitly evaluated for uniformly distributed probabilities.

For any number $B < \Omega$ we can define a coarse graining procedure that yields only B traits by defining the bin size $\Delta = \text{ceil}(\Omega/B)$ and grouping together Δ traits into each bin. The last bin might be smaller than the other bins or even empty. The information loss of this procedure is shown in Fig. 8b for an initially uniform distribution of $\Omega = 1000$ traits. Across certain ranges of B, plateaus can build since our coarse graining might add zero probabilities. However, we can instead start from a continuous distribution and compare the discretization with $\Omega = 1000$ bins to any other binning with $B \leq \Omega$.

Comparing a coarse-grained uniform distribution with B bins to the discretized distribution with Ω bins yields the information loss with respect to the initial discrete distribution $\mathbb{L} = -\log(B/\Omega) \geq 0$. We plot this analytical prediction against the information loss $\mathbb{L}$ associated with coarse graining an already discrete distribution in Fig. 8b, showing them to be well-aligned.

7 Discussion and Conclusions

Quantifying properties of cell-type or sequence distributions is an important aspect of analyzing the immune repertoire in humans and animals. Different methods have been developed to estimate TCR and BCR diversity indices such as the total number of distinct sequences in an organism (i.e., species richness) (Rempala and Seweryn 2013; Kaplinsky and Arnaout 2016; Xu et al. 2020). Another quantity of interest is the number of clones that are considered "public" or "private," indicating how often certain TCR or BCR sequences occur across different individuals.

Public TCRβ and BCR sequences have been reported in a number of clinical studies (Putintseva et al. 2013; Robins et al. 2010; Shugay et al. 2013; Soto et al. 2020; Briney et al. 2019; Soto et al. 2019). However, the terms "public" and "private" clonotypes are often based on different and ambiguous definitions. According to Shugay et al. (2013), a "public sequence" is a sequence that is "*often* shared between individuals" (Shugay et al. 2013), while Greiff et al. (2017) refers to a sequence as public if it is "shared across individuals". In addition to ambiguities in the definition of what constitutes a private/public sequence, overlaps between the immune repertoires of different individuals are often reported without specifying confidence intervals, even though variations may be large given small sample sizes and heavy tailed sequence distributions.

In this work, we provided mathematical definitions for "public" and "private" clones in terms of the probabilities of observing a number of clones across M selected individuals, complementing related work that introduced the notion of "sharing number M" (i.e., the expected number of sequences which will be found in *exactly* M individuals) to quantify the expected overlap between cell-sequence samples (Elhanati et al. 2018; Ruiz Ortega et al. 2023). Besides defining individual repertoire probability distributions, our results include analytic expressions for individual and multi-individual expected richness and expected overlap as given by Eqs. (5), (10), (17), (19), and (20). Additionally, using Eqs. (28) to (33), we derived expressions for the expected richness and expected overlap in subsamples. The variability of quantities (second moments) such as the M-overlap and subsampled overlap were also derived. Studies analyzing the similarities and differences associated with immune repertoires of different individuals (see, e.g., Elhanati et al. 2018; Soto et al. 2020; Ruiz Ortega et al. 2023) may utilize our results on second moments of overlap measures to quantify the statistical significance of their findings. Our results are summarized in Table 1 where we provide expectations and second moments of all quantities as a function the cell population configurations $\mathbf{n}^{(m)}$ or as a function of the underlying clone generation probabilities $\mathbf{p}^{(m)}$, as is generated by models such as SONIA (Elhanati et al. 2014).

Further inference of richness and overlap given sample configurations can be developed using our results. For example, the parametric inference of expected richness in an individual given a sampled configuration $\mathbf{s}$ can be found using the multinomial model and Bayes' rule, as presented in Eq. (44).

While our results depend on knowledge of Ω and N, we show using power-law probability distributions and explicit expressions in Eqs. (47) and (49) that the richness is insensitive to Ω in the large N and Ω/N limits. Therefore, even though Ω may be impossible to accurately estimate, power-law probability distributions generally render our results robust to uncertainty in Ω. Analytic or semi-analytic expressions for the overlap quantities can also be derived. We leave this exercise to the reader.

Finally, in the context of coarse-graining, or spectratyping (Ciupe et al. 2013), we have discussed methods that are useful to quantify the information loss associated with different levels of coarse graining TCR and BCR sequences. The results presented here are based on an assumption of simple multinomial distributions as the underlying population model. A number of mechanistically more realistic probability distributions have been derived for neutral, noninteracting clone populations in steady state (Dessalles et al. 2018). These include log series and negative binomial distributions each requiring tailored calculations for the corresponding richness and overlap.

Acknowledgements LB received funding from the Swiss National Fund (P2EZP2_191888) and the Army Research Office (W911NF-23-1-0129). TC acknowledges funding from the NIH through grant R01HL146552 and the NSF through grant DMS-1814364.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability All source codes are publicly available at https://gitlab.com/ComputationalScience/immune_repertoires.

References

Abbas AK, Lichtman AH, Pillai S (2021) Cellular and molecular immunology, 10th edn. South Asia Edition, Elsevier Health Sciences, New Delhi

Alt FW, Oltz EM, Young F, Gorman J, Taccioli G, Chen J (1992) VDJ recombination. Immunol Today 13(8):306–314

Baez JC, Fritz T, Leinster T (2011) A characterization of entropy in terms of information loss. Entropy 13(11):1945–1957

Briney B, Inderbitzin A, Joyce C, Burton DR (2019) Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature 566(7744):393–397

Casrouge A, Beaudoing E, Dalle S, Pannetier C, Kanellopoulos J, Kourilsky P (2000) Size estimate of the $\alpha\beta$ TCR repertoire of naive mouse splenocytes. J Immunol 164(11):5782–5787

Chao A, Lee S-M (1992) Estimating the number of classes via sample coverage. J Am Stat Assoc 87:210–217

Chao A, Lin C-W (2012) Nonparametric lower bounds for species richness and shared species richness under sampling without replacement. Biometrics 68(3):912–921

Chao A, Kubota Y, Zelený D, Chiu C-H, Li C-F, Kusumoto B, Yasuhara M, Thorn S, Wei C-L, Costello MJ, Colwell RK (2020) Quantifying sample completeness and comparing diversities among assemblages. Ecol Res 35(2):292–314

Chen SX, Liu JS (1997) Statistical applications of the Poisson-Binomial and conditional Bernoulli distributions. Stat Sin 7:875–892

Chiu C-H, Wang Y-T, Walther BA, Chao A (2014) An improved nonparametric lower bound of species richness via a modified Good-Turing frequency formula. Biometrics 70:671–682

Ciupe SM, Devlin BH, Markert ML, Kepler TB (2013) Quantification of total T-cell receptor diversity by flow cytometry and spectra typing. BMC Immunol 14:35

Colwell RK, Chao A, Gotelli NJ, Lin S-Y, Mao C-X, Chazdon RL, Longino JT (2012) Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. J Plant Ecol 5:3–21

Davis MM, Bjorkman PJ (1988) T-cell antigen receptor genes and T-cell recognition. Nature 334(6181):395–402

Davodeau F, Peyrat MA, Romagne F, Necker A, Hallet MM, Vie H, Bonneville M (1995) Dual T cell receptor beta chain expression on human T lymphocytes. J Exp Med 181(4):1391–1398

Dessalles R, D'Orsogna M, Chou T (2018) Exact steady-state distributions of multispecies birth-death-immigration processes: effects of mutations and carrying capacity on diversity. J Stat Phys 173:182–221

Dessalles R, Pan Y, Xia M, Maestrini D, D'Orsogna MR, Chou T (2022) How naive T-cell clone counts are shaped by heterogeneous thymic output and homeostatic proliferation. Front Immunol 12:735135

DeWitt WS, Lindau P, Snyder TM, Sherwood AM, Vignali M, Carlson CS, Greenberg PD, Duerkopp N, Emerson RO, Robins HS (2016) A public database of memory and naive B-cell receptor sequences. PLoS ONE 11(8):0160853

Elhanati Y, Murugan A, Callan CG, Mora T, Walczak AM (2014) Quantifying selection in immune receptor repertoires. Proc Natl Acad Sci 111(27):9875–9880

Elhanati Y, Sethna Z, Marcou Q, Callan CG Jr, Mora T, Walczak AM (2015) Inferring processes underlying B-cell repertoire diversity. Philos Trans R Soc B: Biol Sci 370(1676):20140243

Elhanati Y, Sethna Z, Callan CG Jr, Mora T, Walczak AM (2018) Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. Immunol Rev 284(1):167–179

Fozza C, Barraqueddu F, Corda G, Contini S, Virdis P, Dore F, Bonfigli S, Longinotti M (2017) Study of the T-cell receptor repertoire by CDR3 spectra typing. J Immunol Methods 440:1–11

Gaimann M, Nguyen M, Desponds J, Mayer A (2020) Early life imprints the hierarchy of T cell clone sizes. eLife 9:e61639
Girardi M (2006) Immunosurveillance and immunoregulation by $\gamma\delta$ T cells. J Investig Dermatol 126(1):25–31
GitLab Repository (2022). https://gitlab.com/ComputationalScience/immune_repertoires
Gkazi AS, Margetts BK, Attenborough T, Mhaldien L, Standing JF, Oakes T, Heather JM, Booth J, Pasquet M, Chiesa R et al (2018) Clinical T cell receptor repertoire deep sequencing and analysis: an application to monitor immune reconstitution following cord blood transplantation. Front Immunol 2547
Gorski J, Yassai M, Zhu X, Kissela B, Keever C, Flomenberg N et al (1994) Circulating T cell repertoire complexity in normal individuals and bone marrow recipients analyzed by CDR3 size spectratyping. correlation with immune status. J Immunol 152(10):5109–5119
Gotelli NJ, Chao A (2013) Measuring and estimating species richness, species diversity, and biotic similarity from sampling data
Gotelli N, Colwell R (2011) Estimating species richness 12:39–54
Goyal S, Kim S, Chen ISY, Chou T (2015) Mechanisms of blood homeostasis: lineage tracking and a neutral model of cell populations in rhesus macaques. BMC Biol 13(1):85. https://doi.org/10.1186/s12915-015-0191-8
Grassberger P (2022) On generalized Schürmann entropy estimators. Entropy. https://doi.org/10.3390/e24050680
Greiff V, Weber CR, Palme J, Bodenhofer U, Miho E, Menzel U, Reddy ST (2017) Learning the high-dimensional immunogenomic features that predict public and private antibody repertoires. J Immunol 199(8):2985–2997
Hampton J, Lladser ME (2012) Estimation of distribution overlap of urn models. PLoS ONE 7(11):42368
Hong Y (2013) On computing the distribution function for the Poisson binomial distribution. Comput Stat Data Anal 59:41–51
Jaynes ET (1963) Information theory and statistical mechanics. Stat Phys 3:181
Kaplinsky J, Arnaout R (2016) Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples. Nat Commun 7(1):1–10
Larremore DB (2019) Bayes-optimal estimation of overlap between populations of fixed size. PLoS Comput Biol 15(3):1006898
Laydon DJ, Bangham CRM, Asquith B (2015) Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach. Philos Trans R Soc B: Biol Sci 370(1675):20140291
Lewkiewicz S, Chuang Y-L, Chou T (2019) A mathematical model of the effects of aging on naive T-cell populations and diversity. Bull Math Biol 81:2783–2817
Lythe G, Callard RE, Hoare RL, Molina-París C (2016) How many TCR clonotypes does a body maintain? J Theor Biol 389:214–224
Murugan A, Mora T, Walczak AM, Callan CG Jr (2012) Statistical inference of the generation probability of T-cell receptors from sequence repertoires. Proc Natl Acad Sci U S A 109(40):16161–16166
Padovan E, Giachino C, Cella M, Valitutti S, Acuto O, Lanzavecchia A (1995) Normal T lymphocytes can express two different T cell receptor beta chains: implications for the mechanism of allelic exclusion. J Exp Med 181(4):1587–1591
Putintseva EV, Britanova OV, Staroverov DB, Merzlyak EM, Turchaninova MA, Shugay M, Bolotin DA, Pogorelyy MV, Mamedov IZ, Bobrynina V et al (2013) Mother and child T cell receptor repertoires: deep profiling study. Front Immunol 4:463
Rempala GA, Seweryn M (2013) Methods for diversity and overlap analysis in T-cell receptor populations. J Math Biol 67(6–7):1339–1368
Robins HS, Srivastava SK, Campregher PV, Turtle CJ, Andriesen J, Riddell SR, Carlson CS, Warren EH (2010) Overlap and effective size of the human CD8+ T cell receptor repertoire. Sci Transl Med 2(47):47–644764
Ruiz Ortega M, Spisak N, Mora T, Walczak AM (2023) Modeling and predicting the overlap of B-and T-cell receptor repertoires in healthy and SARS-CoV-2 infected individuals. PLoS Genet 19(2):1010652
Rybakin V, Westernberg L, Fu G, Kim H-O, Ampudia J, Sauer K, Gascoigne NRJ (2014) Allelic exclusion of TCR α-chains upon severe restriction of Vα repertoire. PLoS ONE 9(12):114320
Schuldt NJ, Binstadt BA (2019) Dual TCR T cells: identity crisis or multitaskers? J Immunol 202(3):637–644
Schürmann T (2004) Bias analysis in entropy estimation. J Phys A: Math Gen 37(27):295

Shugay M, Bolotin DA, Putintseva EV, Pogorelyy MV, Mamedov IZ, Chudakov DM (2013) Huge overlap of individual TCR beta repertoires. Front Immunol 4:466

Slabodkin A, Chernigovskaya M, Mikocziova I, Akbar R, Scheffer L, Pavlović M, Bashour H, Snapkov I, Mehta BB, Weber CR et al (2021) Individualized VDJ recombination predisposes the available Ig sequence space. Genome Res 31(12):2209–2224

Soto C, Bombardi RG, Branchizio A, Kose N, Matta P, Sevy AM, Sinkovits RS, Gilchuk P, Finn JA, Crowe JE (2019) High frequency of shared clonotypes in human B cell receptor repertoires. Nature 566(7744):398–402

Soto C, Bombardi RG, Kozhevnikov M, Sinkovits RS, Chen EC, Branchizio A, Kose N, Day SB, Pilkinton M, Gujral M et al (2020) High frequency of shared clonotypes in human T cell receptor repertoires. Cell Rep 32(2):107882

Travers P, Walport M, Shlomchik MJ, Janeway MC (1997) Immunobiology: the immune system in health and disease. Churchill Livingstone, London

Tussiwand R, Bosco N, Ceredig R, Rolink AG (2009) Tolerance checkpoints in B-cell development: Johnny B good. Eur J Immunol 39(9):2317–2324

Venturi V, Price DA, Douek DC, Davenport MP (2008) The molecular basis for public T-cell responses? Nat Rev Immunol 8(3):231–238

Wang JPZ, Lindsay BG (2005) A penalized nonparametric maximum likelihood approach to species richness estimation. J Am Stat Assoc 100:942–959

Xu S, Böttcher L, Chou T (2020) Diversity in biology: definitions, quantification and models. Phys Biol 17(3):031001

Yates A (2014) Theories and quantification of thymic selection. Front Immunol 5:13

Zarnitsyna V, Evavold B, Schoettle L, Blattman J, Antia R (2013) Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire. Front Immunol 4:485