Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Systems Biology

How many different clonotypes do immune repertoires contain?

Thierry Mora and Aleksandra M. Walczak

Abstract

Immune repertoires rely on diversity of T-cell receptors and immunoglobulins to protect us against foreign threats. The ability to recognize a wide variety of pathogens is linked to the number of different clonotypes expressed by an individual. Out of the estimated ~10^{12} different B and T cells in humans, how many of them express distinct receptors? We review current and past estimates for these numbers. We point out a fundamental limitation of current methods, which ignore the tail of small clones in the distribution of clone sizes. We show that this tail strongly affects the total number of clones, but it is impractical to access experimentally. We propose that combining statistical models with mechanistic models of lymphocyte clonal dynamics offers possible new strategies for estimating the number of clones.

Addresses

Laboratoire de physique de l'École normale supérieure (PSL University), CNRS, Sorbonne Université, Université de Paris, 75005 Paris, France

Corresponding authors: Mora, Thierry (thierry.mora@ens.fr); Walczak, Aleksandra M (awalczak@lpt.ens.fr)

Current Opinion in Systems Biology 2019, **18:**104–110

This review comes from a themed issue on **Systems immunology & host-pathogen interaction (2019)**

Edited by **Thomas Höfer and Grégoire Altan-Bonnet**

For a complete overview see the Issue and the Editorial

Available online 2 November 2019

https://doi.org/10.1016/j.coisb.2019.10.001

Keywords

T-cell receptor, B-cell receptor, Immune repertoires, Repertoire sequencing, Repertoire diversity, Species richness, Clone size distribution.

Introduction

The diversity of immune repertoires plays an important role in the host's ability to recognize and control a wide range of pathogens. While actual recognition of an antigen depends on having a relatively specific T-cell receptor (TCR) or immunoglobulin (Ig), multiple experimental examples show that reduced receptor diversity may limit the efficacy of adaptive immune repertoires [1,2]. This effect becomes especially pronounced in individuals infected with cytomegalovirus [3–5], or with immunosenescence [6], when naive clonotypes are significantly reduced, and the organism is left to rely on reduced immune diversity.

The number of different clonotypes, called species richness in ecology, is thus an important quantity to estimate, both biologically and clinically. Here we review some of the experimental and theoretical approaches that have been used to estimate the number of distinct clonotypes in TCR and Ig repertoires, in naive, memory, or unfractionned repertoires. Based on existing repertoire data and computational models, we demonstrate that no statistical method can overcome the limitations of small sampling. We argue that this problem, which is inherent to all existing methods, could be overcome by combining repertoire data with stochastic models of lymphocyte population dynamics, taking into account the caveats of convergent recombination and experimental noise.

Past estimates

The variable part of each TCR beta (and Ig heavy) chain is composed by putting together variable (V), diverse (D) and joining (J) segments, which are encoded in the germline genome. TCR alpha (and Ig light) chains only have V and J genes and no D genes. Additionally each chain experiences additions and deletions of nucleotides at the gene junctions, which increases the diversity. This junctional rearrangements have been identified as the main contributor to sequence diversity [7]. The alpha and beta chain generations in TCR are separated in time, and have been shown to be independent [8,9]. The total number of different alpha-beta pairs that the generation machinery can produce is *much* greater than the total number of receptors in the whole human population [8,10,11]. Thus, each person harbors only a small fraction of the potential diversity of receptors. To estimate the number of distinct receptors, we need to count them.

An early quantitative direct estimate of the size of the TCR repertoire dates back to Arstila et al. [12], following earlier considerations [7,13]. Their approach was to focus on a subclass of receptors (either alpha or

beta) with a specific V-J class and length, and sequence them using low-throughput methods. The number of different sequences in that subclass is then extrapolated back to the get the full diversity by dividing by the known frequency of V-J and length usage. The authors were very careful to verify their estimate in different V-J classes and donors, and to account for rare clonotypes that might have escaped sequencing. A total of $\sim 10^6$ different TCR beta chains were thus estimated in a sample of 10^8 T-cells. Each beta chain paired with 25 different alpha chains, resulting in $\sim 25 \cdot 10^6$ distinct TCR alpha-beta. Much smaller TCR alpha-beta diversity was reported in the memory subset, $\sim 2 \cdot 10^5$, consistent with the idea that memory cells form a selected and thus restricted subset.

With the onset of high-throughput sequencing of immune receptor repertoires [14–22] came the realization of the importance of the sampling problem. In sequenced repertoires, many clonotypes are seen just once, suggesting that there are possibly many more that have similar or slightly smaller sizes but were not sequenced, simply by chance. This issue does not only affect "small" clones, where a clone comprise cells with the same TCR or Ig clonotype. A clone of 10^5 cells among a total of $3 \cdot 10^{11}$ T cells will often not be seen even once in a typical sample of 10^6 cells. To deal with this issue, a commonly adopted approach has been to use statistical estimators (see Ref. [23] for a overview).

Using the Poisson abundance statistical method [24], Robins et al. [15] obtained estimates of $\sim 10^6$ TCR beta nucleotide clonotypes for CD8 and CD4 naive cells, and $\sim 5 \cdot 10^5$-10^6 for CD8 and CD4 memory TCR beta. Qi et al. [6] used another method called the Chao2 estimator [25], which uses multiple replicates of the sequencing experiment, to estimate TCR species richness. They obtained much larger estimates than Robins et al.: $\sim 2 \cdot 10^8$ TCR beta nucleotide clonotypes in CD4 and CD8 naive repertoires, $\sim 1.5 \cdot 10^6$ clonotypes in CD4 memory repertoires, and about 5–10 times fewer in CD8 memory. All those numbers decreased with age. Using the Poisson abundance method, diversities of $\sim$ 1-$2 \cdot 10^9$ for naive and $\sim 5 \cdot 10^7$-10^8 for memory Ig heavy chains were reported [26]. Recent estimates of Ig heavy-chain species richness using an advanced statistical estimator [27] yielded smaller diversities, ranging from $\sim 10^7$ [28] to 10^7-10^9 [29], presumably because they focused on amino acid rather than nucleotide clonotypes and ignored hypermutations in the V and J segments.

The sampling problem

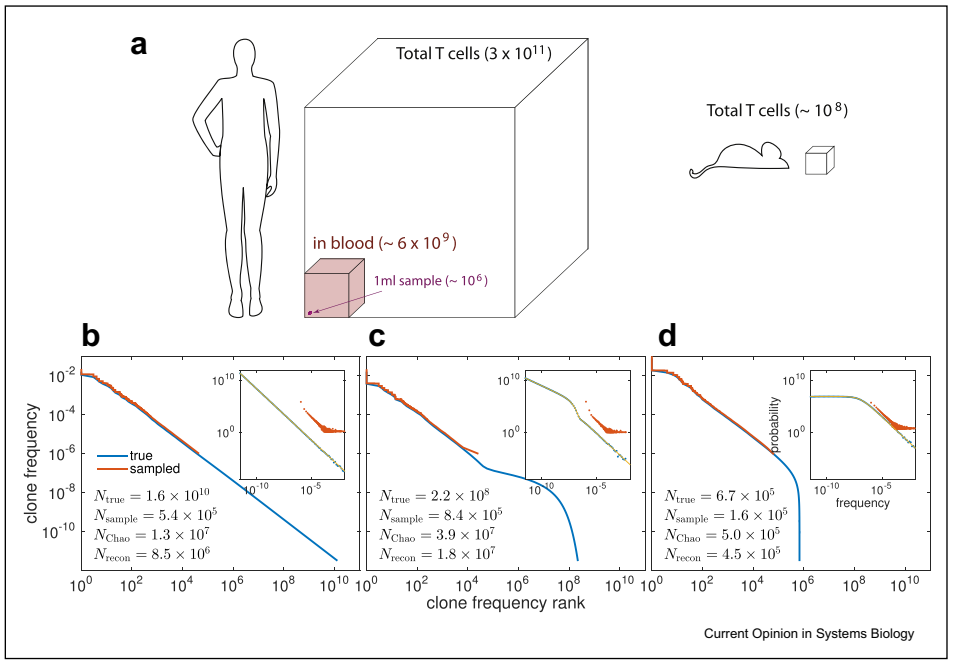
The approaches described above share the common problem that it is impossible to extrapolate what happens for small clones from small samples, which capture the largest clones [23,27]. Getting information on the small clones is in fact impractical: it would require sequencing essentially all lymphocytes in an organism. Humans harbor of the order of $3 \cdot 10^{11}$ T cells (and roughly the same order of B cells). Of these, only a few percents are contained in blood, of which a small fraction ($\sim 10^6$) is sampled in typical experiments (Figure 1A). In addition, blood may not be representative of other tissues. Even in mice, which contain fewer lymphocytes ($\sim 10^8$ T cells) and can be sacrificed to isolate all the body's lymphocytes, cell loss during the experiment hampers this approach.

In Figure 1B–D, we illustrate with simulations what happens when one analyses samples of 10^6 cells from three synthetic repertoires. These repertoires are described by different clone size distributions, corresponding to a widely different number of clones: a pure power law (Figure 1B), a mixture of a power law and neutral model [30] (Figure 1C), and a power law with a low-frequency cut-off (Figure 1D). Their species richness are widely different, ranging from $N \sim 7 \cdot 10^5$ to $1.6 \cdot 10^{10}$. Yet, the sampled repertoires show similar clone size distributions, and comparable observed diversity (10^5–10^6), because they behave similarly for large clone sizes, but drastically differ in the tail of small clone sizes.

Any statistical method that extrapolates from observations assumes, knowingly or implicitly, an underlying model for how the clone size distribution behaves for the smallest clones. The Poisson abundance and Chao estimators [24,25,31] discussed earlier both assume a well peaked distribution of clone sizes, which is not the case in our examples. As a result, the Chao1 [31] estimator can underestimate species richness by up to a 1000-fold factor (Figure 1B). A more advanced and increasingly used estimator such as Recon [27] gives similar estimates, as can be expected because Recon is not meant to capture small clones whose frequency lies below the sampling limit.

Real repertoires are likely affected by this problem. The sampled (large clone) part of their clone size distribution has been shown to follow a power law both for TCR and Ig [10]. Naive subsets display shorter tails of large clones [32], suggesting that the power-law behaviour in unfractioned repertoires is dominated by memory clones. Our three synthetic examples are consistent with power laws for large clones, but differ greatly for small clones, yielding very different species richness. Extrapolating the distribution of clone sizes is the key idea behind DivE [33] and Recon [27], which were proposed to estimate diversity in TCR subsets. However, these approaches work when the behaviour at large clones is informative for small clones, which may not always be true.

Figure 1



Estimating the total number of B or T cell clonotypes from small samples is generally impossible. **a.** Orders of magnitude for the number of T cells. Only one or a few percent of all $T \sim 3 \cdot 10^{11}$ T cells circulate in blood at any given time. Among these, typical sequenced samples contain about a million cells, which is a tiny fraction of the total repertoire. The number of T cells in a mouse is shown for comparison. Similar numbers hold for B cells. **b.-d.** Rank–frequency plots of three synthetic repertoires showing the frequency of B- or T-cell clones *versus* their rank (from most frequent to least frequent). The corresponding clone size distribution is shown in the inset: **b)** power-law distribution; **c)** mixture of power-distribution and neutral distribution (see main text); **d)** power-law distribution with low-frequency cutoff. A random of sample of 10^6 cells (in red) fails to capture most of the true rank–frequency relation (in blue). While the sampled distribution looks similar in all three cases, the true distributions are very different in the domain of low clonal frequencies, and correspond to a widely different number of clones N_{true}. That number is very poorly estimated by the number of sampled clones (N_{sample}), or by statistical estimators such as Chao1 (N_{Chao}) [31], or Recon (N_{recon}) [27].

Proposed solution: stochastic modeling

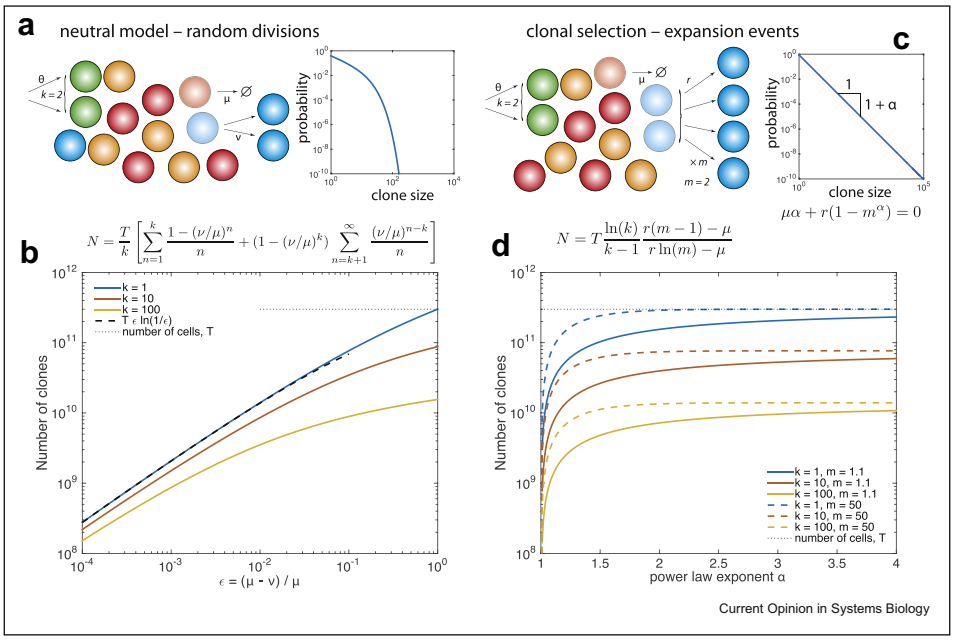
To access small clones that cannot be directly probed experimentally, we need to explicitly model the biological processes that shape these distributions, without having to take a leap of faith. Unlike extrapolation, such models might predict behaviours for small clones that are quantitatively different than the trend suggested by large clones. Of course, model assumptions should be tested experimentally, their parameters estimated from measurements, and confidence intervals put on their predictions. We now briefly review two simple models that have been proposed to describe the dynamics and clone size distributions of naive and memory repertoires.

Cells in naive repertoires have not experienced strong proliferation due to antigen recognition. Nevertheless not all clones are of the same size, in part because clones leave the thymus with different initial sizes, and in part because they undergo stochastic division and death. The simplest model of naive repertoires is Hubbell's neutral model of ecology [30], which assumes constant division and death rates ($\nu < \mu$) for each cell, with new clones introduced with rate θ and constant initial size k (Figure 2A, left). More complex variants of that model may include intrinsic fitness differences between clones or cells, e.g. through competition for self-antigens [34] or cytokines [35].

Under that simple model, the steady state distribution of clones can be computed analytically [35–37], and falls off exponentially for clones larger than k (Figure 2A, right), meaning that large clones are rare. The total number of clones N can also be calculated analytically as a function of the model parameters μ, ν, and k, as well as the total number of cells T (Figure 2C). Unless cell division almost exactly balances death ($\nu \sim \mu$), or the introduction clone size k is large, the typical naive clone size is fairly small. This means that the total number of clones is very large, and comparable to the total number of cells. To get a more precise estimate would require to measure the division rate of naive T cells ν, and initial clone size k.

A limitation of this approach is the assumption that the clone size distribution quickly reaches a steady state. Naive T cells are very long-lived $\mu^{-1} \sim 3$ years [38], and

Figure 2



Lymphocyte population dynamics models can be used to estimate the number of clones. a. Neutral model for lymphocyte dynamics. New clones come out of the thymus (for T cells) or bone marrow (for B cells) with rate θ, with initial clone size k. Then each cell may divide with rate ν, and die with rate $\mu > \nu$. The clone size distribution at steady state can be calculate and falls off rapidly (right). **b.** A minimal clonal selection model. Instead of dividing randomly, cells of the same clone all proliferate m-fold upon immune stimulation, which occurs with rate r. The clone size distribution of this process behaves as a power law for large clones. The exponent of the power law can be expressed as a function of the model parameters. **c.-d.** The total number of clones can be expressed as a function of the model parameters for **c)** the neutral model **d)** the clonal selection model, and the total number of cells in the body, $T = 3 \cdot 10^{11}$. In the neutral model, the typical size of clones increases and diverges when division and death balance each other, $\mu \sim \nu$, leading to reduced diversity for a fixed number of cells. In the selection model, a similar divergence is observed as the power law exponent α gets close to 1.

the size of the naive pool changes with age, so that steady state may never be reached. Transient models of naive repertoires remain to be explored in more detail. A recent experimental study suggests that some T-cell naive clones are much larger than predicted by the neutral theory [36]. However the origin of these outliers is not yet well understood, and may have to do with the inadequacy of our current definitions of naive and memory cells through surface markers.

Modeling memory repertoires requires taking into account the expansion and then contraction dynamics after an infection. These dynamics are driven by new pathogens that infect the host, are recognized and then cleared [39], which leads to a constantly changing antigenic landscape [35]. This random encounter with antigens can be simply modeled by bursts of division events for all cells of the same clone, with rate r, causing each cell to effectively multiply m times into memory following antigen clearance (Figure 2B, left). Again, this model can be solved exactly in the continuous limit at steady state. The clone size distribution follows a power law for large clones (Figure 2B, right). The predicted number of clones N depends critically on the power law exponent α, and can be calculated as a function of the model parameters (Figure 2D). N drops to zero for power law exponent α close to 1. Interestingly, measured exponents α from unfractioned T cell beta chain repertoires (whose large-clone tail is believed to be dominated by memory clones) range from 1 to $\sim$1.5 [10]. This high sensitivity to parameters makes estimates from data very difficult.

Extensions of this model include the emergence of antigenic "niches" [40], where clonal expansion is limited by antigen availability, leading to diminishing returns upon multiple stimulation events [41]. Such mechanisms would limit the size of the largest clones and would cut off the power law behaviour, which is not observed in data.

Connecting models to data

Several caveats and corrections must be taken into account when linking stochastic models of population dynamics such as discussed above to repertoire data. Of importances are the issues of convergent recombination and experimental noise.

Population dynamics models focus on clones, defined as the set of cells originating from a common recombination event. However, two recombination events can lead

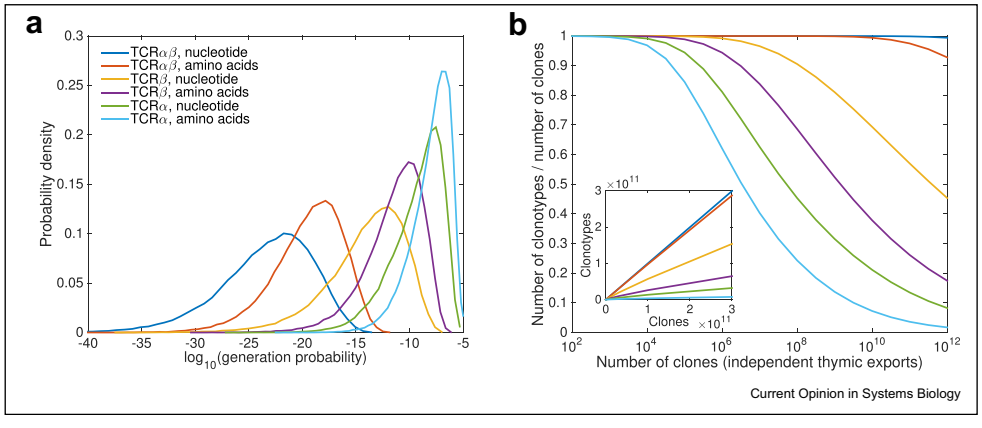
to exactly the same sequence. The two corresponding clones would be indistinguishable, and form a single clonotype in the repertoire. This effect can be corrected for by using models of recombination [42,43]. These models, which are inferred from data, can predict the distribution of generation probabilities of full receptors, or of single chains, both at the level of nucleotide or amino acid sequences, as shown in Figure 3A for TCR. From this distribution, the probability of convergent recombination can be computed to predict the number of distinct clonotypes as a function of the number of "clones", defined as independent recombination events (Eq. (7) of [11]). In Figure 3B we plot that prediction for the alpha, beta, and alpha-beta TCR in humans, for both nucleotide and amino acid clonotypes. These computations show that, for the full TCR alpha-beta clonotypes, convergent recombination is so rare that it hardly affects species richness. For the alpha and beta chains alone, however, the effect is substantial. Since most repertoire data are of single chains, this correction should be applied when linking data to the type of models discussed above, as was done in Ref. [36].

An additionnal issue complicates the comparison of models to data: experimental noise in the observed frequencies of clonotypes. In practice the number of reads (or unique molecular identifiers when they are used) n observed in data for a given clonotype is not simply the result of random sampling, and is not distributed according to a Poisson law, as was assumed in Figure 1 and in all previous work on diversity estimation. Instead, noise is over-dispersed, due to additional noise caused by DNA amplification and library preparation prior to sequencing. This noise model can be fitted using replicates of the repertoire sequencing experiment. This inference is impossible to separate from the inference of the clone size distribution $\rho(f)$. The two must thus be learned simultaneously from replicates by maximizing the likelihood of observed abundances, which depend on both the clone size distribution and the noise properties [44]. Applying this approach to the beta chain of unpartioned T cells with $\rho(f) \propto f^{-1-\alpha}$, yields species richness of $N \sim 10^8 - 10^9$, with power-law exponent $\alpha \approx 1\text{-}1.2$ [44]. In this estimate, the power law is taken as a given, and not linked to a model of clonal dynamics. A full mechanistic model treatment combined with the statistical model remains an interesting direction to explore, which could help shed light on the differences between memory and naive repertoires.

Sampling and repertoire sharing

Several recent papers have focused on shared immune receptors from high throughput Ig repertoire data [28,29,45]. These high profile analyses report absolute percentages of shared clonotypes. However, it was shown in the context of TCR (but the same holds for Ig) that these fractions are not absolute properties of the repertoires, but rather depend on sampling depth and the number of individuals that share the clonotypes [10,11]. Sharing estimates based on samples of the repertoire are bound to grossly underestimate the true sharing fraction. Therefore, reporting sharing percentages without appropriate information about sample sizes is meaningless. Fortunately a new standard for reporting repertoire sequencing data [46] contains this information.

Figure 3



Convergent recombination. a. Distributions of generation probabilities for the alpha and beta chains as well as alpha-beta pairs of T-cell receptors, for both nucleotide and amino-acid sequences, calculated using the OLGA software [43]. Most clonotypes have very low probability and are therefore unlikely to occur in two clones independently. High-probability clonotypes, however, will be generated several times in distinct T-cell clones (convergent recombination), reducing their diversity. **b.** Impact of convergent recombination on clonotype diversity. The ratio of the number of clonotypes to the number of clones is calculated using a model of recombination using OLGA, with the additional assumption that a random fraction q of recombination events fail to pass selection [11]. This ratio decreases as the number of clones increases, as redundant recombination events become more likely. The magnitude of this effect depends on the definition of clonotype (single chains or alpha-beta pairs, amino acids or nucleotides). It is small for full alpha-beta pairs. Inset: number of distinct clonotypes as a function of the number of clones. Selection parameter q: $q_\alpha = 0.046$, $q_\beta = 0.0091$, and $q_{\alpha\beta} = q_\alpha q_\beta$ (taken from Refs. [8,11]).

To assess the true overlap between the repertoires of two or more individuals, we would need to sequence all their lymphocytes, which is impractical. However, statistical model of sequence probabilities can be used to extrapolate sharing estimates to the full repertoire size N, provided that number is known [11]. For instance, clonotypes whose probability p is larger than $1/N$ are expected to be present in $1-e^{-pN} > 63\%$ of individuals, and can be considered "public". Recombination models such as the one of Figure 3A can be used to estimate the fraction of clonotypes that are public. For example, for $N = 10^{10}$, the model predicts that about 15% of TCR beta amino acid clonotypes expressed by human individuals are public [11].

Conclusions

While we have focused our review on the number of distinct clonotypes, what really matters for biological function is the number of different specificities. Due to cross-reactivity, each receptor can recognize many antigens and each antigen can be recognized by many receptors with different strengths. To account for this degeneracy, we would need to define a functional coverage of the antigenic space [47]. However, we currently do not have a comprehensive sequence-to-function maps for TCR and Ig that would allow us to estimate such a quantity.

Simply counting clonotypes also ignores their relative abundances. Clonotypes expressed by very few cells may not be as relevant for immune protection as very frequent clonotypes. Other diversity measures such as Hill numbers account for differences in frequencies [10,48]. Some of these measures are in fact more robust than species richness, because they put more focus on large clones and are less susceptible to sampling noise. Depending on the question, these measures may be better suited than species richness.

Our discussion has focused mostly on T cells, and has ignored the complications of hypermutations in Ig, which cause lineages to split into many clonotypes. Whether diversity is defined at the level of lineages or clonotypes will lead to different answers [26,28,29]. Developping specialized population dynamics models of B cell development and affinity maturation that include hypermutations is an interesting research direction.

We emphasized that estimating species richness cannot be disentangled from estimating the full distribution of clone sizes. As we gain insight into various aspects of lymphocyte dynamics, from thymic output to infection and memory formation, better mathematical descriptions can be leveraged to propose refined forms for the clone size distribution, and to fit their parameters to observations. Only with such a combination of modeling and data will we be able to get a better picture of repertoire diversity and immune coverage.

Conflict of interest statement

Nothing declared.

Acknowledgements

This work was partly supported by ERC CoG 724208.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest

1. Yager EJ, *et al.*: **Age-associated decline in T cell repertoire diversity leads to holes in the repertoire and impaired immunity to influenza virus**. *J Exp Med* 2008, **205**:711–723.

2. Muraro P, Robins H: **T cell repertoire following autologous stem cell transplantation for multiple sclerosis**. *J Clin Investig* 2014, **124**:1168–1172.

3. Khan N, *et al.*: **Cytomegalovirus seropositivity drives the CD8 T cell repertoire toward greater clonality in healthy elderly individuals**. *J Immunol* 2002, **169**:1984–1992.

4. Souquette A, Frere J, Smithey M, Sauce D, Thomas PG: **A constant companion: immune recognition and response to cytomegalovirus with aging and implications for immune fitness**. *GeroScience* 2017, **39**:293–303.

5. Smithey MJ, Venturi V, Davenport MP, Buntzman AS, Vincent BG: **Lifelong CMV infection improves immune defense in old mice by broadening the mobilized TCR repertoire against third-party infection**. *Proc Natl Acad Sci* 2018, **115**.

6. ** Qi Q, *et al.*: **Diversity and clonal selection in the human T-cell repertoire**. *Proc Natl Acad Sci U.S.A* 2014, **111**:13139–13144.
The highest experimental estimate for TCR diversity in humans to date, using high-throughput sequencing on replicates and the Chao2 estimator.

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

8. Dupic T, Marcou Q, Walczak AM, Mora T: **Genesis of the αβ T-cell receptor**. *PLoS Comput Biol* 2019, **15**, e1006874.

9. Grigaityte K, *et al.*: **Single-cell sequencing reveals αβ chain pairing shapes the T cell repertoire**. *bioRxiv* 2017:213462.

10. Mora T, Walczak AM. In *Systems immunology: an introduction to modeling methods for scientists*. Edited by Das J, Jayaprakash C, CRC Press; 2018.

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

12. * Arstila TP, *et al.*: **A direct estimate of the human alphabeta T cell receptor diversity**. *Science* 1999, **286**:958–961.
Early low-throughput estimate of T-cell receptor diversity in humans. This study identified many of the problems in estimating the total number of different clonotypes: limited sampling, repertoire heterogeneity, distinction between potential and expressed repertoires.

13. Davis MM, *et al.*: **Ligand recognition by alpha beta T cell receptors**. *Annu Rev Immunol* 1998, **16**:523–544.

14. Weinstein JA, Jiang N, White RA, Fisher DS, Quake SR: **High-throughput sequencing of the zebrafish antibody repertoire**. *Science* 2009, **324**:807–810.

15. * Robins HS, *et al.*: **Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells**. *Blood* 2009, **114**:4099–4107.

Early estimate of TCR diversity from high-throughput repertoire sequencing, using a statistical estimator to account for unseen species.

16. Boyd SD, *et al.*: **Measurement and clinical monitoring of human lymphocyte clonality by massively parallel {VDJ} pyrosequencing**. *Sci Transl Med* 2009, **1**:12ra23.

17. Six A, *et al.*: **The past, present and future of immune repertoire biology - the rise of next-generation repertoire analysis**. *Front Immunol* 2013, **4**:413.

18. Robins H: **Immunosequencing: applications of immune repertoire deep sequencing**. *Curr Opin Immunol* 2013, **25**:646–652.

19. Georgiou G, *et al.*: **The promise and challenge of high-throughput sequencing of the antibody repertoire**. *Nat Biotechnol* 2014, **32**:158–168.

20. Heather JM, Ismail M, Oakes T, Chain B: **High-throughput sequencing of the T-cell receptor repertoire: pitfalls and opportunities**. *Briefings Bioinf* 2017, **19**:554–565.

21. Minervina A, Pogorelyy M, Mamedov I: **TCR and BCR repertoire profiling in adaptive immunity**. *Transpl Int* 2019. 0–2.

22. Bradley P, Thomas PG: **Using T cell receptor repertoires to understand the principles of adaptive immune recognition**. *Annu Rev Immunol* 2019, **37**:547–570.

23. ** Laydon DJ, Bangham CRM, Asquith B, Crm B: **Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach**. *Philos Trans R Soc Lond B Biol Sci* 2015, **370**: 20140291.
An comprehensive review of existing and recently proposed statistical methods for estimating TCR diversity from samples.

24. Fisher R, Steven Corbet A, Williams C: **The relation between the number of species and the number of individuals in a random sample of an animal population**. *J Anim Ecol* 2016, **12**:42–58.

25. Chao A, Bunge J: **Estimating the number of species in a stochastic abundance model**. *Biometrics* 2002, **58**:531–539.

26. DeWitt WS, *et al.*: **A public database of memory and naive B-cell receptor sequences**. *PLoS One* 2016, **11**, e0160853.

27. * Kaplinsky J, Arnaout R: **Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples**. *Nat Commun* 2016, **7**:1–10.
Advanced statistical method for estimating diversity by extrapolating the clone-size distribution.

28. Soto C, *et al.*: **High frequency of shared clonotypes in human B cell receptor repertoires**. *Nature* 2019, **566**:398–402.

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

30. Hubbell SP: *The unified neutral theory of biodiversity and biogeography*. Princeton: Princeton University Press; 2001.

31. Chao A: **Nonparametric estimation of the number of classes in a population**. *Scand J Stat* 1984, **11**:265–270.

32. Oakes T, *et al.*: **Quantitative characterization of the t cell receptor repertoire of naïve and memory subsets using an integrated experimental and computational pipeline which is robust, economical, and versatile**. *Front Genet* 2017, **8**:1–17.

33. Laydon DJ, *et al.*: **Quantification of HTLV-1 clonality and TCR diversity**. *PLoS Comput Biol* 2014, **10**:1–13.

34. Lythe G, Callard RE, Hoare RL, Molina-parís C: **How many TCR clonotypes does a body maintain ?** *J Theor Biol* 2016, **389**: 214–224.

35. * Desponds J, Mora T, Aleksandra W: **Fluctuating fitness shapes the clone size distribution of immune repertoires**. *Proc Natl Acad Sci USA* 2016, **113**:274.
This study introduces several models of clonal lymphocyte population dynamics, including the neutral model, and models of fluctuating antigenic environments.

36. * Greef PCD, *et al.*: **The naive T-cell receptor repertoire has an extremely broad distribution of clone sizes**. *bioRxiv* 2019: 691501.
Thorough experimental and theoretical investigation of the clone size distribution of naive TCR. The study combines the neutral model of clonal populations with model of recombination to explain observations.

37. ** Altan-Bonnet G, Mora T, Walczak AM: **Quantitative immunology for physicists**. *arXiv* 2019, **1907**, 03891.
This article comprehensively covers models of repertoire dynamics, and calculates predictions for the number of clones used in this article.

38. De Boer RJ, Perelson AS: **Quantifying T lymphocyte turnover**. *J Theor Biol* 2013, **327**:45–87.

39. De Boer RJ, Perelson aS: **Towards a general function describing T cell proliferation**. *J Theor Biol* 1995, **175**:567–576.

40. Desponds J, Mayer A, Mora T, Walczak AM: *Population dynamics of immune repertoires*. 2017:1–9. arXiv:1703.00226.

41. Mayer A, Balasubramanian V, Walczak AM, Mora T: **How a well-adapting immune system remembers**. *Proc Natl Acad Sci* 2019, **116**:8815–8823.

42. Marcou Q, Mora T, Walczak AM: **High-throughput immune repertoire analysis with IGoR**. *Nat Commun* 2018, **9**:561.

43. Sethna 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**:2974–2981.

44. Pogorelyy MV, *et al.*: **Precise tracking of vaccine-responding T-cell clones reveals convergent and personalized response in identical twins**. *Proc Natl Acad Sci* 2018, **115**:12704–12709.

45. Greiff V, *et al.*: **Systems analysis reveals high genetic and antigen-driven predetermination of antibody repertoires throughout B cell development**. *Cell Rep* 2017, **19**:1467–1478.

46. Rubelt F, *et al.*: **Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data**. *Nat Immunol* 2017, **18**:1274–1278.

47. Zarnitsyna VI, Evavold BD, Schoettle LN, Blattman JN, Antia R: **Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire**. *Front Immunol* 2013, **4**:485.

48. Yaari G, Kleinstein SH: **Practical guidelines for B-cell receptor repertoire sequencing analysis**. *Genome Med* 2015, **7**:121.