Book Section:
Ferrarini, M, Molina-Paris, C and Lythe, G (2017) Sampling from T cell receptor repertoires. In: Graw, F, Matthäus, F and Pahle, J, (eds.) Modeling Cellular Systems. Contributions in Mathematical and Computational Sciences, 11. Springer, pp. 67-79. ISBN: 978-3-319-45833-5.

https://doi.org/10.1007/978-3-319-45833-5_3

Sampling from T cell receptor repertoires

Marco Ferrarini, Carmen Molina-París, Grant Lythe

Abstract Modern single-cell sequencing techniques allow the unique TCR signature of each of a sample of hundreds of T cells to be read. The mathematical challenge is to extrapolate from the properties of a sample to those of the whole repertoire of an individual, made up of many millions of T cells. We consider the distribution of the number of repeats of any TCR in a sample, the mean number of samples needed to find a repeat with probability one half, and the relationship between the true distribution of clonal sizes and that experimentally observed in the sample. We consider two special cases, where the distribution of clonal sizes is geometric, and where a subset of clones in the repertoire is expanded.

1 Introduction

Approximately 4×10^{11} T cells circulate in the adult human body [1]. About 30,000 T cell receptors (TCRs) are on the surface of each T cell, usually of only one specificity [2]. T cells are selected in the thymus by binding to self-peptides expressed in association with major histocompatibility complex molecules (self-pMHC) [3, 1, 4, 5]. The set of cells with the same TCR defines a T cell clonotype, and the set of T cells in the body can be thought of as a repertoire of clonotypes. $CD8^+$ T cells recognise peptide bound to MHC class I and $CD4^+$ T cells recognise peptide bound to MHC class II [2]. How many TCR clonotypes are there in humans, mice and other mammals? [6, 7, 8, 9]. Direct measurement is not possible even with the latest developments in sequencing techniques. Estimates of the number of different TCRs that could, in principle, be produced by VDJ gene rearrangement in the

Marco Ferrarini, e-mail: mmmf@leeds.ac.uk · Carmen Molina-París, e-mail: carmen@maths.leeds.ac.uk · Grant Lythe, e-mail: grant@maths.leeds.ac.uk
Department of Applied Mathematics, School of Mathematics, University of Leeds, Leeds LS2 9JT, United Kingdom.

thymus, are about 10^{15} [10, 11, 12, 13]. However, the human body cannot contain even one T cell of 10^{15} possible types: 10^{15} T cells would weigh about 500 kg [14].

The number of distinct TCR clonotypes, N, is equal to the total number of T cells divided by the mean number of cells per clonotype. Equivalently, N is equal to the product of the rate of release of new clonotypes from the thymus to the periphery, θ, and the mean lifetime of a clonotype in the periphery. Lower limits on the number of distinct TCR β chains in the repertoire are about 4×10^6 [15, 16, 17, 18]. If each TCR β chain combines with 25 α chains, then the number of distinct clonotypes in one human is at least 10^8 [19]. An upper limit is the total number of T cells, about 4×10^{11}.

Direct estimates of β chain diversity have been made by PCR amplification of mRNA from pools of cells, but the technique is less suitable for measuring distributions of clonal sizes because numbers of mRNA vary from cell to cell and PCR amplification may depend on the TCR. Single-cell measurements, where PCR and sequencing is performed on one cell at a time, avoid biases. However, their expense means that only hundreds of cells are usually sequenced from a single mammal, and estimates of diversity must therefore rely on mathematical extrapolation from small samples [20, 17, 21, 22].

1.1 Sampling from a repertoire

Suppose that a sample of m cells taken from a total number of cells, S, is divided into N TCR clonotypes. Let us denote by n_i the number of cells of a clonotype labelled i. If $mn_i \ll S$ then

- the probability that none of the m cells in the sample are of clonotype i is $\left(1-\frac{n_i}{S}\right)^m$,
- the probability that exactly one of the m cells in the sample is of clonotype i is $m\frac{n_i}{S}\left(1-\frac{n_i}{S}\right)^{m-1}$,
- the probability that exactly two of the m cells in the sample are of clonotype i is p_i where

$$p_i = \frac{1}{2}m(m-1)\frac{n_i}{S}\frac{n_i-1}{S}\left(1-\frac{n_i}{S}\right)^{m-2}.$$

If $m \gg 1$ but $mn_i/S \ll 1$ then $p_i \simeq r_i$, where

$$r_i = \frac{1}{2}\left(\frac{m}{S}\right)^2 n_i(n_i-1)\,.$$

We say there is a repeat in the sample if two (or more) of the m cells are of the same clonotype. The condition $mn_i/S \ll 1$, or equivalently, $\sqrt{r_i} \ll 1$, ensures that finding three copies of the same type of cell (or clonotype) in the sample is sufficiently small that the probability of a repeat is equivalent to the probability of finding two cells of the given clonotype.

Now consider a set of M identified clonotypes. How many repeats, of clonotypes in this set, will we see in our sample? If the numbers of cells in the identified clonotypes are $n_1, n_2, \ldots, n_M$ and

$$r_i \ll 1 \qquad \forall i = 1, \ldots, M \,,$$

so that the occurrences of repeats in distinct clonotypes can be taken as independent events, then

$$\mathbb{E}(\text{number of repeats of identified clonotypes}) = \lambda \,,$$

where

$$\lambda = \sum_{i=1}^{M} r_i = \frac{1}{2}\left(\frac{m}{S}\right)^2 \sum_{i=1}^{M} n_i(n_i - 1) \,. \tag{2}$$

That is,

$$\mathbb{E}(\text{number of repeats of identified clonotypes}) = \frac{1}{2}\frac{m^2}{S^2} M \, \mathbb{E}(n_i(n_i - 1)) \,, \tag{3}$$

where the expectation is taken over the M clonotypes:

$$\mathbb{E}(n_i(n_i - 1)) = M^{-1} \sum_{i=1}^{M} n_i(n_i - 1) \,.$$

2 Results

2.1 The mean number of repeats

To find the mean number of repeats of *any clonotype from the repertoire* in the sample, we set $M = N$ in (3) and write $S = N\,\mathbb{E}(n_i)$, to obtain

$$\mathbb{E}(\text{number of repeats}) = \sum_{i=1}^{N} r_i = \frac{m^2}{2N} \frac{\mathbb{E}(n_i(n_i - 1))}{\mathbb{E}(n_i)^2} \,. \tag{4}$$

The expression (4) is the product of the factor $\frac{m^2}{2N}$, that does not depend on the distribution of clonal sizes, and the factor $\frac{\mathbb{E}(n_i(n_i-1))}{\mathbb{E}(n_2)^2}$, that does. The latter can be written

$$\frac{\mathbb{E}(n_i(n_i - 1))}{\mathbb{E}(n_i)^2} = \frac{\mathbb{E}(n_i^2)}{\mathbb{E}(n_i)^2} - \frac{1}{\mathbb{E}(n_i)} \,.$$

- If $n_i = \bar{n}$ for every i then

$$\frac{\mathbb{E}(n_i^2)}{\mathbb{E}(n_i)^2} = 1 \qquad \text{and} \qquad \frac{\mathbb{E}(n_i(n_i-1))}{\mathbb{E}(n_i)^2} = 1 - \frac{1}{\bar{n}} \, .$$

- If n_i has a geometric distribution with mean $\bar{n}$ (that is, $\mathscr{P}(n_i \geq k) = (1-\frac{1}{\bar{n}})^{k-1}$, $k = 1, 2, \ldots$) then

$$\frac{\mathbb{E}(n_i^2)}{\mathbb{E}(n_i)^2} = 2 - \frac{1}{\bar{n}} \qquad \text{and} \qquad \frac{\mathbb{E}(n_i(n_i-1))}{\mathbb{E}(n_i)^2} = 2\left(1 - \frac{1}{\bar{n}}\right) \, .$$

See Figure 1.

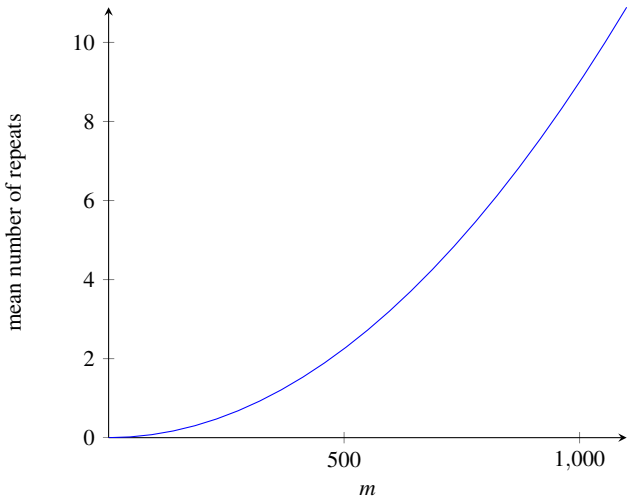

Fig. 1 Mean number of repeats as a function of the number of cells in the sample, from a repertoire of $N = 10^5$ clonotypes and a geometric distribution of clonal sizes, with $\bar{n} = 10$.

2.2 Number of draws to find the first repeat

Let us consider the probability of finding no repeats in a sample of m cells, as a function of m.

$$\mathscr{P}(\text{no repeat in sample of } m \text{ cells}) = \prod_{i=1}^{N} (1 - r_i) \, ,$$

so

$$\log\left(\mathscr{P}\left(\text{no repeat in sample of } m \text{ cells}\right)\right) = \sum_{i=1}^{N} \log(1-r_i)$$
$$\simeq -\sum_{i=1}^{N} r_i \ ,$$

assuming $r_i \ll 1$ for every i. Thus

$$\mathscr{P}\left(\text{no repeat in sample of } m \text{ cells}\right) = \exp\left(-\lambda\right) \ ,$$

where

$$\lambda = \frac{m^2}{2N} \frac{\mathbb{E}(n_i(n_i-1))}{\mathbb{E}(n_i)^2} \ .$$

How many cells do we need to sample in order to have a 50 percent chance of finding a repeat? Let this number be $m_{0.5}$. Then

$$m_{0.5}^2 = \frac{\mathbb{E}(n_i)^2}{\mathbb{E}(n_i(n_i-1))} \, 2N\log 2 \ . \tag{5}$$

In the simplest case, when all clonotypes have the same number of cells, n, we find $\mathscr{P}\left(\text{no repeats}\right) = \exp(-\frac{m^2}{2N}(1-\frac{1}{n}))$ and

$$m_{0.5} = \left(\frac{2N\log 2}{1-\frac{1}{n}}\right)^{\frac{1}{2}} \ .$$

When the distribution of the number of cells per clonotype is geometric,

$$m_{0.5} = \left(\frac{N\log 2}{1-\frac{1}{\bar{n}}}\right)^{\frac{1}{2}} \ .$$

See Figure 2.

2.3 Poisson distribution of number of repeats in a sample

Let k be the total number of repeats in a sample of m cells. We have already seen $\mathscr{P}\left(k=0\right)$. Let us consider $k=1$:

$$\mathscr{P}\left(k=1\right) = \sum_{i=1}^{N} r_i \prod_{\substack{i=1 \\ j\neq i}}^{N} (1-r_j) \ .$$

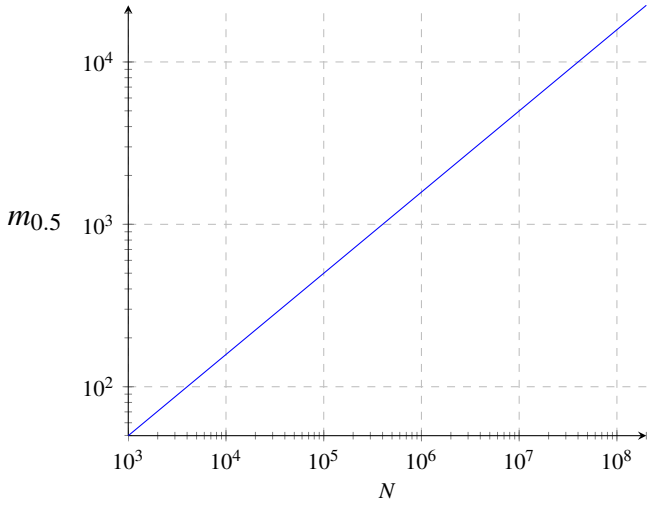

Fig. 2 Mean number of cells that need to be sampled in order to have a 50 percent chance of one repeat, from a repertoire of N clonotypes and a geometric distribution of clonal sizes, with $\bar{n} = 10$.

If $r_i \ll 1$ for every i then $\prod_{\substack{i=1 \\ j\neq i}}^{N}(1-r_j) \simeq \prod_{i=1}^{N}(1-r_i)$ and

$$\mathscr{P}(k=1) = \lambda \mathrm{e}^{-\lambda} .$$

The same argument works for all $k \ll m$, so that it is easy to see that the number of repeats in a sample has a Poisson distribution:

$$\mathscr{P}(\text{number of repeats is } k) = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda} .$$

2.4 Estimating the size of the repertoire from one repeat

Suppose there is one repeat in a sample of m_1 cells. We then estimate that $\lambda = 1$ and, assuming a geometric distribution of clonal sizes, conclude that

$$N = m_1^2 .$$

If we find one repeat per 100 cells, we estimate the size of the repertoire is 10^4. If we find one repeat per 1000 cells, we estimate that the size of the repertoire is 10^6. In practice, the estimate m_1^2 is likely to be conservative, because any clonal expansion will increase the number of observed repeats.

2.5 The observed distribution of clonal sizes

How many times will we see k copies of the same TCR in a sample of m cells?

Let us start at the beginning, with the probability q of finding a single cell in a sample of m cells from a total of S cells, equal to $1-\left(1-\frac{1}{S}\right)^m$. Let us define the Bernoulli random variable B:

$$\mathscr{P}(B=0)=1-q \text{ and } \mathscr{P}(B=1)=q, \qquad \text{where } q=1-\left(1-\frac{1}{S}\right)^m . \quad (6)$$

The probability generating function of B is

$$\phi_B(z)=1-q+qz\,. \quad (7)$$

If n_i is the number of cells of a clonotype labelled i, then the number of cells of type i in the sample is the random variable Y_i, which can be written

$$Y_i=B_0+B_1+\cdots+B_{n_i}\,, \quad (8)$$

where B_j, $j=1,\ldots,n_i$ are random variables with the same distribution as B. With the approximation that the B_j are independent random variables, the probability generating function of Y_i is

$$\phi_{Y_i}(z)=\phi_B(z)^{n_i}=(1-q+qz)^{n_i}\,. \quad (9)$$

Let Y be the number of copies of a randomly-chosen clonotype found in the sample of m cells. We must take the distribution of values of n_i into account. Suppose that the probability generating function of the random variable n_i is $\phi_n(z)$. Then

$$\begin{aligned}\phi_Y(z)&=\sum_k \mathscr{P}(n_i=k)(1-q+qz)^k\\&=\phi_n(1-q+qz)\,.\end{aligned}$$

For example, if n_i has a geometric distribution with mean $\bar{n}$, then $\phi_n(z)=\dfrac{z}{\bar{n}-(\bar{n}-1)z}$, so that

$$\phi_Y(z)=\frac{1-q+qz}{\bar{n}-(\bar{n}-1)(1-q+qz)}=\frac{1-q+qz}{1-(\bar{n}-1)q(1-z)}\,. \quad (10)$$

Therefore, the observed clonal sizes s_1, s_2, ... are distributed as follows:

$$s_k=s_1\gamma^{k-1} \qquad k\geq 1\,, \quad (11)$$

where

$$\gamma=\frac{(\bar{n}-1)q}{1+(\bar{n}-1)\bar{n}q}\,, \quad (12)$$

is the factor by which finding a clone of size $k+1$ in the sample is less frequent than finding a sample of size k in the sample. We conclude that the observed distribution of clonal sizes (the histogram that is obtained by plotting number of TCRs versus number of cells) is also geometric, with mean $1+(\bar{n}-1)q$. See Figure 3.

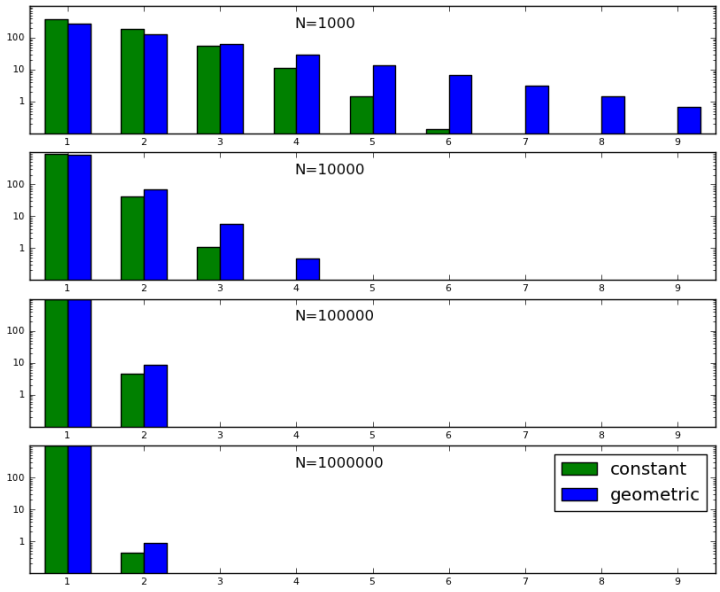

Fig. 3 Observed clonal size distribution in a sample of 1000 cells, from repertoires containing different numbers of clones, N. A "constant" repertoire means that there are 10 cells of each clonotype. In a "geometric" repertoire, the number of cells in each clonotype is drawn from a geometric distribution with mean 10.

3 Expansion of a subset of the repertoire

In this Section, we assume that a fraction $f \ll 1$ of clones are "expanded" by an infection, so that their number of cells is multiplied by a factor $\alpha \gg 1$. When considering an expanded clone, labelled i, we can no longer assume that $mn_i/S \ll 1$. Thus

- the probability that none of the m cells in the sample are of the expanded clonotype i is $(1-\frac{n_i}{S})^m$.

- the probability that exactly one of the m cells in the sample is of clonotype i is $m\frac{n_i}{S}(1-\frac{n_i}{S})^{m-1}$.
- the probability that exactly two of the m cells in the sample are of clonotype i is p_i, where

$$p_i = \frac{1}{2}m(m-1)\frac{n_i}{S}\frac{n_i-1}{S}\left(1-\frac{n_i}{S}\right)^{m-2}. \tag{13}$$

We consider the simplest case where all unexpanded clones contain the same number of cells $\bar{n}$. Then the total number of cells is S' where

$$S' = N\bar{n}\left(1+(\alpha-1)f\right) . \tag{14}$$

That is, $S'/S = 1+(\alpha-1)f$. We write

$$r_i = \begin{cases} \frac{1}{2}(\frac{m}{S'})^2\bar{n}(\bar{n}-1) & \text{unexpanded clones ,} \\ \frac{1}{2}(\frac{m}{S'})^2\alpha^2\bar{n}^2 & \text{expanded clones .} \end{cases} \tag{15}$$

The mean number of repeats is Poisson distributed with mean

$$\lambda' = \sum_{i=1}^{N} r_i = \frac{1}{2}\frac{m^2}{N}\left(\frac{S'}{S}\right)^2\left[1-\frac{1}{\bar{n}}+f\alpha^2\left(1-\frac{\alpha}{N}\frac{S'}{S}\right)^m\right] . \tag{16}$$

4 Discussion

Small samples from a large repertoire, such as are obtained in single-cell sequencing experiments of T cell receptors, present mathematical challenges. Estimates of the diversity of the TCR repertoire, that can be deduced, depend on the distribution of clonal sizes, which is also unknown. However, small sample sizes allow the simplifying approximation that random variables describing quantities of interest, such as the numbers of cells of different types in the sample, are independent. Then, the probability generating function of the distribution of clonal sizes in the sample is the composition of that of a Bernoulli random variable (that takes values 0 or 1) and that of the true distribution of clonal sizes in the repertoire that is being sampled from.

Acknowledgements The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) through the Marie-Curie Action *"Quantitative T cell Immunology"* Initial Training Network, with reference number FP7-PEOPLE-2012-ITN 317040-QuanTI.

References

1. M.K. Jenkins, H.H. Chu, J.B. McLachlan, and J.J. Moon. On the composition of the preimmune repertoire of T cells specific for peptide-major histocompatibility complex ligands. *Annual Review of Immunology*, 28:275–294, 2009.
2. R. Varma. TCR triggering by the pMHC complex: valency, affinity, and dynamics. *Science Signaling*, 1(19):pe21, 2008.
3. BENEDITA Rocha and HARALD von Boehmer. Peripheral selection of the T cell repertoire. *Science*, 251(4998):1225–1228, 1991.
4. I. Bains, R. Antia, R. Callard, and A.J. Yates. Quantifying the development of the peripheral naive $CD4^+$ T-cell pool in humans. *Blood*, 113(22):5480, 2009.
5. François Van Laethem, Anastasia N Tikhonova, and Alfred Singer. MHC restriction is imposed on a diverse T cell receptor repertoire by CD4 and CD8 co-receptors during thymic selection. *Trends in Immunology*, 33(9):437–441, 2012.
6. RE Langman and M Cohn. The ET (elephant-tadpole) paradox necessitates the concept of a unit of B-cell function: the protecton. *Molecular Immunology*, 24(7):675–697, 1987.
7. Joseph N Blattman, Rustom Antia, David JD Sourdive, Xiaochi Wang, Susan M Kaech, Kaja Murali-Krishna, John D Altman, and Rafi Ahmed. Estimating the precursor frequency of naive antigen-specific CD8 T cells. *Journal of Experimental Medicine*, 195(5):657–664, 2002.
8. Stanca M Ciupe, Blythe H Devlin, Mary Louise Markert, and Thomas B Kepler. Quantification of total T-cell receptor diversity by flow cytometry and spectratyping. *BMC Immunology*, 14(1):1–12, 2013.
9. Niclas Thomas, Katharine Best, Mattia Cinelli, Shlomit Reich-Zeliger, Hilah Gal, Eric Shifrut, Asaf Madi, Nir F riedman, John Shawe-Taylor, and Benny Chain. Tracking global changes induced in the CD4 T cell receptor repertoire by immunisation with a complex antigen using local sequence features of CDR3 protein sequence. biorxiv.org, 2014.
10. Andrew K Sewell. Why must T cells be cross-reactive? *Nature Reviews Immunology*, 12(9):669–677, 2012.
11. Janko Nikolich-Žugich, Mark K. Slifka, and Ilhem Messaoudi. The many important facets of T-cell repertoire diversity. *Nature Reviews Immunology*, 4(2):123–132, 2004.
12. Veronika Zarnitsyna, Brian Evavold, Louie Schoettle, Joseph Blattman, and Rustom Antia. Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire. *Frontiers in Immunology*, 4:485, 2013.
13. Anand Murugan, Thierry Mora, Aleksandra M Walczak, and Curtis G Callan. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proceedings of the National Academy of Sciences*, 109(40):16161–16166, 2012.
14. D. Mason. A very high level of crossreactivity is an essential feature of the T-cell receptor. *Immunology Today*, 19(9):395–404, 1998.
15. T.P. Arstila, A. Casrouge, V. Baron, J. Even, J. Kanellopoulos, and P. Kourilsky. A direct estimate of the human $\alpha\beta$ T cell receptor diversity. *Science*, 286(5441):958, 1999.
16. Can Keşmir, José AM Borghans, and Rob J de Boer. Diversity of human $\alpha\beta$ T cell receptors. *Science*, 288(5469):1135–1135, 2000.
17. Harlan S Robins, Paulo V Campregher, Santosh K Srivastava, Abigail Wacher, Cameron J Turtle, Orsalem Kahsai, Stanley R Riddell, Edus H Warren, and Christopher S Carlson. Comprehensive assessment of T-cell receptor β-chain diversity in $\alpha\beta$ T cells. *Blood*, 114(19):4099–4107, 2009.
18. René L Warren, J Douglas Freeman, Thomas Zeng, Gina Choe, Sarah Munro, Richard Moore, John R Webb, and Robert A Holt. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Research*, 21(5):790–797, 2011.
19. Qian Qi, Yi Liu, Yong Cheng, Jacob Glanville, David Zhang, Ji-Yeun Lee, Richard A Olshen, Cornelia M Weyand, Scott D Boyd, and Jörg J Goronzy. Diversity and clonal selection in the human T-cell repertoire. *Proceedings of the National Academy of Sciences*, 111(36):13139–13144, 2014.

20. Vanessa Venturi, Katherine Kedzierska, Stephen J Turner, Peter C Doherty, and Miles P Davenport. Methods for comparing the diversity of samples of the T cell receptor repertoire. *Journal of Immunological Methods*, 321(1):182–195, 2007.
21. N. Sepúlveda, C.D. Paulino, and J. Carneiro. Estimation of T-cell repertoire diversity and clonal size distribution by poisson abundance models. *Journal of Immunological Methods*, 353(1):124–137, 2010.
22. DJ Laydon, CRM Bangham, and B Asquith. Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach. *Philosphical Transactions of the Royal Society B*, 370, 2015.