Journal of Immunological Methods 353 (2010) 124–137

Contents lists available at ScienceDirect

Journal of Immunological Methods

journal homepage: www.elsevier.com/locate/jim

Computational modeling

Estimation of T-cell repertoire diversity and clonal size distribution by Poisson abundance models

Nuno Sepúlveda [a,b,*], Carlos Daniel Paulino [b,c], Jorge Carneiro [a,*]

[a] Instituto Gulbenkian de Ciência, Portugal
[b] Center of Statistics and Applications, University of Lisbon, Portugal
[c] Department of Mathematics, Instituto Superior Técnico, Portugal

ARTICLE INFO

Article history:
Received 6 August 2009
Received in revised form 10 November 2009
Accepted 10 November 2009
Available online 18 November 2009

Keywords:
TCR diversity
Clonal size distribution
Poisson abundance models
Package PAM

ABSTRACT

The answer to many fundamental questions in Immunology requires the quantitative characterization of the T-cell repertoire, namely T cell receptor (TCR) diversity and clonal size distribution. An increasing number of repertoire studies are based on sequencing of the TCR variable regions in T-cell samples from which one tries to estimate the diversity of the original T-cell populations. Hitherto, estimation of TCR diversity was tackled either by a "standard" method that assumes a homogeneous clonal size distribution, or by non-parametric methods, such as the abundance-coverage and incidence-coverage estimators. However, both methods show caveats. On the one hand, the samples exhibit clonal size distributions with heavy right tails, a feature that is incompatible with the assumption of an equal frequency of every TCR sequence in the repertoire. Thus, this "standard" method produces inaccurate estimates. On the other hand, non-parametric estimators are robust in a wide range of situations, but per se provide no information about the clonal size distribution. This paper redeploys Poisson abundance models from Ecology to overcome the limitations of the above inferential procedures. These models assume that each TCR variant is sampled according to a Poisson distribution with a specific sampling rate, itself varying according to some Exponential, Gamma, or Lognormal distribution, or still an appropriate mixture of Exponential distributions. With these models, one can estimate the clonal size distribution in addition to TCR diversity of the repertoire. A procedure is suggested to evaluate robustness of diversity estimates with respect to the most abundant sampled TCR sequences. For illustrative purposes, previously published data on mice with limited TCR diversity are analyzed. Two of the presented models are more consistent with the data and give the most robust TCR diversity estimates. They suggest that clonal sizes follow either a Lognormal or an appropriate mixture of Exponential distributions. According to the ecological interpretation of these models, the T-cell repertoire would be divided in several T-cell niches, themselves created in a series of steps. Definitive conclusions, however, would require larger samples. It is shown here that samples 100-fold larger than hitherto available ones would be sufficient to discriminate candidate models. These large sample sizes are currently affordable using massively parallel sequencing technology. Foreseeing this we provide the package `PAM` for the R software that will facilitate T-cell repertoire data analysis based on Poisson abundance models.

1. Introduction

T cells recognize and respond to antigens via their T cell receptors (TCRs). TCRs are heterodimers with two chains: α and β in $\alpha\beta$ T cells and γ and δ in $\gamma\delta$ T cells. The genes encoding these proteins are generated somatically by V(D)J recombination during thymic T-cell development. By this

Abbreviations: DP, double positive; LN, lymph node; PAM, Poisson abundance model; SP, single positive; TCR, T-cell receptor.

* Corresponding authors. Instituto Gulbenkian de Ciência, Apartado 14, P-2781-901 Oeiras, Portugal. Tel.: +351 21 446 46 14; fax: +351 21 440 79 70.
E-mail addresses: nunosep@igc.gulbenkian.pt (N. Sepúlveda), jcarneir@igc.gulbenkian.pt (J. Carneiro).

0022-1759/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jim.2009.11.009

process, T-cell precursors randomly recombine different V, D, and J gene segments and assemble the mature gene encoding a TCR chain. Additional diversity is obtained by an imprecise joining of those gene segments. In theory, there are 10^{18} different possibilities of generating an $\alpha\beta$ TCR in humans (Janeway et al., 2005) and 10^{15} in mice (Davis and Bjorkman, 1988). However, at any given time, $\alpha\beta$ TCR diversity has been estimated to be $>2\times10^7$ in humans (Arstila et al., 1999; Naylor et al., 2005) and around 2×10^6 in mice (Casrouge et al., 2000).

Many fundamental questions in Immunology are intimately related to T-cell diversity, such as: what are the general structural properties of the T-cell repertoire (Correia-Neves et al., 2001); what is the T-cell diversity of an immune response against a given viral infection (Naumov et al., 2003; Pewe et al., 2004); how diverse is the memory T cell pool (Kedzierska et al., 2006); or how diverse is the regulatory $CD4^+CD25^+Foxp3^+$ T-cell repertoire and how does it relate to that of conventional $CD4^+$ T cells (Hsieh et al., 2004; Hsieh et al., 2006; Pacholczyk et al., 2006; Wong et al., 2007).

To answer the above questions, one generally collects a sample of T cells from an individual and sequences their TCRs by some experimental technique. In this scenario, one defines a clonotype as a distinct TCR sequence, either at the nucleotide or amino-acid level, while the clonal size is the number of cells in the body bearing the same TCR sequence. The so-called clonal size distribution is the frequency of clonotypes with a certain clonal size, and embodies all relevant information about the shape of the T-cell repertoire.

There are several statistical approaches to assess T-cell diversity from TCR sequence data. Simpson's diversity and Shannon's entropy indexes have been used as measures of diversity in samples (Ferreira et al., 2009; Venturi et al., 2007). However, these two indexes are just summary statistics, thus providing limited information about the diversity of the T cell populations from which the samples were taken from. Another approach is to estimate diversity by a model that assumes that all clonotypes are equally represented in the repertoire (Barth et al., 1985; Behlke et al., 1985). This model is usually seen as a standard tool to estimate TCR diversity as judged by its wide application to different data sets (e.g., Casrouge et al., 2000; Hsieh et al., 2004; Hsieh et al., 2006; Pacholczyk et al., 2006; Pacholczyk et al., 2007). However, sampled clonal size distributions show often heavy right tails, which are inconsistent with the above assumption, as noted by Naumov et al. (2003) and Pewe et al. (2004). Therefore the above method is expected to provide unreliable diversity estimates.

Some authors (Hsieh et al., 2006; Pacholczyk et al., 2006) also estimated diversity through the abundance-coverage and incidence-coverage estimators, which take into account some heterogeneity in the clonal size distributions (Chao and Lee, 1992). Although these estimators may be robust in a wide range of situations, they provide no information of the underlying clonal size distribution.

To overcome the limitations of current inferential methods, we propose the usage of Poisson abundance models (PAMs), adopted from Ecology, to estimate both TCR diversity and the underlying clonal size distribution. We focus our attention on five simple PAMs due to the vast ecological literature that can be found on them: the Homogeneous Poisson model, which is mathematically equivalent to the above-mentioned standard method; the Geometric model, which is based on an Exponential clonal size distribution observed in bird communities (MacArthur, 1957); the Poisson–Gamma model, which is an extension of the former by considering a Gamma clonal size distribution (Fisher et al., 1943); the Poisson–Lognormal model, which is often assumed as a null model in Ecology for species abundance distribution (Magurran and Henderson, 2003; McGill, 2003) due to its recurrent observation in many different ecological contexts (Preston, 1948); the Yule model (Yule, 1925), which has been used as a rank abundance distribution and particularly suitable to describe self-organized communities due to its power law tails (Levich, 1980). Any of the above models can be viewed in the light of some ecological mechanism shaping the community structure, and the corresponding immunological interpretation will be briefly addressed in the general Discussion. Yet, the in-depth discussion of these putative mechanisms is beyond the scope of this paper.

For illustrative purposes, the above models are fitted to data from mice with limited TCR diversity (Correia-Neves et al., 2001). However, it is worth noting from the outset that the specific results and estimates obtained in this example should and will be treated with extreme caution due to rather small sample sizes.

2. Poisson abundance models

TCR sequence data refer to the frequency of clonotypes that appear a certain number of times in the sample, the so-called sampled clonal size distribution. In statistical terms, the respective sampling distribution is usually described by the following Multinomial law (Sanathanan, 1972):

$$P[\{m_i\}\,|\,D,\eta] = \frac{D!}{(D-M)!\prod_{i=1}^{n} m_i!}\,[p_\eta(0)]^{D-M}\prod_{i=1}^{n}\,[p_\eta(i)]^{m_i}, \quad (1)$$

where D is the number of distinct clonotypes in the original T-cell population (population diversity), m_i is the number of distinct clonotypes with i copies in the sample, $p_\eta(i)$ is the probability of a clonotype being sampled i times that, in turn, is described by a model with parameter vector η, $M=\sum_{i=1}^{n} m_i$ is the number of distinct clonotypes obtained in the sample (sample diversity), and $n=\sum_{i=1}^{n} i\times m_i$ is the sample size. The first goal of the analysis is to estimate the population diversity D and the parameter vector η.

To derive the probability $p_\eta(i)$, we focus on the event that i cells from a generic clonotype are present in the sample. In this case, if one fixes the sample size by experimental design (or has a random sample size not affected by the composition of the repertoire) and if all cells in the repertoire show the same probability of being sampled, the sampling scheme is conceptually equivalent to collecting n individuals from a finite population with size N where m individuals exhibit a certain characteristic of interest. Therefore, the number of cells belonging to a clonotype in the sample is given by a Hypergeometric distribution with parameters N, m, and n. However, when the sample size is much smaller than the population size, which often occurs in practice, the Poisson

distribution can be used to approximate the Hypergeometric distribution (Dewdney, 1998). By modeling appropriately the Poisson law parameter, hereafter referred to as the sampling rate, one obtains the class of PAMs.

The simplest PAM is to consider that all clonotypes are equally represented in the repertoire, which leads to a homogeneous Poisson distribution across all clonotypes (Homogeneous Poisson model)

$$p_\lambda(i) = \frac{e^{-\lambda}\lambda^i}{i!}, \tag{2}$$

where λ is the sampling rate, defined as the expected value of the Hypergeometric distribution (i.e., $\lambda = n \times m/N$). In this scenario, Barth et al. (1985) have derived the following conditional distribution of $\{m_i\}$ given the sample size n

$$P[\{m_i\}|D,n] = \binom{D}{M}\frac{M!}{\prod_{i=1}^{n} m_i!}\frac{n!}{\prod_{i=1}^{n}(i!)^{m_i}}\frac{1}{D^n}, \tag{3}$$

which might be used alternatively to estimate D, since the sample size n has no information on D (i.e. n is an ancillary statistic for D) . It is worth noting that one can prove that the statistic M embodies all sampling information about D (i.e., M is a specific sufficient statistic for D). In this case, one can use its sampling distribution to estimate D, as done in Behlke et al. (1985).

Clonotypes have different sampling rates if they differ in their clonal sizes. Thus, different models can be obtained by considering a probability distribution for the sampling rate λ. By doing this, one is modeling indirectly the clonal size distribution, as we will see in Section 4.

We first introduce the Poisson–Gamma model (Fisher et al., 1943), in which the probabilistic parameters of the Multinomial law (1) result from integrating out the sampling rate λ by mixing (2) with a Gamma density for λ, that is,

$$\begin{aligned} p_{\alpha,\beta}(i) &= \int_0^\infty \frac{e^{-\lambda}\lambda^i}{i!}\frac{\beta^\alpha\lambda^{\alpha-1}e^{-\lambda\beta}}{\Gamma(\alpha)}d\lambda \\ &= \frac{\Gamma(i+\alpha)}{\Gamma(i+1)\Gamma(\alpha)}\left(\frac{\beta}{\beta+1}\right)^\alpha\left(\frac{1}{\beta+1}\right)^i, \end{aligned} \tag{4}$$

where α and β are the shape and scale parameters of the Gamma distribution, respectively. Note that, when α is an integer, the above equation converts into the well-known Negative-Binomial distribution. Furthermore, if $\alpha=1$, the sampling rate λ is exponentially distributed, leading to the Geometric model

$$p_\beta(i) = \left(\frac{1}{\beta+1}\right)^i\frac{\beta}{\beta+1}. \tag{5}$$

Finally, the Log-Series distribution, which has already been applied to TCR sequence data (Pewe et al., 2004), is the limit distribution of Eq. (4) when α tends to zero (Fisher et al., 1943). However, this model cannot describe the probability of not sampling a clonotype, $p_\eta(0)$, included in Eq. (1). Therefore, under this framework, the Log-Series distribution should not be used to estimate diversity. Instead, one can follow the approach described in detail in Pewe et al. (2004).

The Poisson–Lognormal model considers a Lognormal distribution for the sampling rate λ (Bulmer, 1974). That is, the logarithm of λ follows a Normal distribution, say with mean value μ and variance σ^2. Under this model, the probability of sampling i cells of a given clonotype is described by the following integral

$$p_{\mu,\sigma^2}(i) = \int_0^\infty \frac{e^{-\lambda}\lambda^{i-1}}{i!}\frac{e^{-\frac{(\ln\lambda-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}d\lambda, \tag{6}$$

which cannot be expressed in closed form. To overcome this problem, Bulmer (1974) suggests a numerical algorithm to calculate the above probabilities. It is worth noting that this model shows a heavier right tail than any of the above models, and this feature may be useful to fit many TCR sequence data.

Another model with heavy right tails is the Yule distribution (Yule, 1925) that has been used to model self-organized communities (Levich, 1980). This distribution has been also applied to describe sexual network formation (Jones and Handcock, 2003) and Internet growth (Barabási and Albert, 1999). For what concerns diversity estimation, the Yule distribution can be derived from the following hierarchical structure: the sampling rate λ follows an Exponential distribution with parameter given by $\mu = e^{-\omega}/(1-e^{-\omega})$, where ω is distributed as another Exponential distribution but with parameter ρ. After integrating out λ, one obtains the following distribution

$$p_\rho(i) = \rho\frac{\Gamma(i+1)\Gamma(\rho+1)}{\Gamma(i+\rho+2)}. \tag{7}$$

For sufficiently large i, the above probability is approximately proportional to $1/i^{\rho+2}$, typical of a power-law model. Given this feature, one expects that, under this model, samples would exhibit many rare clonotypes with just one or two copies (singletons and doubletons, respectively) and few highly abundant clonotypes.

3. Estimation of diversity

3.1. Application of the maximum likelihood method

Diversity D and the parameter vector η can be estimated by maximizing likelihood function (1), where $p_\eta(i)$ follows Eqs. (2), (4), (5), (6) and (7) for the Homogeneous Poisson, Poisson–Gamma, Geometric, Poisson–Lognormal and Yule models, respectively. Since diversity D is an integer parameter and the likelihood function for these models is well-behaved and has a single maximum, one can use the following profile likelihood method: estimate first diversity by the number of clonotypes in the sample, $\hat{D}=M$, and then obtain the respective estimates of the parameter vector η; increase the diversity estimate in one unity and re-estimate η; repeat the previous step until the (log-)likelihood function starts to decrease. Considering fixed $\hat{D}$, the estimate of λ in the homogeneous Poisson model is given by $n/\hat{D}$. In the case of the Geometric model, the maximum likelihood estimate of β is given $\hat{D}/(\hat{D}+N)$. To estimate α and β in the Poisson–Gamma, one needs to use a numerical method, such as the

Newton–Raphson. With respect to the Yule model we estimated ρ as follows: (i) assume ρ integer, (ii) find the integer ρ^* that maximizes the log-likelihood given $\hat{D}$, (iii) define the interval (ρ^*-1, ρ^*+1), (iv) apply the well-known bisection method to this interval. Finally, the estimation of the Poisson–Lognormal model is a bit cumbersome, because this model does not show a closed-form expression for the probabilities $p_\eta(i)$. To overcome this problem, we used the Fisher scoring method to estimate μ and σ^2 as proposed by Bulmer (1974).

3.2. Goodness-of-fit and model comparison

After estimating the models, one should then evaluate their goodness-of-fit. At this point, it is important to emphasize the fact that Eq. (1) includes the number of clonotypes not observed in the sample, D–M. Since this number is estimated rather than observed, it should be excluded when evaluating the quality of fit of the models. For this purpose, we note that the unconditional likelihood function (1) can be factorized as follows (see, for example, Chao and Bunge, 2002)

$$P[\{m_i\}|D,\eta] = P[M|D,\eta] \times P[\{m_i\}|M,\eta], \tag{8}$$

where the first factor is given by a Binomial distribution with D trials and probability of success $1-p_\eta(0)$ and the second factor is described by the following Multinomial law

$$P[\{m_i\}|M,\eta] = \frac{M!}{\prod_{i=1}^{n} m_i!} \prod_{i=1}^{n} \left[\frac{p_\eta(i)}{1-p_\eta(0)}\right]^{m_i}. \tag{9}$$

In statistical terms, the above factorization demonstrates that M is a specific sufficient statistic for D, as mentioned before. In this context, it is expected that the conditional distribution (9) of the observed data given M does not depend on either D or D–M. Therefore, one can use it to perform goodness-of-fit tests and model comparisons on the sole basis of the observed data.

To assess the quality-of-fit of a given model, we perform the popular Pearson chi-squared test and select models with p-values higher than 5%. Since TCR sequence data are usually unbalanced (many singletons and doubletons, and few highly abundant TCR sequences), the asymptotic chi-square distribution for the test statistic under the null hypothesis may provide inaccurate results for the true p-value (Agresti, 1992). To avoid this potential problem, it is recommended to group data into fewer categories. In the example of this paper, we consider only five categories for the sampled clonal size distribution: four referring to 1, 2, 3, and 4 copies in the sample and a fifth category for clonotypes with more than 4 copies.

In practice, one may find a set of models that can fit the same data equally well (i.e., at a given significance level). Since the models have different numbers of parameters, they should be compared not only in terms of their goodness-of-fit but also based on some parsimony principle. To this end, one often calculates measures that attempt to capture the goodness-of-fit and the complexity of a model at the same time. In this regard, a popular choice for a model comparison measure is the Akaike's information criterion (AIC; Akaike, 1974) defined by the following equation

$$\mathrm{AIC} = -2\log\mathcal{L}^* + 2p, \tag{10}$$

where $\mathcal{L}^*$ is the likelihood function evaluated at the maximum likelihood estimates and p is the parameter dimension of a model. In this context, the best models are the ones that show the lowest values of AIC. Following the same rationale as in Pearson's goodness-of-fit test, the above measure is calculated according to the likelihood function (9) but evaluated at the parameter estimates that maximize the unconditional likelihood function (1). Since Eq. (9) does not include the parameter D, the number of parameters p of a model is given by the size of the vector η: 1 for the Homogeneous Poisson, Geometric and Yule models, and 2 for the Poisson–Gamma and Poisson–Lognormal models.

3.3. Correcting estimates for heavy right tails

Sampled clonal size distributions show often heavy right tails (Correia-Neves et al., 2001; Hsieh et al., 2006; Naumov et al., 2003). In this situation, the whole data might not be well described by the proposed PAMs. Moreover, diversity estimates might be sensitive to the presence of most abundant clonotypes in the samples. To tackle these problems, it has been recommended that the sampled clonal size distribution should be divided in two parts, one referring to the rare clonotypes and the other to the most abundant ones (Chao et al., 1993; Chao and Bunge, 2002). In this scenario, one fits the PAMs to the former part of the data, estimating the corresponding diversity. The total diversity estimate is obtained by adding to this the observed diversity of abundant clones.

A critical point of the analysis is then to determine the optimal threshold, say τ, that divides data in two parts. To this end, one should vary τ starting on its highest possible value (i.e., the number of copies of the most abundant clonotype) and then decreasing its value until the diversity estimates are reasonably stable and a good fit of the models is achieved. In this analysis, one should select models that fit the highest fraction of the data, producing at the same time stable diversity estimates.

4. Estimation of the clonal size distribution

A great advantage of fitting PAMs to sample data is the possibility of estimating the clonal size distribution in the original T-cell population. The Poisson sampling rate of a clonotype is described as $\lambda = n \times m/N$, where N is the number of cells in the population and m is a random variable describing the size of a generic clonotype (Dewdney, 1998). The quotient between n and N is called the sampling ratio, which should be considered a constant if N is known or if one has a good estimate for it. Therefore, the distribution of the random variable m is that of the sampling rate times the inverse of the sampling ratio ($m = \lambda \times N/n$). Theoretically, the clonal size distribution of an individual's T-cell repertoire refers to a random sample of D clonotypes taken from the random

variable m. For sake of simplicity, the random variable m will be referred to as the true clonal size distribution.

We now provide the clonal size distributions predicted by the different PAMs. These distributions are easily obtained from standard probability techniques for the transformation of random variables. In the Homogeneous Poisson model, the sampling rate is constant among clonotypes, and therefore the clonal size distribution is degenerate (a Dirac pulse) at the inverse of the sampling ratio times the sampling rate. In the Poisson–Gamma model, a Gamma distribution with parameters α and β is considered for the sampling rate, which implies a Gamma distribution for the clonal size distribution with parameters α and $\beta n/N$. As a special case of the Poisson–Gamma, the Geometric model assumes an Exponential distribution with parameter β for the sampling rate. In this case, the clonal size distribution is also Exponential but with parameter $\beta n/N$. The Poisson–Lognormal model describes the sampling rate by a Lognormal distribution, with parameters μ and σ^2. Thus, the clonal size distribution is also a Lognormal distribution but with parameters $\mu+\log N/n$ and σ^2. Finally, the Yule model describes the sampling rate by an appropriate hierarchical structure of two Exponential distributions. The clonal size distribution is a mixture of Exponential distributions with conditional mean (given w) $\mu = n/N \times e^{-w}/(1-e^{-w})$ where w follows an Exponential distribution with parameter ρ. In this case, the clonal size distribution has no analytical expression, but can be approximated by a Monte Carlo simulation: (i) draw a value w from an Exponential distribution with parameter ρ, (ii) calculate $\mu = n/N \times e^{-w}/(1-e^{-w})$, (iii) draw a value x from an Exponential distribution with parameter μ, (iv) repeat this process r times (say $r = 10{,}000$). In the end one can approximate the cumulative distribution function of the clonal sizes with the respective empirical cumulative distribution of the simulated sample $x_1,\ldots x_r$.

As suggested above (Section 3.3), one should estimate TCR diversity with and without the most abundant clonotypes. Since many TCR data show long right tails, it is likely that some clonotypes might be excluded from the analysis in order to obtain stable diversity estimates and good fits of the models. The excluded clonotypes are then treated as outliers under the selected models. In this case, the above estimation of the clonal size distribution needs to be adapted. One first considers the most abundant clonotype that was discarded from the PAM fitting. Its clonal size in the population is estimated to be $N \times m_a/n$, where m_a is the number of copies in the sample, assuming that the proportions of that clonotype in the population and the sample are identical. One discounts the contribution of this clonotype to the population size, obtaining the remaining number of cells $N \times (1 - m_a/n)$. By recursion we remove from the population size all the outlying clonotypes. The remaining cell number, $N \times (1 - \sum m_a/n)$, is then used to calculate the size of the rare clonotypes according to the selected PAM.

5. Package `PAM`

The PAM-based analysis of the experimental data was done using the R software (freely available at http://www.cran.r-project.org/). We developed the package `PAM` (from *Poisson Abundance Models*), which provides simple command lines to fit the models and assess their goodness-of-fit, as well as to estimate diversity and clonal size distribution of sampled populations. The package has been tested using previously published data sets, not only from Immunology but also from Ecology. PAM is publicly available at the following website http://qobweb.igc.gulbenkian.pt/. Instructions, example data, and usage details can be found in the respective documentation within R environment after installation.

6. Example

Correia-Neves et al. (2001) attempted to obtain a broad but manageable view of the thymic and peripheral $CD4^+$ and $CD8^+$ T cell repertoires. With this purpose, a mouse line was engineered to express a TCRβ transgene and a TCRα minilocus transgene composed of a single Vα region and two Jα segments. The expression of the endogeneous TCRα was prevented by inserting an appropriate mutation at the endogeneous α locus, while the blockage of the endogeneous TCRβ relies on allelic exclusion of the β locus. In this experimental setting, T cell diversity is restricted to the imprecise joining of the Vα region with either Jα segments. Further experimental details can be found in the original reference (Correia-Neves et al., 2001).

A large collection of thymic and peripheral CDR3 TCRα sequences is available from different mice (see supplementary material of Correia-Neves et al., 2001). Data under analysis will be of double positive $CD4^+$ $CD8^+$ thymocytes expressing low levels of CD3 protein (DP CD3low) from mouse 17, and of single positive (SP) thymocytes, either $CD4^+$ or $CD8^+$, and lymph node (LN) T cells from mouse 57 (Table 1). Data from other mice were not included in the analysis due to their even smaller sample sizes. Experimentally, TCRα's of SP thymocytes and of LN T cells were sequenced by single-cell RT-PCR, while those of DP CD3low by RT-PCR and cloning. TCR uniqueness

Table 1
Number of distinct CDR3 TCRα sequences with i copies in the samples, where M is the total number of distinct sequences (clonotypes) in the samples, n is the respective sample size, and D_s is the Simpson's diversity index. DP CD3low data are from mouse 17, while remaining data from mouse 57. After proper standardization of sample sizes, Simpson's diversity index was estimated by the median of the diversity distribution resulting from 10,000 random samples of a subset without replacement.

	Thymus			Lymph nodes	
i	DP CD3low	SP $CD4^+$	SP $CD8^+$	LN $CD4^+$	LN $CD8^+$
1	79	33	16	34	17
2	17	6	3	8	8
3	6	2	3	2	1
4	5	2	5	1	2
5	1	0	3	0	1
6	1	0	1	0	0
7	1	0	1	0	0
8		0	1	1	0
10		1	0	1	0
11		0	1	0	0
16		1		0	0
20		0		1	0
21		0			1
28		1			0
52					1
M	110	46	34	48	31
n	169	113	98	98	122
D_s	0.99	0.91	0.96	0.94	0.79

was defined at the amino-acid level. No statistical analysis of the data was previously done.

6.1. Information contained in the samples

Before estimating TCR diversity in each T-cell compartment, we perform a preliminary data analysis with the aim of assessing how much information of a repertoire is contained in its samples. With this in mind, we assume that sequences in the original samples were collected by a sequential sampling scheme (i.e., TCR sequences are added sequentially to the samples). In this scenario, we compute a curve that shows how the probability of obtaining a new TCR variant evolves throughout sampling. It is expected that this probability decreases with the number of accumulated sequences in the samples. More importantly, if the samples are representative of an individual's repertoire, the above-mentioned probability should reach zero in the end of the sampling scheme. We also determine the so-called "species accumulation curve" that reflects how the number of distinct TCR sequences (sampled diversity) evolves throughout sampling, as done in several studies (Casrouge et al., 2000; Pacholczyk et al., 2006; Wong et al., 2007). Likewise, the species accumulation curve should reach a plateau when the samples contain a large information of an individual's repertoire. It is worth noting that similar curves have been reported for the diversity estimated by the Homogeneous Poisson model (Casrouge et al., 2000) or by abundance-coverage and incidence-coverage estimators (Pacholczyk et al., 2006).

To compute these different curves, we used Monte Carlo resampling. We first obtain a random order by which TCR sequences are successively added to the samples. For each sampling step, we calculate the proportion of simulations in which the added TCR sequence increased the sampled diversity (Fig. 1A), the mean of sampled diversity (Fig. 1B), and the mean of estimated diversity according to the Homogeneous Poisson model (Fig. 1C).

As expected, the probability of obtaining a new TCR variant is a decreasing function of the number of accumulated sequences (Fig. 1A). However, it is far from zero at the current sample sizes (from 0.15 to 0.55). This demonstrates that the diversity in the samples at hand is not representative of that at the level of the whole repertoire. Similar evidence can be extracted from both sampled and estimated diversity curves (Fig. 1B and C). In fact, these curves are increasing functions of the sample sizes, but not reaching a plateau. It is worth noting that the estimated diversity curves may not be very accurate since the Homogeneous Poisson model does not hold for the current data, as seen below.

6.2. Estimating TCR diversity

Samples of different T cell subpopulations are mostly composed of many distinct TCR sequences with few copies and few sequences with many copies (Table 1). This is most evident in the LN $CD8^+$ T cell data set where a single clonotype, with 52 copies, represents 40% of the sample. A model with a power-law tail, such as the Poisson–Lognormal or Yule, might capture this feature of the data. Sample of DP CD3low thymocytes exhibits a not so long right tail distribution as the ones from the remaining T-cell populations, which might reflect a moderate selection in the early stages of thymocyte development. This is in contrast to what happens in both SP thymocyte samples with several highly abundant TCR sequences. This observation might be explained by strong biases introduced by positive and negative selection in both $CD4^+$ and $CD8^+$ T cell lineages. Another explanation is suggested by the experimental observation that mature T cells may re-enter the thymus (Bosco et al., 2009; Hale and Fink, 2009). Clones selectively expanded in the periphery could appear as large clones in SP populations upon reentry in the thymus. This hypothesis can in theory be assessed by studying the intersection between SP and LN T cell repertoires using the procedures developed by Sepúlveda (2009) while attempting to quantify the intersection between conventional and regulatory $CD4^+$ T-cell compartments. However, this type of analysis is beyond the scope of this paper.

To quantify the diversity we first used Simpson's diversity index as reported before (Venturi et al., 2007). It is defined as

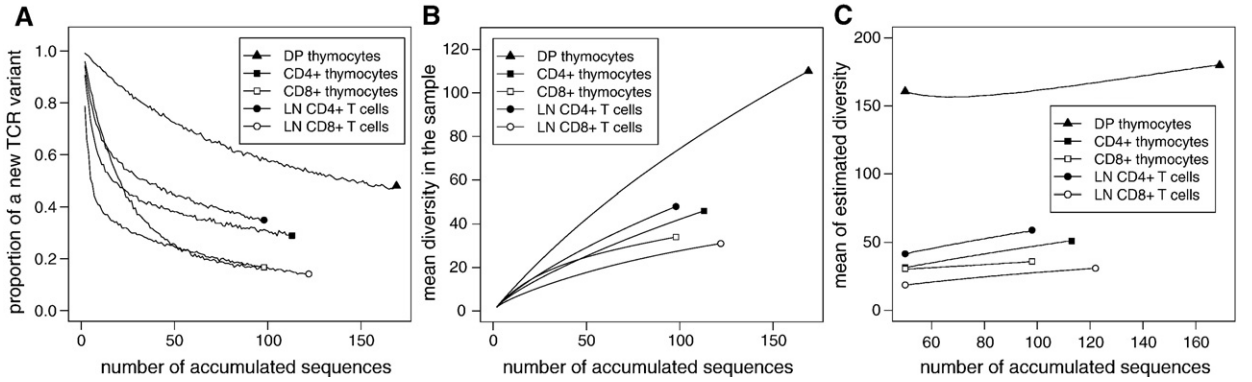

Fig. 1. Assessment of the information contained in the samples assuming that TCR sequences were obtained by a sequential sampling scheme. Proportion of obtaining a new TCR variant in 10,000 simulations (A), mean of sampled diversity (B), and mean of estimated diversity according to the homogeneous Poisson model (C) as a function of the number of TCR sequences accumulated in the samples.

the probability that any two TCR sequences chosen at random from the sample belong to different clonotypes, i.e.,

$$D_S = 1 - \frac{\sum_{i=1}^{n} m_i(m_i-1)}{n(n-1)}. \quad (11)$$

As a probability it ranges from 0 to 1, where 0 means that all TCR sequences belong to the same clonotype (minimal diversity), while 1 indicates that every TCR sequence in the sample is a distinct clonotype (maximal diversity). When one aims to compare this index among samples from different T-cell populations, the current recommended approach advocates the standardization of the sample sizes by the smallest number of TCR sequences in any sample (Venturi et al., 2007). Simpson's diversity index is then applied directly to the smallest sample(s), whereas for the remaining samples it is estimated by the median of the diversity distribution resulting from a random sampling of a subset without replacement. All thymic populations show D_S values close to 1, which suggests almost maximal diversity for these T cell compartments (Table 1). As expected, the sample of the lesser mature DP CD3low population shows the highest value of Simpson's index among all thymic samples. In LNs, $CD4^+$ T cells show an estimate of the index close to that of the thymic SP populations, while $CD8^+$ T cells seem less diverse than the respective thymic SP population according to the same index. This decrease in the estimate of the Simpson's index might be attributed to a great expansion of a few clones in the LNs upon encountering peripheral antigens.

PAMs were then fitted to data (Table 2). According to Pearson's chi-squared test, the DP thymocyte sample can be fitted at 5% significance level by all models with the exception of the Homogeneous Poisson. However, the Poisson–Gamma model led to a quite different diversity estimate for this thymocyte population ($\hat{D}=4697$) when compared to those obtained from other statistically-significant models. It is worth noting that the Geometric model, which is a special case of the former (Eq. (4) with $\alpha=1$), is also not rejected by the data ($p=0.09$), and provides a much lower diversity estimate ($\hat{D}=314$). Yet, its AIC estimate suggests that, in spite of being a simpler model, it does not appear to be more adequate than the Poisson–Gamma itself. Finally, the Poisson–Lognormal and Yule models show a good fit for the data and diversity estimates reasonably in agreement with each other, $\hat{D}=510$ and 625, respectively. Since the Yule model has the smallest AIC among all fitted models, it seems the most appropriate to describe the DP thymocyte sample.

The sample of SP $CD4^+$ thymocytes is well described either by the Poisson–Lognormal or the Yule models, as judged by the *p*-values for Pearson's goodness-of-fit test. However, the former has a significant lower AIC estimate than the latter, suggesting that this model seems formally the best model for the data. Yet, if one considers that some TCR diversity is lost from the DP to SP stage by positive and negative selections (Bouneaud et al., 2000; Ignatowicz et al., 1996; Tourne et al., 1995; Zerrahn et al., 1997), then the Poisson–Lognormal produces an unexpected high diversity estimate (around 17,000). In this line of reasoning, the Yule distribution shows a good fit for the data and the second lowest AIC value, and predicts a lesser diversity estimate of SP $CD4^+$ population than that for the DP population. Thus, this model is considered the most appropriate to describe the SP $CD4^+$ thymocyte data.

Table 2
Diversity estimates of DP CD3low, SP $CD4^+$ and $CD8^+$ thymocytes, and LN $CD4^+$ and $CD8^+$ T cells, and the corresponding AIC measure and *p*-value of Pearson's goodness-of-fit test for different PAMs when fitted to the whole data. *p*-values >0.05 indicate that the respective models are not rejected by the data.

Population	Model	Diversity	*p*-value	AIC
Thymic DP CD3low	Homogeneous Poisson	180	$<10^{-3}$	45.05
	Geometric	314	0.09	28.33
	Poisson–Gamma	4697	0.30	25.70
	Poisson–Lognormal	510	0.32	25.91
	Yule	625	0.52	23.67
Thymic SP $CD4^+$	Homogeneous Poisson	51	$<10^{-3}$	191.08
	Geometric	77	$<10^{-3}$	74.86
	Poisson–Gamma	2388	0.01	53.75
	Poisson–Lognormal	16820	0.69	37.38
	Yule	176	0.30	42.70
Thymic SP $CD8^+$	Homogeneous Poisson	36	$<10^{-3}$	52.55
	Geometric	51	0.10	33.04
	Poisson–Gamma	69	0.07	34.18
	Poisson–Lognormal	49	0.05	35.95
	Yule	93	0.09	35.34
LN $CD4^+$	Homogeneous Poisson	59	$<10^{-3}$	120.89
	Geometric	93	0.02	56.23
	Poisson–Gamma	3074	0.09	43.30
	Poisson–Lognormal	1182	0.76	33.08
	Yule	206	0.54	34.98
LN $CD8^+$	Homogeneous Poisson	31	$<10^{-3}$	284.35
	Geometric	41	$<10^{-3}$	73.72
	Poisson–Gamma	1721	0.02	51.08
	Poisson–Lognormal	260	0.21	40.27
	Yule	91	0.44	40.26

In the case of SP $CD8^+$ thymocytes, TCR diversity estimates appear to be relatively similar in all models and lesser than those of the DP population. All models with the exception, once again, of the Homogeneous Poisson can fit the data at the 5% significance level. Yet, it is worth noting that the quality of the fitting of these models is very close of the borderline for statistical significance, and thus they do not perform particularly well. According to AIC criteria, the Geometric and the Yule models seem to be the most adequate for the data of this thymocyte population.

In both LN populations, the Poisson–Lognormal and Yule models can fit the data well at 5% significance level for Pearson's chi-squared test and show the lowest values for the AIC measure. In the case of LN $CD4^+$ T cells, the Poisson–Gamma model provides also a fairly good fit but shows an intermediate value for the AIC measure and thus should not be considered a good model. The Poisson–Lognormal model predicts a much higher TCR diversity of LN $CD4^+$ T cells than that of their thymocyte counterparts, suggesting an accumulation of $CD4^+$ T cell diversity in the periphery. Nonetheless, it is fair to say that the estimate of 1182 different clones predicted by this model might be explained by the presence of a few highly abundant clones in the LN $CD4^+$ T cell sample, as we will ascertain in the next subsection. With respect to the Yule model, peripheral TCR diversity is estimated more closely to what was estimated in the thymus. For $CD8^+$ T cells, the Poisson–Lognormal model leads once again to a higher TCR diversity estimate than that in thymus, while the Yule predicts a TCR diversity similar to that of SP thymocytes.

6.3. Robustness of TCR diversity estimates

In general, (bio)diversity estimation might be affected by the most abundant species in the sample, typically with more than 10 individuals or copies (Chao et al., 1993). Thus, the next step of the analysis is to evaluate the robustness of diversity estimates coming from samples that have such abundant TCR sequences, namely SP CD4$^+$ thymocytes and both LN CD4$^+$ and CD8$^+$ T cell populations. To this end, we exclude from the data, one by one, the most abundant TCR sequences, and fit again the models and obtain new diversity estimates. In this analysis, one looks for a model that produces reasonably stable diversity estimates as well as a good fit for the data.

In general, the Homogeneous Poisson model cannot describe the data well, even when discarding the most abundant TCR sequences from the analysis (Fig. 2A, B and C). The only exception is the good fit for the LN CD8$^+$ T cell sample, but only for TCR sequences with at most 5 copies (Fig. 2C). Therefore, as expected, a homogeneous clonal size distribution is not realistic. Another model with peculiar

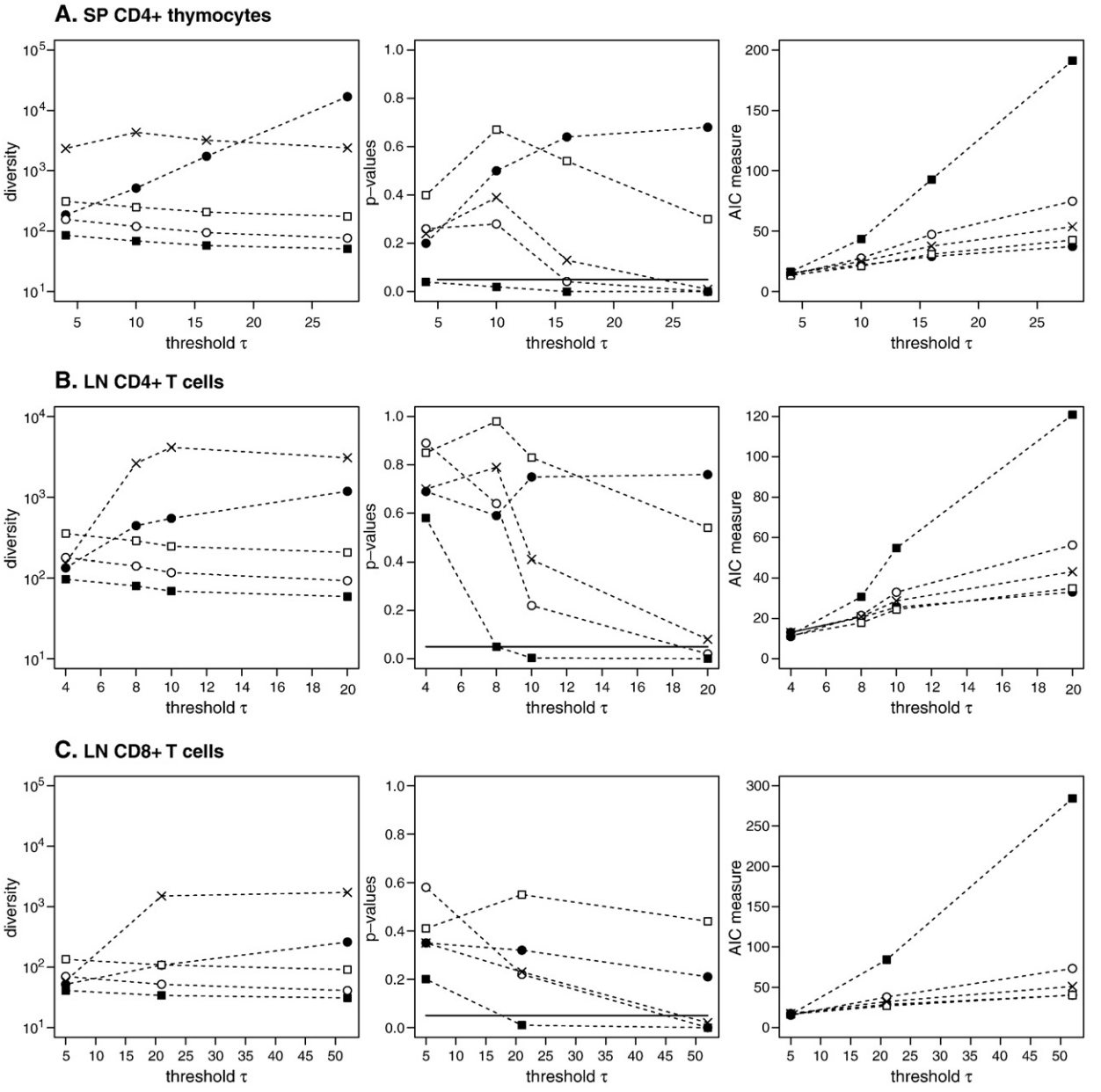

Fig. 2. Robustness of TCR diversity estimates (■ — homogeneous Poisson, ○ — Geometric, × — Poisson–Gamma, ● — Poisson–Lognormal, □ — Yule). The threshold τ in the *x*-axis indicates the maximum number of copies of a TCR sequence allowed in the fitting of the model. In this context, one should look at the plots from right to left, starting from fitting models to the whole data (highest value of τ) followed by the same fitting but to the data excluding successively all TCR sequences with one more than τ copies. *P*-values refer to the Pearson's chi-squared goodness-of-fit test. In this regard, one should select models with *p*-values higher than the 5% significance level (represented by a solid line). Connecting dashed lines are presented just to help in the visualization of the trend of diversity, *p*-values and AIC measure when excluding the most abundant TCR sequences from the data.

results is the Poisson–Gamma, producing the highest TCR diversity estimates in most T cell compartments. This is a consequence of the heavy right tails exhibited by the sampled clonal size distribution, since the diversity estimate tends to decrease dramatically when leaving out the most abundant clonotypes (Fig. 2A, B, and C).

For SP $CD4^+$ thymocytes, the Poisson–Lognormal model can describe the data well according to the *p*-value criterion, with or without the most abundant TCR sequence (Fig. 2A). However, the respective diversity estimates are higher than those of DP thymocytes, and thus this model seems to produce unreliable TCR diversity estimates (Fig. 2A). When the most abundant TCR sequence was not considered in the analysis, the fit of the Yule model improved, and became as adequate as the fit of the Poisson–Lognormal model in terms of the AIC measure. A slight improvement in the Pearson's chi-squared fit was also obtained when additionally excluding the second most abundant clonotype, but not significantly (Fig. 2A). The diversity estimates are always smaller than those of the DP population and reasonably stable (Fig. 2A). With respect to the remaining models, no significant improvement of the Pearson's chi-squared fit was obtained when excluding from the analysis, one by one, the most abundant TCR sequences. For all of this, the best model for the SP $CD4^+$ thymocytes is the Yule, where the most abundant TCR sequence is interpreted as an outlier.

For LN $CD4^+$ T cells, most models can fit the data well if one discards the most abundant TCR sequence from the data set (Fig. 2B). However, as in SP $CD4^+$ thymocytes, the Yule and Poisson–Lognormal models display the highest *p*-values in most cases and, at the same time, the lowest values of AIC. In the latter model, the TCR diversity estimate decreases dramatically without the most abundant TCR sequence but show some stability when discarding further TCR sequences (Fig. 2B). The fit of the Yule model was improved without the most abundant TCR sequence according to the *p*-value for Pearson's chi-squared test, being now the best model with respect to the AIC criterion. Excluding further TCR sequences increases the *p*-value for the goodness-of-fit test but not significantly. In this model, diversity estimates are almost unaffected by the most extreme abundant TCR sequences (Fig. 2B). The Geometric model shows a particular good fit without the three most abundant clonotypes. In fact, this model shows the lowest value of AIC measure in this case. Thus, the Yule and Poisson–Lognormal models seem to describe the data well, but the most abundant TCR sequence might be an outlier. The same can be said for the Geometric model, but the three most abundant TCR sequences should be regarded as putative outliers.

Finally, LN $CD8^+$ T cells are described well by the Yule and Poisson–Lognormal according to Pearson's chi-squared test and AIC measure. Their fits do not dramatically improve when discarding, one by one, the most abundant TCR sequences, while the respective TCR diversity estimates are reasonably stable (Fig. 2C). Therefore, both models should be considered for the whole data. As in the LN $CD4^+$ T cells, the Geometric model also shows a statistically good fit without the three most abundant TCR sequences, being the best regarding the AIC criterion. This model will also be further analyzed.

6.4. Estimating clonal size distributions

All T cell compartments could be explained by the Yule distribution that predicts an appropriate mixture of Exponentials for the clonal size distribution. Another good model was the Poisson–Lognormal that implies a Lognormal distributed clonal sizes. Finally, the Geometric model could describe either LN population and SP $CD8^+$ thymocytes. This model implies Exponential distributed clonal sizes.

To transform the Poisson sampling rate distribution into the respective clonal size distribution, one needs to know the total number of cells in the different T cell compartments. In this mouse line, the thymus contain about 200 million thymocytes, of which DP CD3low, SP CD4+ and SP $CD8^+$ represent approximately 95%, 0.1% and 0.25%, respectively. LNs contain in total around 15 million T cells (data provided by the authors of the original reference). As shown in Correia-Neves et al. (2001), the frequency of peripheral $CD4^+$ and $CD8^+$ T cells changes in time. At week 6, when samples were collected, $CD4^+$ and $CD8^+$ T cells represent approximately 4% and 10% of total LN cells, respectively. Table 3 summarizes all relevant information.

In all T-cell populations, the clonal size distributions have the common feature of being highly skewed to the right (Fig. 3 and 4 A and C). As a consequence, the median and mean values provide different summaries of the underlying clonal size distributions. However, the differences between these two summary measures are in most cases less than one order of magnitude (Table 3). The best-fitted models entail different TCR diversity estimates. This difference is reflected

Table 3
Mean, median and standard deviation (SD) of clonal size distributions for selected PAMs: N is the total number of cells in the respective population and $\hat{D}$ is the diversity estimate. The number of outlier TCR sequences is also given as well as their contribution to the respective T cell compartment (in parentheses).

Population	N [a]	Model	$\hat{D}$	Mean [a]	Median [a]	SD [a]	Outliers
Thymic DP CD3low	8.28	Poisson–Lognormal	510	5.58	5.22	5.89	0 (0%)
		Yule	625	5.48	5.02	5.81	0 (0%)
Thymic SP $CD4^+$	5.18	Yule	207	2.83 [b]	2.34 [b]	3.18 [b]	1 (25%)
Thymic SP $CD8^+$	5.70	Geometric	51	3.99	3.83	3.99	0 (0%)
		Yule	93	3.83	3.17	4.58	0 (0%)
LN $CD4^+$	5.78	Geometric	178	3.31 [b]	3.16 [b]	3.31	3 (39%)
		Poisson–Lognormal	545	2.92 [b]	2.27 [b]	3.55 [b]	1 (20%)
		Yule	245	3.28 [b]	2.80 [b]	3.64 [b]	1 (20%)
LN $CD8^+$	6.18	Geometric	70	3.94 [b]	3.78 [b]	3.94 [b]	2 (60%)
		Poisson–Lognormal	260	3.74	2.42	5.06	0 (0%)
		Yule	91	4.10	3.49	4.73	0 (0%)

[a] Base 10 logarithm.
[b] Estimates without outliers.

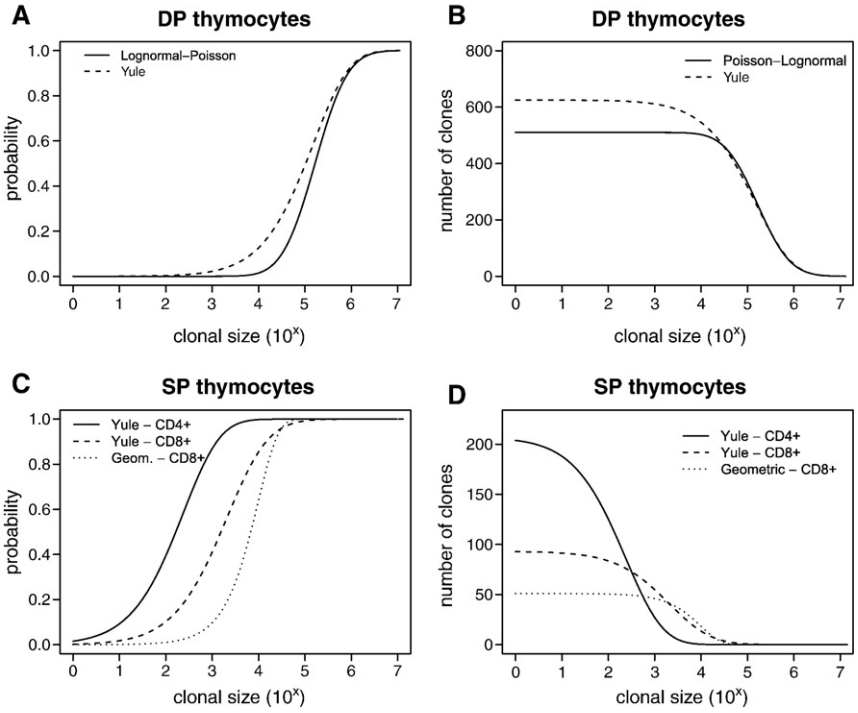

Fig. 3. Clonal size cumulative distribution functions for thymocyte populations (A and C) — those predicted by the Yule model were obtained by Monte Carlo method with 10,000 simulations. Expected number of clones with clonal size above a certain value (B and D).

in the clonal size distribution by rarer clonotypes in models with high TCR diversity estimates (Fig. 3B and D for the thymocyte populations and Fig. 4B and D for LN T cells). This is most evident for LN $CD8^+$ T cells when comparing the Yule and the Poisson–Lognormal models, where the latter provides higher TCR diversity estimates than the former

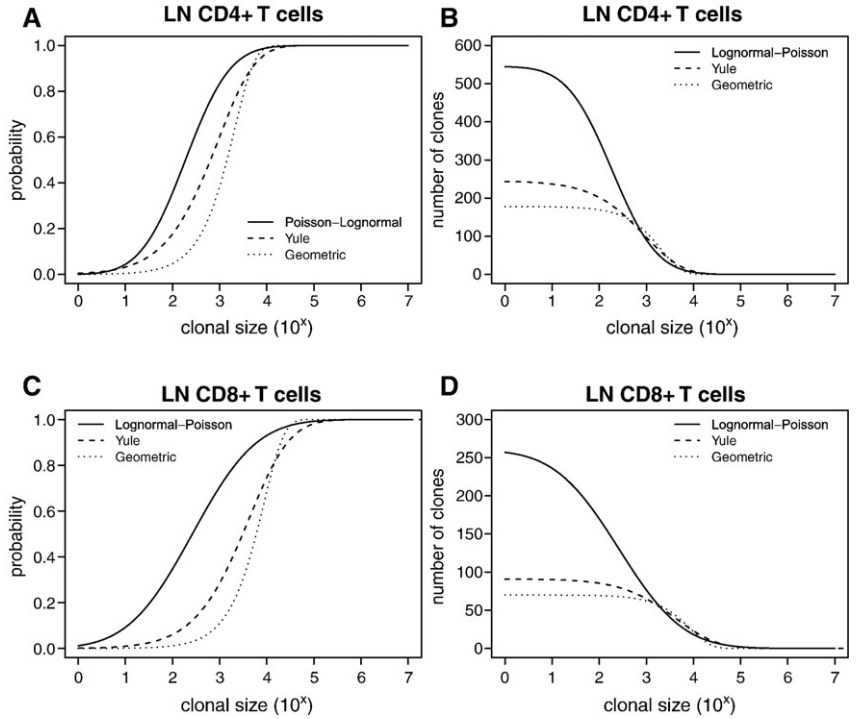

Fig. 4. Clonal size cumulative distribution functions for LN T cell populations (A and C) — those predicted by the Yule model were obtained by Monte Carlo method with 10,000 simulations. Expected number of clones with clonal size above a certain value (B and D).

(Fig. 4D). In spite of this difference, a similar number of extremely abundant clonotypes is predicted from all models.

6.5. Discriminating the models requires large samples

As seen earlier, each T-cell compartment could be well fitted by more than one PAM. For example, data from LN CD4$^+$ T cells are reasonably described either by Geometric, Yule or Poisson–Lognormal (Table 3). This raises the question: What is then the "true" model?

A straightforward way of addressing this question would be to increase the sample size. In theory, it is expected that only the true model would hold when compared to a sufficiently large data set. With this in mind, we performed a simulation study to determine the sample size required to discriminate the above-mentioned models for the LN CD4$^+$ T cells. To this end, we simulated data from each model using as true diversity the corresponding estimate of the original data set (Table 3). We then fitted the remaining models to the simulated data, and test their goodness-of-fit according to Pearsons's chi-squared test. Since PAMs assume a Poisson sampling scheme, the sample size is not fixed when simulating a data set from the repertoire. To increase the sample size, we vary the parameters of each model in order to increase the respective sampling rates of each TCR variant in the repertoire. We then divided the simulated sample sizes in distinct categories, calculating in each one the proportion of simulations that reject each model at 5% significance level (Fig. 5A, B and C). As a control of the simulations, we confirmed that the model that generated the data was rejected in approximately 5% of the generated data sets (Fig. 5A, B and C). To complement this analysis, we also compute the median of the respective distribution of estimated diversities in each sample size category. Again as demonstration of the internal consistency of the simulations, the median of estimated diversities obtained by fitting the true model agrees with the TCR diversity specified during data generation (Fig. 5D, E and F).

In general, large sample sizes are required to reject all wrong models (Fig. 5). This is clearly shown in the Yule-generated data, where the Poisson–Lognormal model can be at best rejected in 20% of the simulated samples with nearly 10,000 TCR sequences (Fig. 5C). This result might be explained by two facts. On one hand, it is known that the Lognormal distribution can give good fit even when the totality of the left tail of the Lognormal distribution is absent from the data, the so-called "veiled line" phenomenon (Dewdney, 1998). On the other hand, the right tails of Yule and Poisson–Lognormal distributions are approximately power-laws. Putting both facts together the great acceptance proportion of the Poisson–Lognormal might reflect a sort of veiled-line phenomenon, where one can only infer the power-law-like right tail of the Lognormal sampling rate distribution from the data. Notwithstanding, the TCR diversity predicted by Poisson–Lognormal is usually an overestimate of the true diversity, being worse when increasing the sample size (Fig. 5F). In contrast, 8000 TCR sequences are sufficient to completely reject the Yule model for data generated from the Poisson–Lognormal (Fig. 5B). This might be due the fact that increasing the sample size leads to a larger information on the whole Lognormal sampling rate distribution, specially its left tail, which cannot be captured

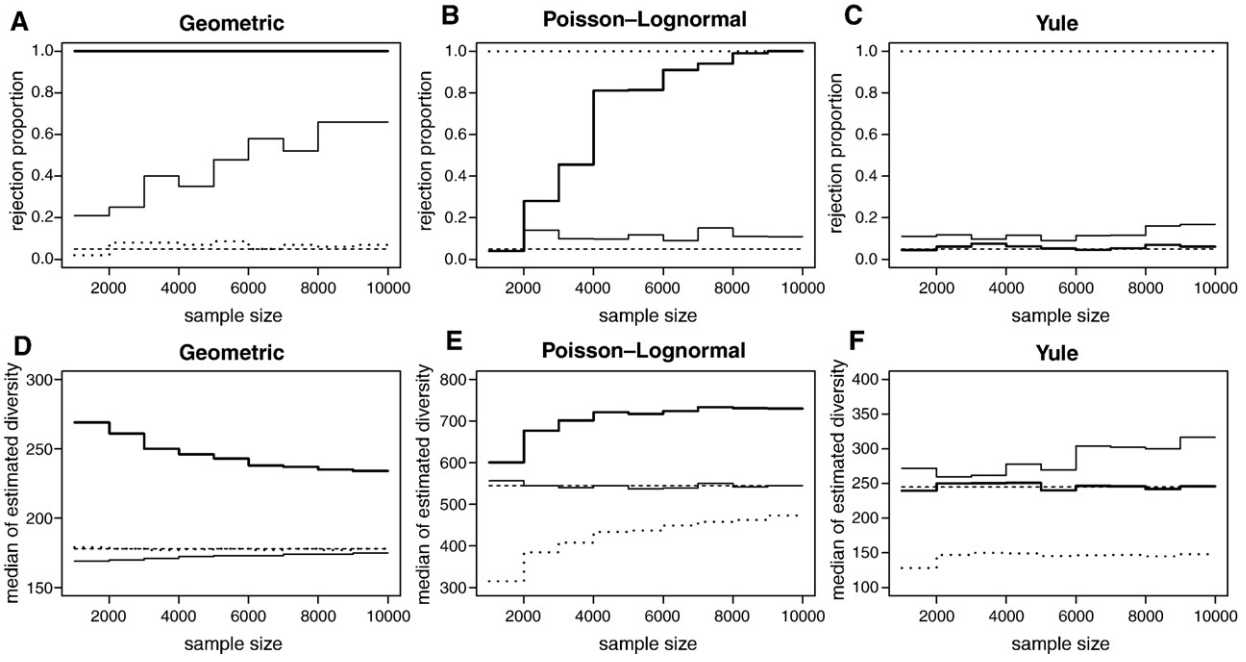

Fig. 5. Large samples are required to discriminate PAMs from each other when simulating data from LN CD4$^+$ T cell compartment (Geometric model — dotted line; Poisson–Lognormal model — thin solid line; Yule model — thick solid line). (A and C) Data generated from the Geometric model (with $\hat{D}=178$); (B and D) Data generated from the Poisson–Lognormal model (with $\hat{D}=545$); (C and F) Data generated from the Yule model (with $\hat{D}=245$). The rejection proportion is defined by the fraction of simulations that rejected a model by the Pearson's chi-squared test at 5% significance level. As a control of the simulations, the significance level is shown (dashed lines in A, B, and C) as well as the "true" diversity (dashed lines in D, E, and F). Each step in the plots was obtained according to 100 simulated samples within each sample size category.

by the Yule distribution. Moreover, Yule-based TCR diversity estimates tend to be higher than the underlying true diversity (Fig. 5E). However, when the true model is either Yule or Poisson–Lognormal, samples of 1000 TCR sequences seem sufficient to reject a Geometric distribution (Fig. 5B and C). This is related to the fact that Yule and Poisson–Lognormal have power-law right tails that are much heavier than the exponential one associated with the Geometric model. Finally, the Poisson–Lognormal model is wrongly accepted in a high proportion of data sets generated from a Geometric distribution, but provides good estimates for the true TCR diversity (Fig. 5D). In contrast, Geometric-generated data with sample sizes higher than 1000 TCR sequences cannot be captured by the Yule distribution.

Facing the above results, we recommend that future studies on TCR repertoire should contemplate a large number of TCR sequences in the samples. Only in this way one could obtain a clear insight on the shape of the TCR repertoire. However, it is worth mentioning that these large sample sizes should not be obtained by pooling data together from different animals, as often done in TCR repertoire studies (Lathrop et al., 2008; Pacholczyk et al., 2006; Pacholczyk et al., 2007). On one hand, TCR diversity estimates of pooled data tend to be an overestimate of that for each individual TCR repertoire by the presence of TCR variants belonging to each individual exclusive repertoire in the same data set. On the other hand, pooled data provide meaningless inferences over clonal size distribution of an individual TCR repertoire. Specifically, by pooling data together, one is artificially increasing the clonal sizes of clonotypes shared among each individual T cell population, giving rise to a clonal size distribution that might not resemble those from each individual.

To illustrate these problems, we pooled together LN $CD4^+$ data from three distinct animals and fit different PAMs to them (Table 4). As expected, higher diversity estimates are predicted when analyzing pooled data even for models that are rejected by the data. The most dramatic effect is observed in the Poisson–Lognormal model, which can fit well every mouse data as well as pooled data but with different parameters. In this model, diversity estimates for each animal range approximately from 1200 to 7500, as opposed to roughly 32,000 different TCR sequences predicted for the pooled data, which is much higher than the sum of diversities expected for each mouse repertoire. Interestingly, the Yule model can fit well every individual mouse data but not the pooled data. This last result shows that pooling data may give rise to a different clonal size distribution than those for each individual TCR repertoire.

7. Discussion

In recent years several TCR diversity estimates have been put forward in reports comparing the repertoires of T-cell subsets (Ferreira et al., 2009; Hsieh et al., 2004; Hsieh et al., 2006; Lathrop et al., 2008; Pacholczyk et al., 2006; Pacholczyk et al., 2007; Wong et al., 2007). A common feature in these studies is an incipient statistical analysis, such that diversity estimates are typically provided without proper evaluation of their (statistical) quality. Another drawback of these analyses is that they do not assess the underlying clonal size distribution and how it impinges on the diversity estimates. In this work, several PAMs were presented and used to estimate TCR diversity and clonal size distributions in mice with limited TCR diversity based on a maximum likelihood framework. By fitting these models to these data, we have now a better quantitative picture of the T-cell repertoire in mice with limited TCR diversity. After a productive imprecise joining of V and J of the α locus, DP CD3low have no more than 600 distinct TCR sequences. With positive and negative selection, this number drops to less than 100 in both SP populations, even though SP $CD4^+$ thymocytes seem more diverse than their $CD8^+$ counterparts. In the periphery, TCR diversity seems to accumulate in both $CD4^+$ and $CD8^+$ T-cell compartments, since these peripheral populations appear to be more diverse than their thymic counterparts. There are only a few clones that will expand greatly in the periphery, due to a better usage of resources, perhaps during immune responses. In this regard, samples of $CD8^+$ T cells are more prone to exhibit clonotypes with large clonal sizes than those from $CD4^+$ T cells, in agreement with what was previously described during viral infections (Maini et al., 2000).

We showed that some data sets can be fitted by more than one model, notably the Geometric, Yule or Poisson–Lognormal models for LN $CD4^+$ T-cell compartment. In this case, the theory developed here indicates that the clonal sizes could be distributed as an Exponential, an appropriate mixture of Exponentials, or a Lognormal. Facing this result, it is fair to wonder what is the true model. A definitive answer to this question cannot be drawn from the data due to limited sample sizes, but our simulation study indicates that sample sizes should be large and not less than 10,000 independent TCR sequences. Sample sizes of this order of magnitude were prohibitive using the experimental techniques of the original reference but they are now within reach using new high-throughput sequencing methods (Freeman et al., 2009; Holt and Jones, 2008; Shendure and Ji, 2008). Anticipating the availability of these large data sets and the opportunities that

Table 4
Diversity estimates $\hat{D}$ of LN $CD4^+$ T cells in three mice and in the respective pooled data, where p is the corresponding p-value of Pearson's goodness-of-fit test for each PAM. One should select models with p-values higher than 0.05. Sample sizes for each mouse and pooled data are 98 (mouse 57), 69 (mouse 63), 98 (mouse 91) and 212 (pooled). The respective numbers of distinct TCR variants in the samples are 48, 33, 28 and 95.

Model	Mouse 57		Mouse 63		Mouse 91		Pooled	
	$\hat{D}$	p	$\hat{D}$	p	$\hat{D}$	p	$\hat{D}$	p
Homogeneous Poisson	59	$<10^{-2}$	39	$<10^{-2}$	42	$<10^{-2}$	111	$<10^{-2}$
Geometric	93	0.02	62	0.01	73	0.03	171	$<10^{-2}$
Poisson–Gamma	3074	0.08	2046	0.02	2615	0.06	4978	$<10^{-2}$
Yule	206	0.54	153	0.23	158	0.18	385	0.01
Poisson–Lognormal	1182	0.76	7619	0.10	2998	0.29	32342	0.46

they will bring, we developed and provided the package PAM for the R software. As it stands, the package could handle the large simulated data sets generated with relatively small diversity as the ones estimated from the small samples. It may turn out that the maximum likelihood algorithms need to be further optimized to process large samples obtained from repertoires that are 100-fold more diverse than the estimates presented here. Yet, in this case, we might need not only a better software but even larger samples. It is an open question whether or not wild type animals and humans will be in this more difficult scenario.

The quantitative characterization of clonal size distributions afforded by large samples will in turn give information on the mechanisms underlying the selection and maintenance of the T-cell repertoire. PAMs may also be a good tool here since in Ecology the Lognormal distribution of species abundances or mixtures of Exponential distributions as the ones identified above have mechanistic interpretations. Thus, to understand the origin of these distributions in the T-cell repertoire, one might conceptualize the immune system as a community divided into niches (Bautista et al., 2009; Carneiro et al., 1995; Freitas and Rocha, 2000; Leung et al., 2009; Stephens et al., 2007). A niche can be defined either as an anatomical site or as the availability of a certain resource. Different mechanisms of niche apportionment between T cell "species" entail different clonal size distributions.

The Exponential clonal size distribution can be derived from the "broken stick" model (MacArthur, 1957). In this model, a community is compared to a stick with unit length. If there are S species in a community, the stick is broken simultaneously into S pieces, the niches, each one occupied by a single species. The relative abundance of the species is proportional to the lengths of the segments they belong to. The simultaneous breaking of the stick implies that the niches are already predefined in the community. It is known that the architectures of thymic microenvironments and peripheral lymphocyte environment change with time (van Ewijk et al., 1999; Freitas and Rocha, 2000). Therefore, a static definition of an immunological niche is rather difficult to envisage. For all of this, the "broken stick" model might not be a good candidate mechanism for the T-cell repertoire.

A more realistic scenario is the so-called sequential "broken stick" model (Bulmer, 1974; Sugihara, 1980) in which new niches are created by fragmenting the preexisting ones in a series of stages. When the number of stages is sufficiently large, the clonal size distribution tends to be Lognormal. This sequential niche apportionment is reasonable for the periphery, since the arrival of new clones from the thymus can be regarded as the stages of the model. For the thymus, this might be also true, since it is known that thymocytes affect the microenvironment that supports them (van Ewijk et al., 1999). Thus, new thymic microenvironments might be created with the production of cells with new TCR sequences during the V(D)J recombination process. It is worth noting that other mechanisms generating a Lognormal distribution were put forward (Diserud and Engen, 2000; Engen and Lande, 1996), but their immunological interpretation is not so straightforward.

One can also extend the "broken-stick" model by breaking the stick first into a certain number of large pieces, superniches, that will be further divided according to their size. Thus, larger superniches will tend to be more fragmented than small ones. We expect that this two-stage niche apportionment leads to a mixture of Exponential distributions for the clonal size distribution, as the one predicted by the Yule model. A superniche might be defined by the expression of certain homing receptors, such as CCR4 and CCR7 (Sallusto et al., 1999; Tang and Cyster, 1999), or by a certain motif in the TCR sequence (Wallace et al., 2000).

This discussion on the rationale and interpretation of the clonal size distributions puts the emphasis on the necessity to better define and characterize immunological niches experimentally. Lathrop et al. (2008) have made a step in this direction by studying the repertoire in different anatomical locations. However, one can anticipate that characterizing static or sequential niche apportionment processes would require not only the study of the T-cell repertoire at different anatomical locations but at different time points.

As a concluding remark it is worth emphasizing that the analysis of TCR data does not end with the estimation of diversity and clonal size distributions. The next step is to quantify the intersection among repertoires of different T-cell compartments. The well-known Morisita–Horn similarity index, which takes into account not only the number of clonotypes shared by several samples but also their clonal sizes, has been used before (Venturi et al., 2008). Yet, this measure is a summary statistic and thus does not truly quantify the intersection at the level of the repertoires. In theory, the PAMs might be further extended to tackle this inferential problem. To this end, models might include parameters that reflect how clonal sizes of the clonotypes belonging to the intersection set are related to each other. The Poisson–Lognormal model seems to be the easiest candidate to such extension, since the relationship between clonal sizes could be modeled directly by correlation coefficients, as illustrated by Sepúlveda (2009). The Yule model can also be extended to the two-sample case, as done by Xekalaki (1986), but restricted to a scenario in which each sample follows the same Yule distribution. For the remaining models, the relationship between clonal sizes requires an extra level of modeling. Thus, additional work needs to be done to harvest the full potential of PAMs in the analysis of TCR repertoire data.

Acknowledgments

The authors are grateful to José Faro for the challenging discussions that originated this work and to Margarida Correia–Neves for the invaluable insights on the experimental data. The authors would like also to thank Rui Gardner and Eurico de Sepúlveda for reading the manuscript. This work was supported by Fundação para a Ciência e Tecnologia through grants (POCI/SAU-MMO/60333-2004 and PTDC/SAU-MII/71402-2006) and a fellowship to Nuno Sepúlveda (SFRH/BD/19810).

References

Agresti, A., 1992. A survey of exact inference for contingency tables. Stat. Sci. 7, 131.
Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19, 716.

Arstila, T.P., Casrouge, A., Baron, V., Even, J., Kanellopoulos, J., Kourilsky, P., 1999. A direct estimate of the human αβ T cell receptor diversity. Science 286, 958.
Barabási, A.L., Albert, R., 1999. Emergence of scaling in random networks. Science 286, 509.
Barth, R.K., Kim, B.S., Lan, N.C., Hunkapiller, T., Sobieck, N., Winoto, A., Gershenfeld, H., Okada, C., Hansburg, D., Weissman, I.L., Hood, L., 1985. The murine T-cell receptor uses limited repertoire of expressed V_β gene segments. Nature 316, 517.
Bautista, J.L., Lio, C.W., Lathrop, S.K., Forbush, K., Liang, Y., Luo, J., Rudensky, A.Y., Hsieh, C.S., 2009. Intraclonal competition limits the fate determination of regulatory T cells in the thymus. Nat. Immunol. 10, 610.
Behlke, M.A., Spinella, D.G., Chou, H.S., Sha, W., Hartl, D.L., Loh, D.Y., 1985. T-cell receptor β-chain expression: dependence on relatively few variable region genes. Science 229, 566.
Bosco, N., Kirberg, J., Ceredig, R., Agens, F., 2009. Peripheral T cells in the thymus: have they just lost their way or do they do something? Imm. Cell Biol. 87, 50.
Bouneaud, C., Kourilsky, P., Bousso, P., 2000. Impact of negative selection on the T cell repertoire reactive to a self-peptide: a large fraction of T cell clones escapes clonal deletion. Immunity 13, 829.
Bulmer, M.G., 1974. On fitting the Poisson Lognormal distribution to species-abundance data. Biometrics 30, 101.
Carneiro, J., Stewart, J., Coutinho, A., Coutinho, G., 1995. The ontogeny of class-regulation of CD4+ T lymphocyte populations. Int. Immunol. 7, 1265.
Casrouge, A., Beaudoing, E., Dalle, S., Pannetier, C., Kanellopoulos, J., Kourilsky, P., 2000. Size of the αβ TCR repertoire of naive mouse splenocytes. J. Immunol. 164, 5782.
Chao, A., Bunge, J., 2002. Estimating the number of species in a stochastic abundance model. Biometrics 58, 531.
Chao, A., Lee, S.-M., 1992. Estimating the number of classes via sample coverage. J. Am. Stat. Assoc. 87, 210.
Chao, A., Ma, M.-C., Yang, M.C.K., 1993. Stopping rule and estimation for recapture debugging with unequal detection rates. Biometrika 80, 193.
Correia-Neves, M., Waltzinger, C., Mathis, D., Benoist, C., 2001. The shaping of the T cell repertoire. Immunity 14, 21.
Davis, M.M., Bjorkman, P.J., 1988. T-cell antigen receptor genes and T-cell recognition. Nature 334, 395.
Dewdney, A.K., 1998. A general theory of the sampling process with applications to the veil line. Theor. Popul. Biol. 54, 294.
Diserud, O.H., Engen, S., 2000. A general and dynamic species abundance model, embracing the lognormal and the gamma models. Am. Nat. 155, 497.
Engen, S., Lande, R., 1996. Population dynamic models generating the lognormal species abundance distribution. Math. Biosci. 132, 169.
Ferreira, C., Singha, Y., Furmanskia, A.L., Wong, F.S., Gardena, O.A., Dysona, J., 2009. Non-obese diabetic mice select a low-diversity repertoire of natural regulatory T cells. Proc. Natl. Acad. Sci. U. S. A. 106, 8320.
Fisher, R.A., Corbet, A.S., Williams, C.B., 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. J. Anim. Ecol. 12, 42.
Freeman, J.D., Warren, R.L., Webb, J.R., Nelson, B.H., Holt, R.A., 2009. Profiling the T-cell receptor β-chain repertoire by massively parallel sequencing. Genome Res. 19, 1817.
Freitas, A.A., Rocha, B., 2000. Population biology of lymphocytes: the flight for survival. Annu. Rev. Immunol. 18, 83.
Hale, J.S., Fink, P.J., 2009. Back to the thymus: peripheral T cells come home. Imm. Cell Biol. 87, 58.
Holt, R.A., Jones, S.J., 2008. The new paradigm of flow cell sequencing. Genome Res. 18, 839.
Hsieh, C.-S., Liang, Y., Tyznik, A.J., Self, S.G., Liggitt, D., Rudensky, A.Y., 2004. Recognition of the peripheral self by naturally arising $CD25^+$ $CD4^+$ T cell receptors. Immunity 21, 267.
Hsieh, C.-S., Zheng, Y., Liang, Y., Fontenot, J.D., Rudensky, A.Y., 2006. An intersection between the self-reactive regulatory and nonregulatory T cell receptor repertoires. Nat. Immunol. 7, 401.
Ignatowicz, L., Kappler, J., Marrack, P., 1996. The repertoire of T cells shaped by a single MHC/peptide ligand. Cell 84, 521.
Janeway, C.A., Travers, P., Walport, M., Shlomchik, M.J., 2005. Immunobiology. Garland, New York.
Jones, J., Handcock, M.S., 2003. An assessment of preferential attachment as a mechanism for human sexual network formation. Proc. Roy. Soc. London B 270, 1123.
Kedzierska, K., Venturi, V., Field, K., Davenport, M.P., Turner, S.J., Doherty, P.C., 2006. Early establishment of diverse T cell receptor profiles for influenza-specific $CD8^+$ CD62Lhi memory T cells. Proc. Natl. Acad. Sci. USA 103, 9184.
Lathrop, S.K., Santacruz, N.A., Pham, D., Luo, J., Hsieh, C.-S., 2008. Antigen-specific peripheral shaping of the natural regulatory T cell population. J. Exp. Med. 205, 3105.
Leung, M.W.L., Shen, S., Lafaille, J.L., 2009. TCR-dependent differentiation of thymic Foxp3+ cells is limited to small clonal sizes. J. Exp. Med. 206, 2121.
Levich, A.P., 1980. A structure of ecological communities. Moscow University Press, Moscow.
MacArthur, R.H., 1957. On the relative abundance of bird species. Proc. Nat. Acad. Sci. USA 43, 293.
Magurran, A.E., Henderson, P.A., 2003. Explaining the excess of rare species in natural species abundance distributions. Nature 422, 714.
Maini, M.K., Gudgeon, N., Wedderburn, L.R., Rickinson, A.B., Beverley, P.C.L., 2000. Clonal expansions in acute EBV infection are detectable in the CD8 and not the CD4 subset and persist with a variable CD45 phenotype. J. Immunol. 165, 5729.
McGill, B.J., 2003. A test of the unified neutral theory of biodiversity. Nature 442, 881.
Naumov, Y.N., Naumova, E.N., Hogan, K.T., Selin, L.K., Gorski, J., 2003. A fractal clonotype distribution in the $CD8^+$ memory T cell repertoire could optimize potential for immune responses. J. Immunol. 170, 3994.
Naylor, K., Li, G., Vallejo, A.N., Lee, W.W., Koetz, K., Bryl, E., Witkownski, J., Fullbright, J., Weyand, C.M., Goronzy, J.J., 2005. The influence of age on T cell generation and TCR diversity. J. Immunol. 70, 7446.
Pacholczyk, R., Ignatowicz, H., Kraj, P., Ignatowicz, L., 2006. Origin and T cell receptor diversity of $Foxp3^+$ $CD4^+$ $CD25^+$ T cells. Immunity 25, 249.
Pacholczyk, R., Kern, J., Singh, N., Iwashima, M., Kraj, P., Ignatowicz, L., 2007. Nonself-antigens are the cognate specificities of $Foxp3^+$ regulatory T cells. Immunity 27, 493.
Pewe, L.L., Netland, J.M., Heard, S.B., Perlman, S., 2004. Very diverse CD8 T cell clonotypic responses after virus infections. J. Immunol. 172, 3151.
Preston, F.W., 1948. The commonness and rarity of species. Ecology 29, 254.
Sallusto, F., Lenig, D., Forster, R., Lipp, M., Lanzavecchia, A., 1999. Two subsets of memory T lymphocytes with distinct homing potentials and effector functions. Nature 401, 708.
Sanathanan, L., 1972. Estimating the size of a multinomial population. Ann. Math. Stat. 43, 142.
Sepúlveda, N., 2009. How is the T-cell repertoire shaped? Ph.D. thesis, University of Oporto, Oporto, Portugal.
Shendure, J., Ji, H., 2008. Next-generation DNA sequencing. Nat. Biotechnol. 26, 1135.
Stephens, G.L., Andersson, J., Shevach, E., 2007. Distinct subsets of $FoxP3^+$ regulatory T cells participate in the control of immune responses. J. Immunol. 178, 6901.
Sugihara, G., 1980. Minimal community structure: an explanation of species abundance patterns. Am. Nat. 116, 770.
Tang, H.T., Cyster, J.G., 1999. Chemokine up-regulation and activated T cell attraction by maturating dendritic cells. Science 284, 819.
Tourne, S., Naoko, N., Viville, S., Benoist, C., Mathis, D., 1995. The influence of invariant chain on the positive selection of single T cell receptor specificities. Eur. J. Immunol. 25, 1851.
van Ewijk, W., Wang, B.-P., Hollander, G., Kawasoto, H., Spanopoulou, Itoi, M., Amagai, T., Jiang, Y.-F., Germeraad, W. T., Chen, W.-F., Katsura, Y., 1999. Thymic microenvironments, 3-D versus 2-D? Semin. Immunol. 11, 57.
Venturi, V., Kedzierska, K., Tanaka, M.M., Turner, S.J., Doherty, P.C., Davenport, M.P., 2008. Method for assessing the similarity between subsets of the T cell receptor repertoire. J. Immunol. Methods 329, 67.
Venturi, V., Kedzierska, K., Turner, S.J., Doherty, P.C., Davenport, M.P., 2007. Methods for comparing the diversity of samples of the T cell receptor repertoire. J. Immunol. Methods 321, 182.
Wallace, M.E., Bryden, M., Cose, S.C., Coles, R.M., Schumacher, T.N., Brooks, A., Carbone, F.R., 2000. Junctional biases in the naive TCR repertoire control the CTL response to an immunodominant determinant of HSV-1. Immunity 12, 547.
Wong, J., Obst, R., Correia-Neves, M., Losyev, G., Mathis, D., Benoist, C., 2007. Adaptation of TCR repertoires to self-peptides in regulatory and nonregulatory $CD4^+$ T cells. J. Immunol. 178, 7032.
Xekalaki, E., 1986. The bivariate Yule distribution and some of its properties. Statistics 17, 311.
Yule, G., 1925. A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis FRS. Philos. Trans. Roy. Soc. London Ser. B - Biol. Sci. 213, 21.
Zerrahn, J., Held, W., Raulet, D.H., 1997. The MHC reactivity of T cell repertoire prior to positive and negative selection. Cell 88, 627.