



Bayesian multivariate Poisson abundance models for T-cell receptor data

Joshua Greene^a, Marc R. Birtwistle^b, Leszek Ignatowicz^c, Grzegorz A. Rempala^{d,*}

^a Department of Biostatistics, Georgia Health Sciences University, Augusta, GA 30912, USA

^b Department of Pharmacology and Systems Therapeutics, Mount Sinai School of Medicine, New York, NY 10029, USA

^c Center for Biotechnology and Genomic Medicine, Georgia Health Sciences University, Augusta, GA 30912, USA

^d Division of Biostatistics, The Ohio State University, Columbus, OH 43210, USA

ARTICLE INFO

Article history:

Received 5 June 2012

Received in revised form

2 November 2012

Accepted 14 February 2013

Available online 1 March 2013

Keywords:

T-cell antigen receptors

Species diversity estimation

Poisson abundance models

Lognormal distribution

MAP estimation

ABSTRACT

A major feature of an adaptive immune system is its ability to generate B- and T-cell clones capable of recognizing and neutralizing specific antigens. These clones recognize antigens with the help of the surface molecules, called antigen receptors, acquired individually during the clonal development process. In order to ensure a response to a broad range of antigens, the number of different receptor molecules is extremely large, resulting in a huge clonal diversity of both B- and T-cell receptor populations and making their experimental comparisons statistically challenging. To facilitate such comparisons, we propose a flexible parametric model of multivariate count data and illustrate its use in a simultaneous analysis of multiple antigen receptor populations derived from mammalian T-cells. The model relies on a representation of the observed receptor counts as a multivariate Poisson abundance mixture (*m* PAM). A Bayesian parameter fitting procedure is proposed, based on the complete posterior likelihood, rather than the conditional one used typically in similar settings. The new procedure is shown to be considerably more efficient than its conditional counterpart (as measured by the Fisher information) in the regions of *m* PAM parameter space relevant to model T-cell data.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Recent advances in experimental molecular genetics, such as multigene, single cell quantitative PCR assays, DNA microarrays, or high-throughput RNA sequencing (Wang et al., 2009), have challenged the traditional approaches to analyze biological data and pushed the boundaries of modern statistical science. This seems especially the case in the field of immunology, where it is now possible to massively sample the *antigen receptors* expressed by the body's lymphocytes known as T-cells (or, similarly, B-cells) and designed to help them recognize foreign, antibody-generating molecules or *antigens*. In principle, the immune responses to antigens could be evaluated quantitatively by examining the counts of different T-cell clonotypes collected into the samples, according to their unique T-cell receptors (TCRs). However, in view of the potentially enormous diversity (i.e., the number of distinct clonotypes) of the underlying populations and the complicated data collection process (Correia-Neves et al., 2001; Wang et al., 2010), a method relying solely on the empirically observed counts may be very unreliable, due to the under-sampling bias and the frequent clonotype miscalls in the sampling procedures (cf., e.g., Mamedov et al., 2011). Whereas the clonotype miscalls, which

typically occur due to the mechanistic sequencing errors, can often be corrected via the post-processing procedures (see, e.g., Bolotin et al., 2012), the under-sampling of the TCR clonotypes presents a statistical challenge, and seems to require a special approach. Motivated by the experimental data arriving from TCR populations in mammalian studies, the current paper presents one possible solution, in a form of a unified, model-based statistical methodology designed to comprehensively address the issue of TCR profiling.

In order to appreciate the difficulty of sampling TCR populations, let us briefly recall that the T-cell receptors are surface molecules of heterodimer proteins with two chains: α and β (in $\alpha\beta$ T-cells) or γ and δ (in $\gamma\delta$ T-cells). The genes encoding these proteins are generated by the so-called V(D)J DNA recombination in the TCR *variable domain* during thymic T-cell development. In this process, the T-cell precursors or thymocytes randomly recombine different V (variable), D (diverse), and J (joint) gene segments and assemble the mature gene encoding a specific TCR chain. By enumerating all such possible recombinations, one concludes that there are possibly 10^{18} distinct TCR chains in humans (Murphy et al., 2011) and 10^{15} in mice (Davis and Bjorkman, 1988). This enormous size of a TCR repertoire (i.e., a collection of distinct TCR molecules or clonotypes) enables the immune system to identify a vast number of antigens by complementing their molecular shapes in the so-called *complementarity determining regions* or CDRs. The CDRs determine the T-cell affinity and specificity for particular antigens. In the amino acid

* Corresponding author. Tel.: +1 614 247 6289.

E-mail address: rempala.3@osu.edu (G.A. Rempala).

sequence of the variable domain of an antigen receptor there are three CDRs (CDR1, CDR2 and CDR3), arranged non-consecutively: CDR1 and CDR2 are found in the V region of a polypeptide chain, and CDR3 includes some of V, all of D and J, as well as some other TCR regions. This complicated CDR structure creates additional challenges in properly identifying different receptors.

Another important molecular mechanism responsible for the diversification of the TCR populations is based on a *major histocompatibility complex* (MHC)—a cell surface molecule which regulates T-cells lineage commitment during their thymic development process. During that process, the thymocytes with TCRs selected by MHC class II become the so-called CD4+ T-cells, and those with TCRs selected by MHC class I become the so-called CD8+ T-cells (see, e.g. Wong and Janeway, 1999). For a more in-depth overview of the TCR biology, the reader is referred to one of the standard references, like, e.g., Murphy et al. (2011).

The frequency of the unactivated (or naive) T-cell clones in normal individuals is minuscule; however, once a naive T-cell expressing appropriate TCR encounters antigen, it becomes activated and expands, forming a clonal population of cells, all expressing the same TCR. The analysis of such clonal expansions allows one to study the conversion between different functionally committed T-cells, which is a crucial step in understanding the basic molecular mechanisms of adaptive immunity functions in health and disease, such as infection, autoimmunity or transplantation (Kedzierska et al., 2008). Moreover, high individual convergence of TCR CDR3 amino acid sequences (Venturi et al., 2011) and our ever-expanding knowledge on their specificities (Venturi et al., 2008), suggest the potential in the near future for TCR-based diagnostics of various infectious and pathological states (cf., e.g., Kedzierska et al., 2008).

The goals of a typical statistical TCR profiling study are to determine (i) a diversity (i.e., the number of different clonotypes), (ii) an associated clonal distribution, and (iii) an overlap (or similarity) among different TCR counts sampled from various T-cell populations, often referred to as *TCR repertoires* in the immunological literature. With this in mind, the purpose of the current paper is twofold. First, we would like to briefly introduce to the quantitative biology community a method for comparison of multiple populations via the hierarchical clustering algorithm based on a class of multivariate count distributions, known as the multivariate *Poisson abundance (mixture) models* or *m PAMs*. As we show below, this approach may be broadly applied to simultaneous comparison of TCR repertoires obtained with the help of many standard biological assays, in most circumstances of practical interest. Secondly, and perhaps more importantly, we would like to propose a flexible, and in some sense optimal, inferential procedure for identifying the *m PAMs* parameters and subsequently carrying out the repertoires' clustering. We argue that our proposed method of TCR profiling allows one to not only compare graphically the clustering patterns, but also to analyze them quantitatively and, in particular, to attach a measure of uncertainty to the clustering hierarchy via the model-based, explicit confidence bounds.

Whereas the multivariate Poisson abundance models have been introduced in the context of TCR studies before (e.g. Rempala et al., 2011), they have suffered from a potentially very inefficient method of parameter estimation based on the truncated conditional (or partial) likelihood. Consequently, although they were able to deal with (iii) and to some extent with (ii) above, it was much harder for the conditionally fitted *m PAMs* to properly address the clonal diversity (i). Moreover, the inefficiency of the conditional parameter estimation often resulted in poor clustering. An example of the latter is provided below, when we compare different methods of analysis on some experimental data.

The paper is organized as follows. In the next section (Section 2) we briefly recall the basic notions related to hierarchical clustering

algorithms and *m PAMs*, as well as propose an MCMC-based algorithm (Algorithm 1) for the required statistical inference. In the same section, we also describe a specific *m PAM* known as the *multivariate Poisson-lognormal model* (MPLN) which is used for all our numerical examples throughout the paper. At the end of Section 2, we briefly discuss some computational aspects and limitations of our proposed inference method. In Section 3 we show, with the help of the MPLN model, how one may apply the tools of Section 2 to TCR data. Specifically, we analyze a dataset consisting of four TCR repertoires obtained from CD8+ $\alpha\beta$ T-cells in transgenic mice with compromised MHC-restriction and chain rearrangement abilities (Ignatowicz et al., 1996). The results of the analysis via the MPLN model and Algorithm 1 are compared with those from the alternative approaches used earlier in the literature, and the differences are discussed in the context of their biological implications. A review of our main points, along with some conclusions, is given in Section 4. For completeness, we discuss in the Appendix the statistical efficiency, as measured by the Fisher information, gained with the new proposed fitting method.

2. Clustering with multivariate abundance models

2.1. Hierarchical clustering

When studying the development of TCR populations it is often desirable to compare them simultaneously against some fixed baseline. This is the case, for instance, in clinical studies where one is interested in quantifying the “divergences” of TCR repertoires, sampled at various disease stages, from a control one. An attractive quantitative approach to this problem is offered by the *hierarchical clustering* approach, as often applied to DNA microarrays (Thalamuthu et al., 2006). Let us briefly recall some basic facts related to hierarchical clustering. For additional details, one may refer, for example, to Chapter 14 of Hastie et al. (2009).

For a given set of $m > 1$ TCR samples of interest, their hierarchical clustering depend on a particular pairwise dissimilarity (distance) index $Q(i,j)$ calculated between all distinct pairs of samples (i,j) . In our setting, $Q(i,j)$ will be derived from *m PAMs* as discussed below, but, in general, any appropriate distance measure may be used. For instance, some crude dissimilarity indices could be based simply on the joint TCR species presence/absence data, i.e., the number of TCR species shared by two samples and the number of species unique to each of them (see discussion in Legendre and Legendre, 1998). The two oldest and most widely used examples of such indices are based on the classical Jaccard and Sørensen similarity indices prevalent in ecological biodiversity studies (Magurran, 2005). For a given dissimilarity index Q , a stepwise procedure is employed, which results in a tree-like cluster structure (graphically summarized by a *dendrogram* or a “tree diagram”) with the clusters at each level of the tree created by merging or splitting clusters at the next level, according to the appropriately aggregated values of Q . Unlike in some other clustering algorithms, here there is no need to specify in advance the number of clusters to be created at each level. The hierarchical clustering is performed via either agglomerative methods, which proceed by series of fusions of the original *m TCR repertoires* into larger groups, or divisive methods, which successively separate repertoires into finer ones. As agglomerative methods are more commonly used, we also choose one of them for the TCR data below.

The extent to which the hierarchical structure produced by a dendrogram actually represents the data itself can be judged by the *cophenetic correlation* coefficient. This is the correlation between the $m(m-1)/2$ values of $Q(i,j)$ and the corresponding

cophenetic dissimilarities derived from the dendrogram. The cophenetic dissimilarity between any two samples (ij) is the value of the intergroup dissimilarity index at which i and j are first clustered together. The cophenetic correlation coefficient may be used to numerically assess to what extent various dissimilarity indices reflect the true pattern of the data, with higher positive values indicating better agreement.

2.2. Multivariate Poisson abundance models

The idea of modeling count data with a Poisson abundance model (PAM) goes back to Fisher et al. (1943) where it was proposed as an extension of a univariate Poisson model to the over-dispersed data (see, e.g., Sepúlveda et al., 2010). In turn, m PAM may be viewed as a multivariate extension of PAM.

As all vertebral T-cell receptor populations originate in the thymus, we denote the size of the initial thymic TCR population (i.e. the number of distinct clonotypes) by M and regard it as a parameter in the subsequent analysis of the m TCR repertoires of interest (in our specific data example below, $m=4$). Since PAM assumes (see, e.g., Chao, 2006; Rempala et al., 2011) that the recorded counts are arriving from a mixture of Poisson processes in some time interval, we consider therefore the observed counts of the i th clonotype ($i = 1, \dots, M$) as arriving according to an m -variate Poisson process with the marginal rates $(\lambda_1, \lambda_2, \dots, \lambda_m)$. If the detectability of individuals is assumed to be equal across all clonotypes (which is typically the case in TCR repertoire sequencing), then the rates may be interpreted as clonal species abundances (Nayak, 1991).

Since antigen receptor clonal size distributions are generally regarded as having heavy right tails, the population-specific species abundance rates $(\lambda_1, \lambda_2, \dots, \lambda_m)$ may be modeled jointly as a random vector from a mixing distribution with density $g_\theta(\lambda_1, \dots, \lambda_m)$, where θ is a vector of parameters. According to m PAM, the clonal counts represent therefore a multivariate sample from a mixture distribution of m conditionally independent of Poisson variates (cf., Rempala et al., 2011)

$$p_\theta(k_1, \dots, k_m) = \int_0^\infty \dots \int_0^\infty \prod_{i=1}^m \left[\frac{\lambda_i^{k_i} \exp(-\lambda_i)}{k_i!} \right] g_\theta(\lambda_1, \dots, \lambda_m) d\lambda_1, \dots, d\lambda_m, \quad (1)$$

$k_i = 0, 1, \dots, i = 1, \dots, m,$

where $p_\theta(k_1, \dots, k_m)$ is the probability that a TCR clonotype is present k_i times in the sample from the i th TCR population. Let f_{k_1, \dots, k_m} be the observed (empirical) count of such clonotypes¹ with $D = \sum_{k_1, \dots, k_m} f_{k_1, \dots, k_m}$ and $f_{0, \dots, 0} = 0$, so that $E(f_{k_1, \dots, k_m}) = M p_\theta(k_1, \dots, k_m)$. Under m PAM, the data consist of a collection of all observed counts $\{f_{k_1, \dots, k_m}\}$ with the likelihood function given by the multinomial distribution

$$\begin{aligned} \mathcal{L}(M, \theta | \{f_{k_1, \dots, k_m}\}) &= \frac{M!}{(M-D)! \prod_{k_1, \dots, k_m \geq 0} (f_{k_1, \dots, k_m})!} [p_\theta(0, \dots, 0)]^{M-D} \\ &\times \prod_{k_1, \dots, k_m \geq 0} [p_\theta(k_1, \dots, k_m)]^{f_{k_1, \dots, k_m}} \\ &= \frac{M!}{(M-D)! D!} [p_\theta(0, \dots, 0)]^{M-D} [1 - p_\theta(0, \dots, 0)]^D \\ &\times \frac{D!}{\prod_{k_1, \dots, k_m \geq 0} (f_{k_1, \dots, k_m})!} \prod_{k_1, \dots, k_m \geq 0} \left[\frac{p_\theta(k_1, \dots, k_m)}{1 - p_\theta(0, \dots, 0)} \right]^{f_{k_1, \dots, k_m}}. \end{aligned} \quad (2)$$

Note that the above implies that the likelihood function for M and θ can be factored as

$$\mathcal{L}(M, \theta | \{f_{k_1, \dots, k_m}\}) = \mathcal{L}_b(M, \theta | D) \mathcal{L}_c(\theta | \{f_{k_1, \dots, k_m}\}, D), \quad (3)$$

where $\mathcal{L}_b(M, \theta | D)$ is the likelihood with respect to D , the binomial random variable with parameters $(M, 1 - p_\theta(0, \dots, 0))$, and $\mathcal{L}_c(\theta | \{f_{k_1, \dots, k_m}\}, D)$ is the (conditional) multinomial likelihood with respect to $\{f_{k_1, \dots, k_m}, \sum_i k_i > 0\}$ with the total sample size D and the zero-truncated cell probabilities $\{p_\theta(k_1, \dots, k_m) / [1 - p_\theta(0, \dots, 0)]\}_{k_1, \dots, k_m, \sum_i k_i > 0}$. Since the likelihood function $\mathcal{L}_c(\theta | \{f_{k_1, \dots, k_m}\}, D)$ does not involve M , the inference for θ may be based on the partial MLE $\hat{\theta}_c = \arg\max_\theta \mathcal{L}_c(\theta | \{f_{k_1, \dots, k_m}\}, D)$ (Rempala et al., 2011; Engen et al., 2002), particularly when the value of M is not of immediate interest. In Section 1, we have referred to this method as the “partial” or “conditional” inference. Unfortunately, this approach leaves out the factor $\mathcal{L}_b(M, \theta | D)$ from the model fitting considerations and, depending upon the particular form of the distribution p_θ , may be far from optimal (see Appendix). Despite this drawback, the conditional inference method seems quite popular in the literature, as it provides a viable alternative to the generally difficult problem of directly maximizing the full likelihood $\mathcal{L}(M, \theta | \{f_{k_1, \dots, k_m}\})$ in pursuit of the MLE for (M, θ) .

Whereas the direct frequentist inference for (M, θ) is challenging, it seems that adopting a Bayesian viewpoint may circumvent some of the difficulties. Indeed, considering m PAM within the Bayesian framework, we may regard g_θ as a prior distribution on the rates vector $(\lambda_1, \dots, \lambda_m)$ and, accordingly, (M, θ) as the underlying hyperparameter. Consequently, under the hierarchical Bayesian model with the improper (uniform) prior on (M, θ) , we may regard the likelihood function $\mathcal{L}(M, \theta | \{f_{k_1, \dots, k_m}\})$ as the marginalization of the posterior distribution of $(M, \theta, \lambda_1, \dots, \lambda_m | \{f_{k_1, \dots, k_m}\})$ over g_θ . This interpretation leads to a straightforward MCMC approach in the (M, θ) inference problem (cf., also Barger and Bunge, 2008 who, in a different context, applied a similar approach or Rodrigues et al., 2001 who applied this idea for the univariate case).

2.3. MCMC inference for PAM

First, note that if $\mathcal{L}(M, \theta | \{f_{k_1, \dots, k_m}\})$ is viewed as the posterior distribution of $(M, \theta | \{f_{k_1, \dots, k_m}\})$ (with a non-informative prior), then the conditional $(M | \theta, \{f_{k_1, \dots, k_m}\})$ has a negative binomial (nb) distribution. Indeed, with this interpretation M simply describes the number of trials needed for D successes to occur, where each success has a probability $1 - p_\theta(0, \dots, 0)$. In abbreviation, $(M | \theta, \{f_{k_1, \dots, k_m}\}) \sim nb(D, 1 - p_\theta(0, \dots, 0))$. On the other hand, the conditional $(\theta | M, \{f_{k_1, \dots, k_m}\})$ may be obtained by the marginalization of the joint posterior distribution $(\theta, \lambda_1, \dots, \lambda_m | M, \{f_{k_1, \dots, k_m}\})$ over g_θ , with the sampling from that distribution done via the usual Metropolis–Hastings method (see, for instance, Andrieu et al., 2003). These considerations may be summarized in the Gibbs-sampler algorithm with a nested Metropolis–Hastings step described below. Upon convergence, the algorithm produces samples from the posterior distribution $(M, \theta, \lambda_1, \dots, \lambda_m | \{f_{k_1, \dots, k_m}\})$ and, in particular, from $(M, \theta | \{f_{k_1, \dots, k_m}\})$. The empirical mode of the latter may be viewed as an approximate maximal a posteriori estimator (MAP) which, in this particular case, is also an approximate MLE based on the complete likelihood $\mathcal{L}(M, \theta | \{f_{k_1, \dots, k_m}\})$. The advantages of utilizing the estimate of θ based on the complete rather than the partial likelihood (\mathcal{L} vs. \mathcal{L}_c) may be seen in terms of the gain in the Fisher information, as briefly described in the Appendix.

Algorithm 1. Hybrid Gibbs sampler.

1. Initiate with $M=D$.
2. Perform a Metropolis–Hastings step for the target distribution $(\theta, \lambda_1, \dots, \lambda_m | M, \{f_{k_1, \dots, k_m}\})$.

¹ That is, the count of all clonotypes which appeared k_1 times in sample one, k_2 times in sample two, etc.

3. Evaluate $p_\theta(0, \dots, 0)$ for the sampled value of θ .
4. Sample M value from $nb(D, 1-p_\theta(0, \dots, 0))$ and return to Step 2.
5. Repeat 2–4 until convergence.

Assuming that the draws from g_θ are easily obtained, the mode of the empirical posterior distribution produced by the algorithm above will give an approximate full likelihood MLE $(\hat{M}, \hat{\theta})$ for any PAM of the form (1).

Note that the (profile) pointwise M estimate is approximately $\hat{M} \approx D/(1-p_\theta(0, \dots, 0))$ (4)

which is the estimate considered in Rempala et al. (2011), but without a clear justification (for some earlier discussion, see also Sanathanan, 1972; Bulmer, 1974).

It is perhaps also worthy to notice that the use of a proper prior in the above algorithm generally may not be beneficial. On one hand, in order to retain the simplicity of the conditional sampling in Step 4, a prior distribution conjugates to the negative binomial should be used, for instance, a univariate PAM distribution. However, it is easy to see that in our setting this class of priors would make the posterior distribution unduly sensitive to the hyper-parameters, leading to a biased inferential procedure. On the other hand, a more sophisticated, non-conjugate prior distribution would necessarily add to the computational overhead, as it would require an additional Metropolis–Hastings step.

In what follows, we apply Algorithm 1 in the case when g_θ is the density of a multivariate lognormal random variable, which in turn makes p_θ the mass function of a multivariate Poisson-lognormal distribution (see, e.g., Aitchison and Ho, 1989).

2.4. Poisson-lognormal dissimilarity index

Let us now derive a dissimilarity index Q associated with the particularly convenient (for our purpose) m PAM model which we shall use in our numerical examples in the next sections.

Although there are many possible parametric multivariate mixture models, recent works by Sepúlveda et al. (2010) and Rempala et al. (2011) suggest that lognormal mixing distributions may often be especially appropriate for TCR repertoires modeling. Because of this, we consider henceforth a multivariate model based on lognormal variates, which is a straightforward extension of the bivariate model considered in Rempala et al. (2011). Recall that under the Poisson abundance model discussed above, the number of individuals sampled from any given receptor clone species with abundance λ is Poisson distributed with mean λ . If one assumes that $\ln \lambda$ is normally distributed with mean μ and variance σ^2 , then the vector of clonotypes sampled from all M species comprises a sample from the Poisson-lognormal distribution with parameters $\theta = (\mu, \sigma^2)$, where μ and σ^2 are the mean and variance of the log-abundances. The corresponding mass function may be written as

$$p(k; \mu, \sigma^2) = \int_{-\infty}^{\infty} g_k(\mu, \sigma, u) \phi(u) du, \quad (5)$$

where $\phi(\cdot)$ is the standard normal density function and

$$g_k(\mu, \sigma, u) = \frac{\exp(u\sigma k + \mu k + e^{-(u\sigma + \mu)})}{k!}, \quad k \geq 0$$

is the re-parameterized Poisson distribution. Similarly, the m -tuples of clonotype counts from m different repertoires constitute a random sample (of size M) from the multivariate Poisson-lognormal distribution (MPLN), a special case of m PAM given in (1). The log m -abundances of the clonotypes have therefore the multivariate normal distribution with mean vector μ and variance-covariance matrix Σ . Let $\phi(u_1, \dots, u_m; \rho)$ denote the normal multivariate density with correlation matrix $\underline{\rho} = [\rho_{ij}]$, zero means, and unit

variances. The distribution of MPLN is given in terms of the multivariate probability mass function $p_\theta(k_1, \dots, k_m) = p(k_1, \dots, k_m, \underline{\mu}, \Sigma)$ for $k_i \geq 0$ where

$$p(k_1, \dots, k_m, \underline{\mu}, \Sigma) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \left[\prod_{i=1}^m g_{k_i}(\mu_i, \sigma_i, u_i) \right] \times \phi(u_1, \dots, u_m; \rho) du_1, \dots, du_m.$$

From the above formula we can obtain any other MPLN of lower dimension by integrating out (marginalizing over) an appropriate subset of m variables.

Under the assumed MPLN model, the dissimilarity index Q of Section 2.1 may be conveniently defined in terms of the entries of the rescaled matrix Σ . This choice corresponds to the correlation-based dissimilarity measure discussed, e.g., in Rempala et al. (2011).

Denoting the mean and variance of the i th MPLN marginal by α_i, β_i^2 , respectively, we have

$$\begin{aligned} \alpha_i &= \exp(\mu_i + \sigma_i^2/2) \\ \beta_i^2 &= \alpha_i + \alpha_i^2[\exp(\sigma_i^2) - 1] \end{aligned} \quad (6)$$

for $i=1, \dots, m$. The correlation coefficient values evaluated for all the possible pairs (ij) of MPLN distribution marginals give therefore the following dissimilarity index²:

$$Q_\theta(i, j) = 1 - \frac{\alpha_i \alpha_j |\exp(\rho_{ij} \sigma_i \sigma_j) - 1|}{\sqrt{\alpha_i \alpha_j (1 + \alpha_i [\exp(\sigma_i^2) - 1]) (1 + \alpha_j [\exp(\sigma_j^2) - 1])}}. \quad (7)$$

In practice, the above quantity needs to be approximated via the sample estimate $Q_\theta(i, j)$ with $\hat{\theta} = (\hat{\underline{\mu}}, \hat{\Sigma})$ taken as the mode of an empirical posterior distribution obtained via Algorithm 1, that is, an approximate MLE for θ . In this way, the fitted MPLN distribution defines uniquely the values of the dissimilarity index Q_θ and, consequently, the corresponding hierarchical clustering for the analyzed TCR repertoires. We shall illustrate this process with a data example in Section 3.

Whereas for the particular dataset considered in Section 3, the computational cost was only moderately high, in general, the overall processing time is seen to increase quickly with the value of m . To illustrate this for some benchmark CPU processing times, the relative computational cost of fitting an MPLN model using Algorithm 1, as a function of the dimension m of the count distribution, is provided in Table 1. Note that while the processing time increases with m , the amount of CPU effort per model parameter remains relatively steady. From the computational perspective, it seems therefore preferred to fit a single high-dimensional model, rather than multiple lower-dimensional ones. For instance, based on the values in the table, for $m=10$ the pairwise fitting algorithm described in Rempala et al. (2011), which uses the multiple conditional fitting of the bivariate model in order to obtain the final MPLN fit, is seen to require on average $(^{10}_2) = 45$ min of CPU time, as opposed to about 10 min for a single fitting of the full MPLN model.

3. Application to TCR data analysis

3.1. Biological data

The animal TCR repertoire samples considered here are obtained from a thymus of a TCRmini mouse (see, e.g., Pacholczyk et al., 2006) in which all T-cells have a restricted range of possible $\alpha\beta$ TCR

² Note the typo in the similar formula given in (3.14) of Rempala et al. (2011) where the absolute value needs to be added to the numerator and the factor of 2 needs to be removed.

clonotypes. Specifically, the receptors on the $\alpha\beta$ T-cells from TCRmini mouse differ from each other only in the CDR3 region of the α chain (all CDR1 and CDR2 regions and CDR3 β chain are fixed). This restricted rearrangements model allows one to more easily analyze the diversity of the TCR repertoires as well as to track in vivo the fate of individual T-cell clones. Consequently, as our experimental dataset, we take a collection of four different TCR samples harvested from the TCR repertoires of CD8+ T-cells in different TCRmini mice populations. The first two sampled repertoires were (i) a “baseline” TCRmini (denoted WT) which expresses both all class I MHC as well as a particular class II MHC (called AbEp) and (ii) a modified TCRmini where all Ab molecules are bound with a single Ep peptide (denoted AbEp). Two additional TCR repertoires were sampled from the two populations of irradiated AbEp and RAG−/− mice (Kawano et al., 1997) with reconstituted bone marrows taken from the WT mice population. The irradiated AbEp and RAG−/− mice are referred to below as radiation chimeras and denoted by AbEp(c) and RAG(c), respectively. The TCR samples from the first two populations of the CD8+ T-cells (WT and AbEp) were collected using labor-intensive, but also more reliable, single cell sequencing (see, e.g., Warren et al., 2009). In contrast, the samples from the TCRs of radiation chimeras were obtained via high-throughput sequencing using a 454 platform (see, e.g., Lai et al., 2012). This difference in the collection method is expected to influence the relative sampling intensity (see above), but not directly the dissimilarity index pattern across the four repertoires. Due to the biology of the specific animal models, one would expect to see strong dissimilarity between AbEp (single peptide) and the remaining three populations. The relations among other repertoires are less clear a priori.

3.2. Summary statistics

The complete dataset consisting of all sequenced receptors in four repertoire samples (for a total of $D=310$ different

clonotypes) is provided as supplementary material, which may be downloaded from the journal site. The pictorial summary of the empirical frequencies for each of the four TCR repertoires considered is presented in Fig. 1, as a set of four frequency plots. Each plot represents the number of observed counts for each clonotype (some possibly zero, with frequencies over 200 truncated for better visibility) observed in the respective population. The ordering of the clonotypes remains the same across plots so as to allow for direct comparisons. The summary statistics for the data are provided in Table 2. In the notation of Section 2.2, let $D_i = \sum_{k>0} f_k^{(i)}$ and $n_i = \sum_k k f_k^{(i)}$, where $f_k^{(i)}$ is the number of clonotypes observed k times in the repertoire i , $i=1, \dots, 4$. The observed values of D_i and n_i , based on all the observed clonotypes, are presented in Table 2. The observed overlap between (i.e., presence in both) the combined chimera (c) and mini (WT, AbEp) populations was found at around 48% (which is consistent with earlier findings, see, e.g., Pacholczyk et al., 2007).

It seems intuitively clear from the frequency plots that the two chimera populations and the two remaining ones should be clustered together. The goal of our analysis below is to quantify and formally test this intuition.

3.3. Results

The results of the MCMC analysis in terms of the approximate MLE values for $(M, \underline{\mu}, \Sigma^{-1})$ based on the MAP estimates obtained via Algorithm 1 along with their credibility bounds are presented in Table 3. The required computational analysis was performed with the help of the R library *rjags* and the JAGS software (R Team, 2010; Plummer, 2003) where for technical reasons, rather than Σ , the log-abundances precision matrix Σ^{-1} was estimated. For the purpose of the analysis, the non-informative independent improper priors were used for $(M, \underline{\mu}, \Sigma^{-1})$. Recall that the marginal parameters μ and σ^2 are related to the marginal means and variances of the Poisson-lognormal variates by the formula (6).

The marginal values of the repertoire-specific parameters in both types of repertoires (chimera and mini) were found to be of similar magnitude (with estimated values of μ parameters between -3.6 and -1.7 and σ^{-2} , denoted in the table as σ_{ii}^{-1} ($i=1, \dots, 4$), between 0.4 and 0.7. Overall, the numerical values of the parameters indicated the smaller Poisson-lognormal means for the restricted-repertoire mice, as compared with the wild-type, although this finding may be confounded with the fact that the sampling intensity of the chimera populations was markedly different. In general, the issue of unequal sampling intensity may present a challenge for the type of data considered here (see, e.g., discussion in Rempala et al., 2011 on the effects of sampling

Table 1
CPU usage in MPLN model fitting. Average increase of CPU processing time in Algorithm 1 for every 1 min (or about 100 iterations) of the CPU processing for the bivariate model fit ($m=2$). The comparison is based on the average time runs for two chains in a simulation study using a HP Pavilion dv7 Notebook PC with a 64-bit version of Windows 7, 3.75 GB of usable RAM, and Dual-Core M620 AMD Turion 2.50 GHz processor.

Dim. (m)	No. of params	Avg. CPU time (min)
2	5	1.00
3	9	1.47
4	14	2.38
5	20	3.40
10	65	10.05

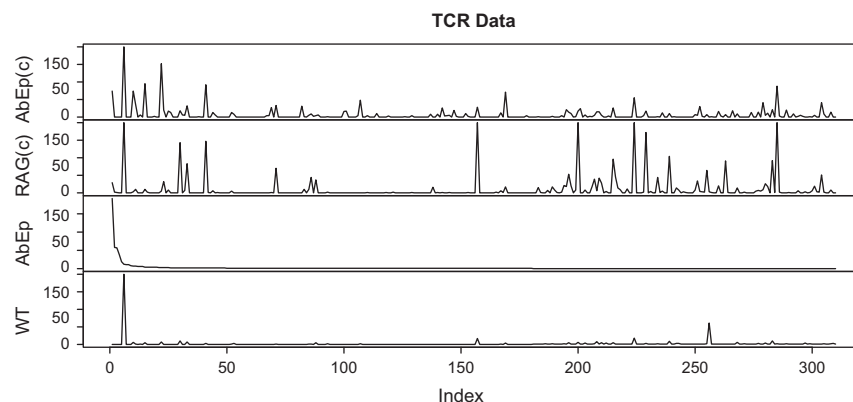


Fig. 1. Frequency plots for TCR data. For better visibility, the plot for the AbEp population was truncated at 200.

Table 2
TCR summary statistics. The observed values of D_i and n_i for each sample in the TCR data. The clonotype overlaps between the pooled chimera (c) populations and the pooled “mini” ones (WT, AbEp) was approx. 48% or 148 out of a total of $D=310$ different receptors.

	AbEp(c)	RAG(c)	AbEp	WT
D_i	110	111	180	150
n_i	2009	6160	657	628

Table 3
MPLN model estimates for TCR data. The estimated posterior values of the log-intensity means, the missing species number and the entries of the precision matrix (Σ^{-1}) for the log-abundances in four TCR repertoires considered, along with the bias-corrected, two-sided 95% credibility bounds generated via MCMC. For the purpose of comparison, the corresponding values obtained by maximizing the conditional likelihood are provided in parenthesis.

Value	μ_1	μ_2	μ_3	μ_4	$M-D$
95%Lo	−3.38	−3.62	−1.70	−2.37	266.70
95%Up	(−7.11)	(−8.52)	(−4.91)	(−6.82)	–
	−4.21	−4.46	−2.19	−2.90	177
	−2.62	−2.84	−1.28	−1.88	379
	σ_{11}^{-1}	σ_{22}^{-1}	σ_{33}^{-1}	σ_{44}^{-1}	σ_{12}^{-1}
95%Lo	0.41	0.50	0.72	1.75	−0.17
95%Up	(1.12)	(1.07)	(0.51)	(1.64)	(−0.48)
	0.27	0.34	0.42	1.04	−0.3
	0.64	0.72	1.31	3.05	−0.04
	σ_{13}^{-1}	σ_{14}^{-1}	σ_{23}^{-1}	σ_{24}^{-1}	σ_{34}^{-1}
95%Lo	−0.30	−0.28	−0.07	−0.63	0.57
95%Up	(−0.44)	(−0.49)	(−0.02)	(−0.51)	(0.24)
	−0.64	−0.79	−0.29	−1.00	0.14
	−0.14	−0.01	0.10	−0.34	1.44

intensity on the MPLN estimates). Fortunately, for the diversity analysis via hierarchical clustering as presented below, the sampling intensity by itself was irrelevant and had been therefore incorporated into the estimates of the means for marginal log-abundances. The MLE value for the missed receptors ($M-D$) was found as approximately 267, based on the posterior mode. This value is consistent with the previous studies on TCR diversity in mini mice (Pacholczyk et al., 2006) and gives the approximate percentage of the unseen species as 50%. The diversity estimate along with its credibility bounds is presented in the right-most column at the top of Table 3.

In Table 3, for the purpose of comparison, we have also provided (in parenthesis) the estimated values of the MPLN model parameters obtained using the conditional likelihood model (i.e., via maximizing \mathcal{L}_c in (3)). These values were calculated with the help of the simple Newton–Raphson optimization algorithm with random restarts (as provided by the *poilog* R library, see Rempala et al., 2011). With the resampling option for the error analysis turned on, the algorithm was computationally comparable to MCMC in terms of the CPU usage. However, as seen by comparing the entries in Table 3, the two sets of final estimated values differ considerably, particularly with respect to the MPLN shift parameters μ .

The goodness-of-fit analysis for the MCMC-fitted model was conducted by means of the Bayesian χ^2 statistic proposed recently by Johnson (2004) and based on the MCMC samples obtained via Algorithm 1. The difference between the MCMC samples and the χ^2 distribution was not-significant (all p -values < 0.05), and indicated a good fit of the MPLN to the data. In addition to the

goodness-of-fit testing, we have also performed more qualitative comparisons of data against the random sample from the fitted MPLN model via four marginal QQ-plots (Fig. 2). Except for the first population, where an outlier is distorting the larger quantiles, the remaining data quantiles plots seem to give a reasonable indication of agreement with a random sample from MPLN.

The results of the MCMC-based hierarchical clustering analysis of the four mice TCR repertoire samples are presented in Fig. 3. The figure top-left panel shows the dendrogram obtained by the agglomerative hierarchical clustering with a complete link function (see, e.g., Hastie et al., 2009, Chapter 14 for a definition) under Q_θ dissimilarity index (7) where $\theta = (\mu, \Sigma)$ is fitted by the MCMC procedure outlined in Algorithm 1. For the purpose of comparison, the remaining two top panels depict the alternative dendrograms obtained (in the top-middle panel) under the fully non-parametric index defined as $Q_1(i, j) = 1 - |\varrho_{np}(i, j)|$, where $\varrho_{np}(\cdot)$ is the non-parametric Pearson correlation coefficient between samples i and j , and (in the top-right panel) under the index Q_θ where the MPLN parameters θ are fitted via the conditional likelihood inference (i.e., using the values listed in parenthesis in Table 3). The latter two dendrograms are seen as representing a very similar clustering pattern which is markedly different from that obtained with the full-likelihood MPLN model. Although both patterns describe a biologically plausible hierarchical structure, which keeps the AbEp population away from the remaining ones, the overall dissimilarity values based on the full MPLN are seen as larger, which is somewhat more consistent with the biological model. For a quantitative assessment of the consistency of all three dendrograms with the data, we have calculated their respective cophenetic correlations (see Section 2.1), along with their confidence intervals. The sample-based values of the cophenetic correlation for different clusterings are presented in their corresponding dendrogram plots. These values are also given in Table 4, along with their respective credibility or confidence bounds, based on the MCMC samples for the full MPLN, and on the bootstrap method for both the Pearson correlation and the conditionally fitted MPLN. In all three cases, the high values of the correlations indicate the internal consistency of the clustering with their corresponding dendrogram structures. However, the value for the index Q_θ estimated via the full-likelihood MPLN MLE is seen as slightly higher than the remaining ones, and hence that clustering is preferred.

The remaining (bottom) panels of Fig. 3 depict the upper and lower 95% credibility dendrograms (bottom-middle and bottom-right panels, respectively). The fact that both display the identical hierarchy tree indicates a strong robustness of the hierarchical clustering in the full MPLN model against the sampling fluctuations of the dissimilarity index. As in Rempala et al. (2011), these credibility bounds were obtained by inverting the upper and lower 2.5% quantiles of the Frobenius norm³ of the dissimilarity matrix $[Q_\theta(i, j)]$. The density estimator of the Frobenius norm distribution, with the upper and lower 2.5% quantiles marked as the vertical lines, is shown in the bottom-left panel.

The result of the MCMC procedure in Algorithm 1, which produced the samples used for clustering analysis above, is presented in the left panel of Fig. 4 as the Gibbs sampler's trace plot based on 6000 steps, after discarding the initial 4000 steps as the burn-in. For better visualization, only the projection of the posterior samples onto the bi-variate subspace $(\|\underline{\mu}\|, \|\Sigma^{-1}\|)$ is plotted, where $\|\cdot\|$ stands for the matrix Frobenius norm. The Gelman–Rubin statistic R was used to diagnose the sampler

³ Recall that for any real matrix A its Frobenius norm $\|A\|$ is $\sqrt{\text{Tr}(A^T A)}$. See also the Appendix or, e.g., Golub and Van Loan (1996) for a general reference on matrix norms.

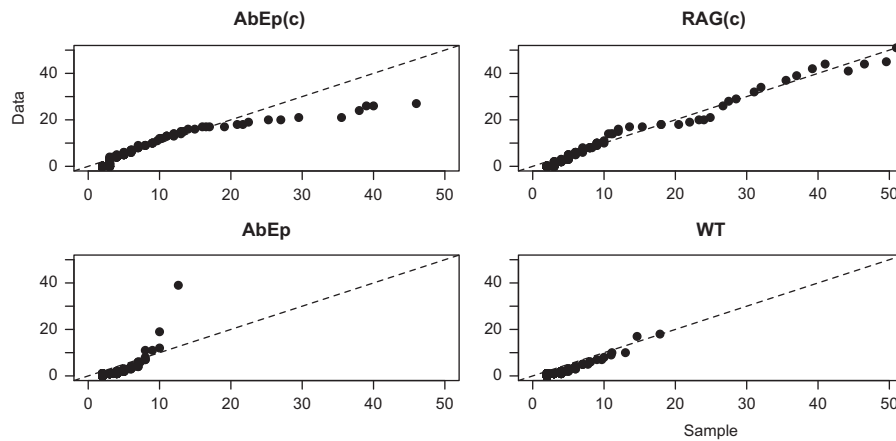


Fig. 2. Goodness of fit analysis for TCR data. Four panels illustrate the quantile plots of the data against a zero-truncated random sample from the MPLN model with parameter values taken at the mode of the posterior distribution of (M, θ) . The dashed line represents $y=x$ function. The agreement in the upper quantiles in AbEp(c) populations is seen to be distorted by an outlier in the observed frequencies (outside the plot range).

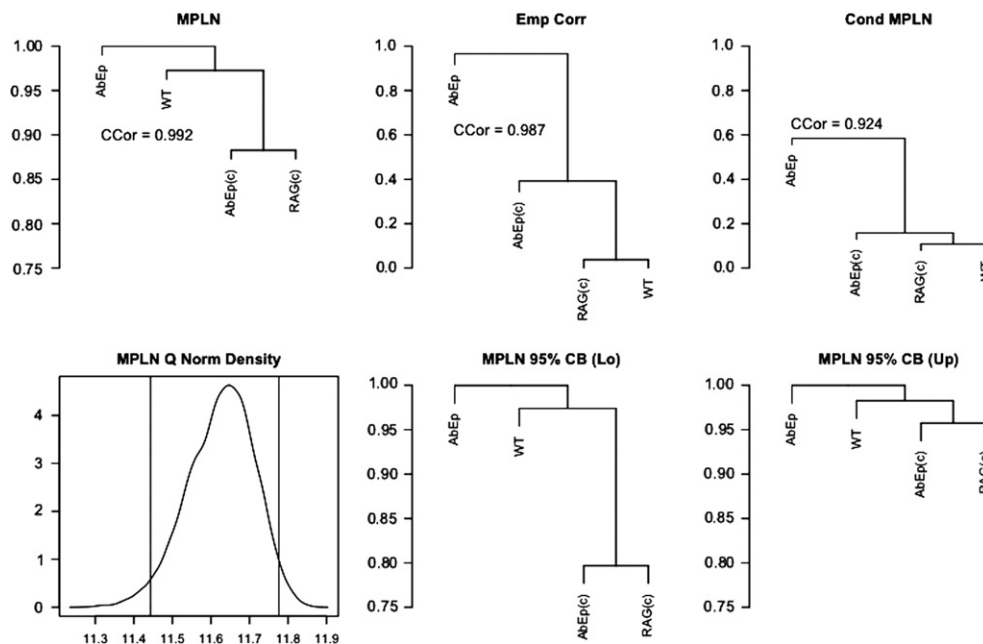


Fig. 3. Analysis results for TCR data. Top panels illustrate the final clustering via MPLN model (left) and the two alternative clusterings using (i) the empirical correlation coefficient (middle), and (ii) the conditional MPLN. The corresponding values of the cophenetic correlations are given for comparison. Bottom panels show the Frobenius norm density of the dissimilarity matrix Q_D (left) as well as the credibility bounds for the MPLN clustering (middle and right).

Table 4

The estimated cophenetic correlation coefficients for clustering under Q_D via the full (MPLN) and conditional (Cond. MPLN) models as well as using the non-parametric dissimilarity based on the Pearson correlation (Emp. Corr). The higher value indicates better clustering. The 95% credibility interval based on the MCMC runs is reported for the MPLN model coefficient, while for the remaining two the 95% confidence interval based on the bias-corrected bootstrap percentile method is reported.

Dissimilarity index	CC value	95%Lo	95%Up
MPLN	0.992	0.987	0.997
Cond. MPLN	0.924	0.794	0.986
Emp. Corr	0.987	0.659	1.0

convergence (R values close to unity indicate convergence, see Gelman and Rubin, 1992). For further illustration, the posterior density of the unobserved species $(M-D)$ is presented in the right panel, indicating in particular a good agreement between the

posterior mode (the MAP estimate) and the posterior mean (the Bayesian estimate) of $M-D$.

Based upon the cophenetic correlation values, the clustering analysis with the full-likelihood MPLN model is seen as the most likely representation of the true relationships among the four TCR populations. The inspection of the top-left panel dendrogram in Fig. 3 along with its credibility bounds (the bottom-middle and bottom-left panels) provides one with several biologically interesting and, very importantly, *statistically significant* findings. First, as expected, the dissimilarity pattern revealed by the top-right dendrogram indicates that the TCRs samples from the original AbEp mice CD8+ T-cells were the most different from the remaining three sets of TCR data. This is likely due to the fact that they were the only ones not in contact with class II MHC bound with multiple peptides. Although this difference is also visible in the alternative dendrograms (top-middle and top-right panels) constructed under the competing dissimilarity indices, in both cases it turns out not to be significant at 5% level, as

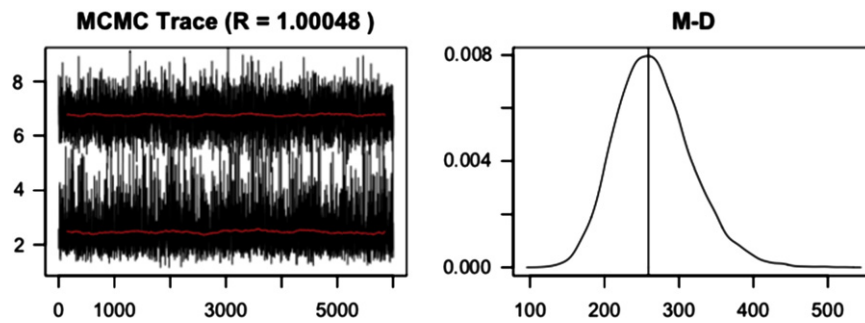


Fig. 4. MCMC plots for TCR data. Left panel—the bivariate projection of the trace in the converged Gibbs sampler described in Algorithm 1 based on 6000 iterations and 4000 burn-in steps. The value of Rubin's R statistic (above) and moving average plots both indicate convergence. Right panel—the posterior density of $M-D$ based on 6000 iterations of the converged Gibbs sampler. The vertical line drawn at the mode gives an approximate value of the MAP for $M-D$ given in Table 3.

measured by the alternative dendrograms confidence bounds (in order to conserve space, we forgo the details of this additional analysis here). Secondly, the MPLN clustering also indicates that CD8+ T-cell TCRs of the AbEp chimera mice (c), which are reconstituted with TCRmini WT bone marrow, more closely resemble the TCRs of the WT mice than those of the AbEp mice. This finding, if confirmed with other experiments, is potentially quite significant as it may indicate that the thymus negative selection (i.e., the self-destruction of certain clonotypes) is an important discriminating factor between mice populations TCRmini AbEp (MHC class II-one peptide) and TCRmini wild-type (MHC class II-many peptides).

Finally, the fitted MPLN model indicates that the overall frequency patterns of chimera mice made them more similar to each other than to the remaining two TCR populations. This seems to be also consistent with the summary plots presented in Fig. 1 reflecting, among others, the fact that the different TCR harvesting method was used to obtain data from the chimera populations. It is important to note that this last conclusion cannot be reached by means of the dissimilarity analysis based on either the empirical correlation method, or the partial likelihood fit of MPLN, as both are unadjusted for the unseen clonotypes and sensitive to the large spikes of the empirical frequencies. In contrast, the full MPLN model is seen to down-weight the spikes as it adjusts for the estimated distribution of the $M-D$ unobserved TCR types (Fig. 4, right panel).

4. Summary and discussion

We have presented a method of analyzing multiple TCR repertoires with a parametric model based on a multivariate Poisson distribution, and more generally, on a class of multivariate count distributions known as the multivariate Poisson abundance models (m PAMs). Whereas m PAMs have been used to model T-cell counts in previous works, they have been typically fitted by means of the truncated likelihood function, conditioned on the number of observed species. Since this method leaves out a factor in the complete likelihood function, such a fitting procedure is likely to be less efficient, as illustrated here via the Fisher information analysis discussed in the Appendix. As a remedy, we have proposed a likelihood profiling method, which allows one to fit the model using a complete likelihood function. Moreover, we have shown that by introducing a set of improper non-informative prior distributions on the m PAM parameters, the fitting procedure may easily be carried out as a hybrid Gibbs sampler. Such a sampler may be constructed using popular open source software, like, e.g., JAGS or BUGS (Lunn et al., 2009). The samples from the converged sampler may be then used to perform the desired model diagnostics and error analysis, with

the marginal modes of the posterior samples corresponding to the approximate respective MLEs. Computationally, the method is seen as only marginally more demanding than the conditional one, which in turn requires parametric resampling for the error analysis.

Our proposed new approach to TCR analysis was illustrated on biological samples from four CD8+ T-cell populations harvested via two different laboratory techniques and involving two TCRmini mice populations (one with and another without the TCR recombination restrictions) and two chimera mice populations in which the bone marrow reconstitution was performed. In order to identify the differences among these TCR populations, the complete likelihood method was applied to fit a multivariate Poisson-lognormal model. The resulting clustering based on the covariance between the fitted model components was seen to correctly identify the main underlying biological pattern (a population with TCR restricted recombinations was seen as markedly different from the remaining ones) as well as to provide additional insights into the similarities among the remaining repertoires which were missed by both a less sophisticated, conditional MPLN analysis, and a simple non-parametric one based on the Pearson-correlation.

Overall, the proposed MCMC method was found to be conceptually straightforward and computationally feasible when applied to TCR data from a sequencing experiment. As analyzing such data starts to play a central role in modern studies on acquired immune response, we hope the statistical methodology proposed here could become standard in many circumstances of practical interest.

Acknowledgments

The research was partially funded by the NIH under Grant R01CA152158 to GAR. The authors would like to thank the members of the Ignatowicz's research laboratory for providing TCR data for the analysis and for helpful discussions. We would also like to thank the reviewers and the associate editor for their comments and improvement suggestions made on the earlier drafts of the manuscript.

Author disclosure statement: no competing financial interests exist.

Appendix A. Efficiency in PAM inference

The relative gain in efficiency when fitting the PAMs described in Section 1 by maximizing the complete likelihood $\mathcal{L}(M, \theta | \{f_{k_1, \dots, k_m}\})$ rather than $\mathcal{L}_c(\theta | \{f_{k_1, \dots, k_m}\}, D)$, may be formally quantified by analyzing the Fisher information about θ contained in $\mathcal{L}_c(\theta | \{f_{k_1, \dots, k_m}\}, D)$

relative to that contained in $\mathcal{L}(M, \theta | \{f_{k_1, \dots, k_m}\})$. For the purpose of such an analysis, the latter may be treated as a function of the PAM parameter θ only, with M remaining fixed (to emphasize this fact, we write $\mathcal{L}^M(\theta | \{f_{k_1, \dots, k_m}\})$ below).

In order to simplify notation and facilitate the numerical calculations below, assume that there are $K+1$ classes $k=0, \dots, K < \infty$ to which we assign the empirical counts $\{f_{k_1, \dots, k_m}\}$, denoted now by x_0, \dots, x_K , with each class having multinomial probability $q_k(\theta)$. In particular, let $q_0(\theta)$ be the probability of $k=0$ or an “unobserved” class (that is, $p_0(0, \dots, 0)$ in the notation of Section 1). Then denote

$$h(\underline{x} | \theta, M) := \mathcal{L}^M(\theta | \{f_{k_1, \dots, k_m}\}) = \frac{M!}{\prod_{k=0}^K x_k!} \prod_{k=0}^K q_k(\theta)^{x_k}.$$

Denote by ∂_i the differentiation with respect to θ_i , the i th component of the θ vector, and recall that the Fisher information $I_h(\theta)$ associated with $h(\underline{x} | \theta, M)$ is given by the matrix

$$I_h(\theta) = -E[\partial_i \partial_j \log h(\underline{x} | \theta, M)]. \quad (\text{A.1})$$

Note

$$\begin{aligned} \partial_i \partial_j \log h(\underline{x} | \theta, M) &= \sum_{k=0}^K \partial_i \left(\frac{\partial \log h(\underline{x} | \theta, M)}{\partial q_k(\theta)} \partial_j q_k(\theta) \right) \\ &= \sum_{k=0}^K \partial_i \left(\frac{x_k}{q_k(\theta)} \partial_j q_k(\theta) \right) \sum_{k=0}^K \frac{x_k}{q_k(\theta)} \partial_i \partial_j q_k(\theta) \\ &\quad - \sum_{k=0}^K \left(\frac{x_k}{q_k^2(\theta)} \partial_i q_k(\theta) \partial_j q_k(\theta) \right). \end{aligned}$$

Substituting this into (A.1) and applying the facts that $EX_k = Mq_k(\theta)$ and $\sum_{k=0}^K \partial_i \partial_j q_k(\theta) = \partial_i \partial_j (\sum_{k=0}^K q_k(\theta)) = \partial_i \partial_j (1) = 0$, one obtains

$$I_h(\theta) = \sum_{k=0}^K Mq_k(\theta)^{-1} \partial_i q_k(\theta) \partial_j q_k(\theta). \quad (\text{A.2})$$

The information decomposition, corresponding to the likelihood decomposition

$$\mathcal{L}^M(\theta | \{f_{k_1, \dots, k_m}\}) = \mathcal{L}_b^M(\theta | D) \mathcal{L}_c(\theta | \{f_{k_1, \dots, k_m}\}, D)$$

(cf., (3) in Section 1), is seen to be

$$I_h(\theta) = I_b(\theta) + I_c(\theta), \quad (\text{A.3})$$

where $I_b(\theta)$ and $I_c(\theta)$ are, respectively, the Fisher information matrices corresponding to $\mathcal{L}_b^M(\theta | D)$ and $\mathcal{L}_c(\theta | \{f_{k_1, \dots, k_m}\}, D)$ given by

$$\begin{aligned} I_b(\theta) &= \frac{M}{q_0(\theta)(1-q_0(\theta))} \partial_i q_0(\theta) \partial_j q_0(\theta) \quad \text{and} \\ I_c(\theta) &= \sum_{k=1}^K \frac{M(1-q_0(\theta))^2}{q_k(\theta)} \partial_i \left(\frac{q_k(\theta)}{1-q_0(\theta)} \right) \partial_j \left(\frac{q_k(\theta)}{1-q_0(\theta)} \right). \end{aligned}$$

The natural statistic which measures the loss of efficiency between $I_h(\theta)$ and $I_c(\theta)$, regardless of the value of M (which is typically unknown), is the relative information

$$RI(\theta) = \frac{\|I_c(\theta)\|}{\|I_h(\theta)\|}, \quad (\text{A.4})$$

where $\|\cdot\|$ is any reasonable matrix norm. For instance, in our numerical example below we take it to be the Frobenius norm, already discussed in the main body of the paper. Recall that it is defined for any $A=[a_{ij}]$ as the Euclidean norm of the matrix entries. That is, $\|A\| = \sqrt{\sum_{ij} a_{ij}^2} = \sqrt{\text{Tr}(A^T A)}$ (cf., e.g., Golub and Van Loan, 1996). Note that with this choice of $\|\cdot\|$ and in view of (A.3)

$$0 < RI(\theta) \leq 1$$

for any θ for which the matrix $I_h(\theta)$ is positive definite. For a given PAM, the loss of efficiency at any such point θ may be defined

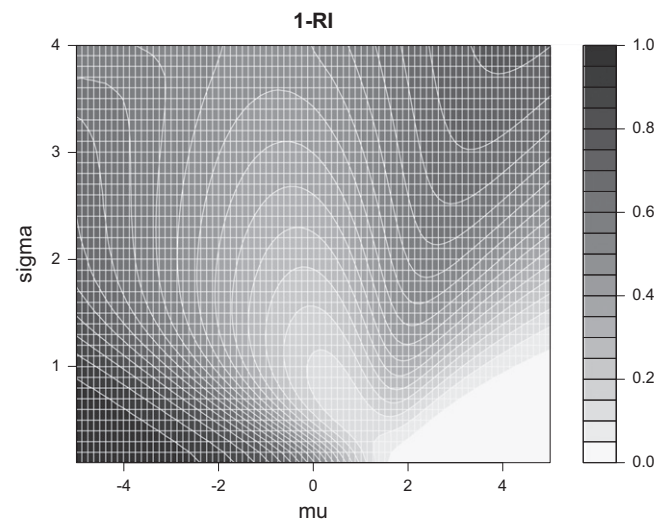


Fig. A1. Loss of efficiency in the conditional Poisson-lognormal model. The values of the relative information statistic $1-RI$ plotted as 2-d contours against the grid of values (μ, σ) in the one dimensional Poisson-lognormal model given in (5). The dark areas correspond to the very severe loss of efficiency and are located in the region of the parameter space which contains the values fitted in Table 3.

therefore as $1-RI(\theta)$. An example for the Poisson-lognormal model follows.

Appendix B. Example: truncated Poisson-lognormal model

The statistic RI may be used in a quite general setting. For illustration, we shall apply it to the particular PAM defined by the Poisson-lognormal model introduced in Section 2.4. Consider the one dimensional model (5) with $\theta = (\mu, \sigma)$ stratified into seven classes ($K=6$) with class probabilities $q_k(\theta) = p(k; \mu, \sigma)$ for $k=0, \dots, 6$ and $q_7(\theta) = \sum_{k \geq 7} p(k; \mu, \sigma)$. The contour plot of the values of $1-RI$ as a function of μ and σ is presented in Fig. A1. The grey scale is increasingly darker as the value of $1-RI$ approaches unity. It is easily observed that when $\mu < -1$, the loss of efficiency is anywhere between 25% and almost 100% (black region in the left lower corner), depending on the value of σ . A quick inspection of Table 3 reveals that this is the region of interest in our TCR data study.

Appendix C. Supplementary material

Supplementary data associated with this article can be found in the online version at <http://dx.doi.org/10.1016/j.jtbi.2013.02.009>.

References

- Aitchison, J., Ho, C., 1989. The multivariate Poisson-log normal distribution. *Biometrika* 76, 643.
- Andrieu, C., De-Freitas, N., Doucet, A., Jordan, M., 2003. An introduction to MCMC for machine learning. *Mach. Learn.* 50, 5–43.
- Barger, K., Bunge, J., 2008. Bayesian estimation of the number of species using noninformative priors. *Biom. J.* 50, 1064–1076.
- Bolotin, D.A., Mamedov, I.Z., Britanova, O.V., Zvyagin, I.V., Shagin, D., Ustyugova, S.V., Turchaninova, M.A., Lukyanov, S., Lebedev, Y.B., Chudakov, D.M., 2012. Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. *Eur. J. Immunol.* 42 (11), 3073–3083.
- Bulmer, M., 1974. On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics* 30, 101–110.
- Chao, A., 2006. Species richness estimation. In: Balakrishnan, N., Read, C., Vidakovic, B. (Eds.), *Encyclopedia of Statistical Sciences*. Wiley, New York.
- Correia-Neves, M., Waltzinger, C., Mathis, D., Benoist, C., 2001. The shaping of the T-cell repertoire. *Immunity* 14, 21–32.

- Davis, M.M., Bjorkman, P.J., 1988. T-cell antigen receptor genes and T-cell recognition. *Nature* 334, 395–402.
- Engen, S., Lande, R., Walla, T., DeVries, P., 2002. Analyzing spatial structure of communities using the two-dimensional Poisson lognormal species abundance model. *Am. Nat.* 160, 60–73.
- 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, 42–58.
- Gelman, A., Rubin, D., 1992. Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–472.
- Golub, G., Van Loan, C., 1996. *Matrix Computations*. Johns Hopkins University Press.
- Hastie, T., Tibshirani, R., Friedman, J., 2009. *Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second ed. Springer-Verlag, New York.
- Ignatowicz, L., Kappler, J., Parker, D.C., Marrack, P., 1996. The responses of mature T-cells are not necessarily antagonized by their positively selecting peptide. *J. Immunol.* 157, 1827–1831.
- Johnson, V., 2004. A Bayesian χ^2 test for goodness-of-fit. *Ann. Stat.* 32, 2361–2384.
- Kawano, T., Cui, J., Koezuka, Y., Toura, I., Kaneko, Y., Motoki, K., Ueno, H., Nakagawa, R., Sato, H., Kondo, E., et al., 1997. CD1D-restricted and TCR-mediated activation of $\nu\alpha 14$ nkT-cells by glycosylceramides. *Science* 278, 1626–1629.
- Kedzierska, K., La Gruta, N.L., Stambas, J., Turner, S.J., Doherty, P.C., 2008. Tracking phenotypically and functionally distinct T-cell subsets via T-cell repertoire diversity. *Mol. Immunol.* 45, 607–618.
- Lai, B., Ding, R., Li, Y., Duan, L., Zhu, H., 2012. A de novo metagenomic assembly program for shotgun DNA reads. *Bioinformatics* 28, 1455–1462.
- Legendre, P., Legendre, L., 1998. *Numerical ecology. Developments in Environmental Modelling*, English edition, vol. 20, Elsevier Science B.V., Amsterdam (translated and revised from the second French (1984) edition).
- Lunn, D., Spiegelhalter, D., Thomas, A., Best, N., 2009. The BUGS project: evolution, critique and future directions. *Stat. Med.* 28, 3049–3067.
- Magurran, A.E., 2005. Biological diversity. *Curr. Biol.* 15, R116–R118.
- Mamedov, I.Z., Britanova, O.V., Bolotin, D.A., Chkalina, A.V., Staroverov, D.B., Zvyagin, I.V., Kotlobay, A.A., Turchaninova, M.A., Fedorenko, D.A., Novik, A.A., Sharonov, G.V., Lukyanov, S., Chudakov, D.M., Lebedev, Y.B., 2011. Quantitative tracking of T-cell clones after haematopoietic stem cell transplantation. *EMBO Mol. Med.* 3, 201–207.
- Murphy, K., Travers, P., Walport, M., et al., 2011. *Janeway's Immunobiology*. Taylor & Francis.
- Nayak, T., 1991. Estimating the number of component processes of a super-imposed process. *Biometrika* 78, 75–81.
- Pacholczyk, R., Ignatowicz, H., Kraj, P., Ignatowicz, L., 2006. Origin and T-cell receptor diversity of Foxp3+CD4+CD25+ T-cells. *Immunity* 25, 249–259.
- Pacholczyk, R., Kern, J., Singh, N., Iwashima, M., Kraj, P., Ignatowicz, L., 2007. Nonspecific antigens are the cognate specificities of Foxp3+ regulatory T-cells. *Immunity* 27, 493–504.
- Plummer, M., 2003. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, March, pp. 20–22.
- R Team, 2010. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rempala, G.A., Seweryn, M., Ignatowicz, L., 2011. Model for comparative analysis of antigen receptor repertoires. *J. Theor. Biol.* 269, 1–15.
- Rodrigues, J., Milan, L., Leite, J., 2001. Hierarchical Bayesian estimation for the number of species. *Biom. J.* 43, 737–746.
- Sanathanan, L., 1972. Estimating the size of a multinomial population. *Ann. Math. Stat.* 142–152.
- Sepúlveda, N., Paulino, C.D., Carneiro, J., 2010. Estimation of T-cell repertoire diversity and clonal size distribution by Poisson abundance models. *J. Immunol. Methods* 353, 124–137.
- Thalamuthu, A., Mukhopadhyay, I., Zheng, X., Tseng, G.C., 2006. Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics* 22, 2405–2412.
- Venturi, V., Chin, H.Y., Asher, T.E., Ladell, K., Scheinberg, P., Bornstein, E., van-Bockel, D., Kelleher, A.D., Douek, D.C., Price, D.A., Davenport, M.P., 2008. TCR β -chain sharing in human CD8+ T-cell responses to cytomegalovirus and EBV. *J. Immunol.* 181, 7853–7862.
- Venturi, V., Quigley, M.F., Greenaway, H.Y., Ng, P.C., Ende, Z.S., McIntosh, T., Asher, T.E., Almeida, J.R., Levy, S., Price, D.A., Davenport, M.P., Douek, D.C., 2011. A mechanism for TCR sharing between T-cell subsets and individuals revealed by pyrosequencing. *J. Immunol.* 186, 4285–4294.
- Wang, C., Sanders, C., Yang, Q., Schroeder Jr., H., Wang, E., Babrzadeh, F., Gharizadeh, B., Myers, R., Hudson Jr., J., Davis, R., et al., 2010. High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T-cell subsets. *Proc. Natl. Acad. Sci.* 107, 1518–1523.
- Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63.
- Warren, R., Nelson, B., Holt, R., 2009. Profiling model T-cell metagenomes with short reads. *Bioinformatics* 25, 458.
- Wong, F.S., Janeway Jr., C.A., 1999. The role of CD4 vs. CD8 T-cells in IDDM. *J. Autoimmun.* 13, 290–295.