# Model for Diversity Analysis of Antigen Receptor Repertoires

Grzegorz A. Rempala*, Michał Seweryn†, and Leszek Ignatowicz‡

February 24, 2010

**Abstract**

In modern molecular biology one of the most common ways of studying a vertebrate immune system is to statistically compare the counts of sequenced antigen receptor clones (either immunoglobulins or T-cell receptors) derived from various tissues under different experimental or clinical conditions. The problem is difficult and does not fit readily into the standard statistical framework of contingency tables primarily due to serious under-sampling of the receptor populations. This under-sampling is caused on one hand by the extreme diversity of antigen receptor repertoires maintained by the immune system and, on the other, by the high cost and labor intensity of the receptor data collection process. In most of the recent immunological literature the differences across antigen receptor populations are examined via non-parametric statistical measures of species overlap and diversity borrowed from ecological studies. While this approach is robust in a wide range of situations, it seems to provide little insight into the underlying clonal size distribution and the overall mechanism differentiating the receptor populations. As a possible alternative, the current paper presents a parametric method which adjusts for the data under-sampling as well as provides a unifying approach to simultaneous comparison of multiple receptor groups by means of the modern statistical tools of unsupervised learning. The parametric model is based on a flexible multivariate Poisson-lognormal distribution and is seen to be a natural generalization of the univariate Poisson-lognormal models used in ecological studies of biodiversity patterns. The procedure for evaluating model's fit is described along with the public domain software developed to perform the necessary diagnostics. The model-driven analysis is seen to compare favorably vis a vis traditional methods when applied to the data from T-cell receptors in transgenic mice populations.

**Keywords**: T-cells, antigen receptors, computational immunology, species diversity estimation, Poisson abundance models, Lognormal distribution, dissimilarity measure, dendrogram, mutual information.

**2010 AMS Subject Classification**: 62P10, 92B05

* Corresponding author. Department of Biostatistics and the Cancer Center, Medical College of Georgia, Augusta, GA 30912. E-mail:grempala@mcg.edu

† Wydział Matematyki, Universytet Łódzki, Łódz, Poland. E-mail msewery@math.uni.lodz.pl

‡ Department of Medicine, Center for Biotechnology and Genomic Medicine, Medical College of Georgia. E-mail: lignatowicz@mcg.edu

# 1 Introduction

The major feature of the adaptive immune system is its capacity to generate clones of B and T-cells that are able to recognize and neutralize specific antigens. Both cell types recognize antigens by a special class of surface molecules called B- and T-cell receptors. The methodology developed in this paper will apply to both types of receptors, for the sake of clarity and simplicity, we describe the background and the overall problem in terms of T-cell receptors. For a general introduction to the molecular biology of the immune system, we refer interested reader to e.g., Janeway (2005).

A single T-cell receptor (TCR) is composed of two chains, $\alpha$ and $\beta$, that are formed during T-cell differentiation. Both chains are formed by rearrangements of genetic segments, V$\alpha$ and J$\alpha$ for TCR$\alpha$ chain and V$\beta$, D$\beta$ and J$\beta$ for TCR$\beta$ chain. Since there are a number of segments of each type in the genomic DNA , a great number of different $\alpha$ and $\beta$ chains are generated. This chain diversity is further increased by the recombination process when individual nucleotides might be added or deleted at the junctional sites. The region containing these highly variable junctions is the third of three complementarity-determining regions (CDRs) that are seen crystallographically to contact antigen. The sources of TCR diversity are thus naturally broken down hierarchically into gene segment family (library), segment within family, CDR3 length and CDR3 nucleotide diversity. Both combinatorial and insertional re-arrangements result in the huge TCR repertoire ensuring that immune system has a potential to recognize a large number of antigens. For instance, it is estimated that in mice the number of different TCRs that can be formed exceeds $10^{15}$ (Davis and Bjorkman, 1988; Casrouge et al., 2000). For humans, it is estimated that over $10^{18}$ different TCRs can be produced and the number of different TCR species (*TCR richness*) in a human at any given time has been estimated to exceed $10^7$ (Arstila et al., 1999; Naylor et al., 2005). This *clonal diversity* of TCR populations makes them particularly challenging objects to analyze statistically.

In what follows, we are concerned with the statistical analysis of the diversity of TCR repertoire samples obtained from various subsets of T-cells as *counts* of different TCR clones. These T-cell subsets are generally defined based on the expression of cell surface markers leading to different functions in the immune response. For example, naive T-cells are defined as cells that did not encounter an antigen in their lifetime, while memory T-cells are cells that previously responded to an antigenic stimulation and underwent clonal expansion. The frequency of individual T-cell clones in normal individuals is very low. However, once a naive T-cell expressing appropriate TCR encounters antigen, it becomes activated and expands forming multiple clone cells. Thus, the analysis of TCR repertoire *clonal size distribution* has become a crucial element in many studies aimed at better understanding the evolution of the immune responses in vaccinated individuals or patients suffering from autoimmune diseases or cancer (cf., e.g., Butz and Bevan 1998). The clonal size estimates has been recently applied to quantify differences in TCR repertoires of normal and infected individuals and to help determine which T-cell clones persisted as memory T-cells (McHeyzer-Williams et al. 1999; Sebzda et al. 1999; Busch and Pamer 1999; Savage et al. 1999). The comparisons of the diversity of TCR repertoires has been also used to determine the origin of T-cells and to study heterogeneity of memory T-cells (Sallusto et al. 1999, 2004). For the purpose of the current paper, under the term "clonal diversity" we will understand both the TCRs clonal size distribution (abundance pattern) as well as their richness (number of different TCRs).

There are several statistical approaches to assess T-cell diversity based on the TCR clones counts. For instance, in the non-parametric approach, Simpson's diversity and Shannon's entropy indices have been applied to measure the diversity of collected samples (Ferreira et al., 2009; Venturi et al., 2008). However, such indexes are typically just summary statistics and, due to frequent data under-sampling, provide only limited information about the true diversities of the populations. Another approach seen

in immunological studies relies on modeling diversity parametrically assuming that all clonotypes are equally represented in the repertoire (Barth et al. 1985; Behlke et al. 1985). The advantage of this homogenous model is its computational and conceptual simplicity which contributes to its wide use in analyzing TCR data (e.g., Casrouge et al. 2000; Hsieh et al. 2004, 2006; Pacholczyk et al. 2007, 2006). However, this simple model is called into question by the empirical evidence (see e.g., Naumov et al. 2003; Pewe et al. 2004) suggesting heavy right tails of the clonal size distributions. To account for this heterogeneity, the homogenous model has been expanded with a variety of mixture models, typically under the assumption of the Poisson-distribution of the TCR clones. These so-called *Poisson abundance mixture models* (Chao, 2006) assume that each TCR variant (i.e., each clone family) is sampled according to the Poisson distribution with a specific sampling rate, itself varying according to a prescribed parametric (mixing) distribution e.g., exponential, gamma, or lognormal (Ord and Whitmore 1986; Sepúlveda et al. 2009; Bulmer 1974). A recent detailed comparative study of Sepúlveda et al. (2009) identified one of such models, the Poisson-lognormal mixture (PLN), as particularly well suited for modeling clonal diversity. The special appeal of the PLN is in that it may be naturally extended to multivariate setting, allowing therefore for the simultaneous analysis of abundance patterns of several repertoires (Engen et al., 2002). In particular, the bivariate extension of the PLN model may be used to derive a class of pairwise dissimilarity measures between repertoires and to construct tree-based hierarchy relating various TCR repertoires.

The purpose of this article is to describe a new general approach to analyzing and comparing TCR-type data arriving from multiple repertoires. The approach we develop here relies on the TCR repertoires dissimilarities analysis where the appropriate tree-distance measures are derived under a simple, easily testable and interpretable model of clonal abundance based on the parametric bivariate Poisson-lognormal distribution (BPLN). By means of examples derived from real TCR data, we argue that under BPLN both the moment-based and the information-based parametric measures of dissimilarity yield consistent and biologically meaningful results. The paper is organized as follows. In the next section (Section 2) we give a brief overview of the Poisson abundance models both in univariate and multivariate (bivariate) settings. In Section 3 we discuss one popular method for deriving dissimilarity measures which is particularly relevant for TCR data studies and provide the formal definitions of the four different measures considered in the paper. In Section 4 we present the application of our method to TCR data obtained from the populations of naive and so-called regulatory T-cell receptors in healthy and immune-deficient mice. This data set was described in detail and analyzed by different methods in Pacholczyk et al. (2006). We re-analyze it using clustering algorithms derived under both non-parametric and BPLN models and compare the results. In Section 5 we provide a concise summary of our findings and offer some concluding remarks. Some elementary derivations related to the entropy function are provided for readers convenience in the appendix.

## 2 Poisson Models of Abundance

Poisson abundance models arrive naturally in the biodiversity studies if we assume (see, e.g., Chao 2006) that the species sampling is done by a "continuous type of effort" i.e., data is recorded as arriving from a mixture of Poisson processes in time interval from 0 to $T$ (in what follows we take $T = 1$). This type of model approach can be traced back to Fisher, Corbet and Williams (Fisher et al., 1943). Consider $M$ species labeled from 1 to $M$. Individuals of the $i$-th species arrive in the sample according to a Poisson process with a discovery rate $\lambda_i$. If the detectability of individuals can be assumed to be equal across all species (which is typically a case in TCR repertoires analysis), then the rates can be interpreted as

species abundances (Nayak, 1991). In this sampling scheme, the sample size $n$ (the number of individuals observed in the experiment) is a random variable. Since the conditional frequencies follow a multinomial distribution with class total $n$ and class probabilities given by relative frequencies $\lambda_i/\sum_{k=1}^{M}\lambda_k$, many estimators are shared in both the continuous-type Poisson models and the discrete-type (multinomial) models where $n$ is assumed to be a constant. We note here that in case of antigen receptors data, the constant $n$ is sometimes known (e.g., DNA sequencing data) and sometimes not known (e.g., spectratype data, see, Kepler et al. 2005). In the latter case the prior distributional form of $n$ is typically assumed and the posterior model is investigated based on the Bayesian tools (see, e.g., Rodrigues et al. 2001; Lewins and Joanes 1984; Barger and Bunge 2008; Solow 1994 ). Since the present paper is motivated by the single–cell DNA sequencing data, we are assuming throughout that $n$ is known. The extension of our model to unknown $n$ along the lines of Rodrigues et al. (2001) is reasonably straightforward but not pursued here.

## 2.1 Univariate mixture models

Since it has been generally accepted that the antigen receptor clonal size distributions have heavy right tails, to adjust for the over-dispersion the species rates $(\lambda_1, \lambda_2, \ldots, \lambda_M)$ are typically modeled as a random sample from a mixing distribution with density $f(\lambda;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a low-dimensional vector of parameters. Following the famous paper by Fisher and his colleagues (Fisher et al., 1943) many researchers have adopted a gamma density as a mixing model. Other parametric models include among others the log-normal (Bulmer, 1974), inverse-Gaussian (Ord and Whitmore, 1986), and generalized inverse-Gaussian (Sichel, 1997) distributions as well as many others (Sepúlveda et al., 2009). An obvious advantage of such parametric models is that the inference problem reduces to estimating only a few relatively low-dimensional parameters for which the traditional estimation procedures can be typically applied. For any mixture density $f(\lambda;\boldsymbol{\theta})$, define $p_\theta(k), k = 0, 1, \ldots$ as the probability that any TCR species is observed $k$ times in the sample, that is

$$p_\theta(k) = \int_0^\infty [\lambda^k e^{-\lambda}/k!]\, f(\lambda;\boldsymbol{\theta}) d\lambda \qquad k = 0, 1, \ldots. \tag{2.1}$$

Denoting by $f_k$ $(k = 1, 2, \ldots, n)$ the number of receptor species observed exactly $k$-times in the sample we have $E(f_k) = Mp_\theta(k)$. Setting $D = \sum_k f_k$ the likelihood function for $M$ and $\boldsymbol{\theta}$ can be written as

$$L(M, \boldsymbol{\theta}|\{f_k\}) = \frac{M!}{(M-D)!\prod_{k\geq 1}(f_k!)}[p_\theta(0)]^{M-D}\prod_{k\geq 1}[p_\theta(k)]^{f_k}.$$

The (unconditional) MLEs for $M$ and $\boldsymbol{\theta}$ and their asymptotic variances are obtained based on the above likelihood which as we can see depends on the data only through the observed values of $\{f_k\}$. The likelihood can be factored as

$$L(M, \boldsymbol{\theta}|\{f_k\}) = L_b(M, \boldsymbol{\theta}|D) L_c(\boldsymbol{\theta}|\{f_k\}, D)$$

where $L_b(M, \boldsymbol{\theta}|D)$ is a likelihood with respect to $D$, a binomial $(M, 1-p_\theta(0))$ variable, and $L_c(\boldsymbol{\theta}|\{f_k\}, D)$ is a (conditional) multinomial likelihood with respect to $\{f_k; k \geq 1\}$ with cell total $D$ and zero-truncated cell probabilities $p_\theta(k)/[1-p_\theta(0)], k \geq 1$, i.e.,

$$L_c(\boldsymbol{\theta}|\{f_k\}, D) = \frac{D!}{\prod_{k\geq 1}(f_k!)}\prod_{k\geq 1}\left[\frac{p_\theta(k)}{1-p_\theta(0)}\right]^{f_k}. \tag{2.2}$$

The MLE obtained from this likelihood can be regarded as a (conditional) *empirical Bayes estimator* if we think of the mixing distribution as a prior distribution having unknown parameters that must be estimated. See, e.g., Rodrigues et al. (2001) for further reference.

### 2.2 Extension to bivariate models

As we shall see in the next section, it is of interest to also consider multivariate models of abundance. For the purpose of our discussions below we focus on the bivariate models but the modifications for higher dimensions are rather straightforward. For simplicity assume that we have the same $M$ species in both populations. In direct analogy with the notation of the previous section, define now $p_\theta(k,l)$ to be the probability that any TCR species (i.e., a TCR clone) is present $k$ times in the sample from the first population (repertoire) and $l$ times in the sample from the second one. Let $f_{k,l}$ be the empirical count and set now $D = \sum_{k,l\geq 0} f_{k,l}$ (assuming $f_{0,0} = 0$). Let $f(\lambda_1, \lambda_2, \boldsymbol{\theta})$ be the bivariate mixture distribution. The likelihood formulae from previous section extends to the bivariate case as

$$L(M, \boldsymbol{\theta}|\{f_{k,l}\}) = \frac{M!}{(M-D)!\prod_{k,l\geq 0}(f_{k,l}!)}[p_\theta(0,0)]^{M-D}\prod_{k,l\geq 0}[p_\theta(k,l)]^{f_{k,l}}.$$

where

$$p_\theta(k,l) = \int_0^\infty [\lambda_1^k e^{-\lambda_1}\lambda_2^l e^{-\lambda_2}/(k!\,l!)]\, f(\lambda_1,\lambda_2;\boldsymbol{\theta})d\lambda_1 d\lambda_2 \qquad k,l = 0,1,\ldots. \tag{2.3}$$

Note that, as before, $E(f_{k,l}) = Mp_\theta(k,l)$. The likelihood function for $M$ and $\boldsymbol{\theta}$ can be again factored as

$$L(M, \boldsymbol{\theta}|\{f_{k,l}\}) = L_b(M, \boldsymbol{\theta}|D)L_c(\boldsymbol{\theta}|\{f_{k,l}\}, D)$$

where, in obvious analogy with (2.2), $L_b(M, \boldsymbol{\theta}|D)$ is now a likelihood with respect to $D$, a binomial variable with parameters $(M, 1-p_\theta(0,0))$ and $L_c(\boldsymbol{\theta}|\{f_{k,l}\}, D)$ is a (conditional) multinomial likelihood with respect to $\{f_{k,l}, k+l>0\}$ with cell total $D$ and the bivariate, zero-truncated cell probabilities $\{p_\theta(k,l)/[1-p_\theta(0,0)]\}_{k,l}, k+l>0$, i.e.,

$$L_c(\boldsymbol{\theta}|\{f_{k,l}\}, D) = \frac{D!}{\prod_{k,l\geq 0} f_{k,l}!}\prod_{k\geq 1}\left[\frac{p_\theta(k,l)}{1-p_\theta(0,0)}\right]^{f_{k,l}}. \tag{2.4}$$

## 3 Diversity Analysis and Clustering

When studying evolution of TCR species it is of interest to compare their diversity, by which we mean herein (cf. Section 1) the clonal size distribution $\{p_\theta(k)\}$ and the species number $M$. Such repertoire diversity comparisons are of great interest for instance in clinical studies, where the quantities of interest are the "divergences" of multiple observed TCR repertoires from some control or asymptotic one. The individual repertoires of antigen receptors can be then characterized in terms of their divergence from the control (Chen et al., 2003; Komatsu et al., 2009; Pacholczyk et al., 2007, 2006). Under our definition the TCR repertoire diversity is completely determined by the parameters $(M, \boldsymbol{\theta})$. This agrees with the original concept of "species diversity" known from the field of ecology where the term itself relates both to the number of species (richness) and to their apportionment within the sequence (evenness or equitability, see Sheldon 1969). A sensible method of comparing diversity of multiple repertoires simultaneously is based on a concept of (pairwise) *diversity dissimilarity measure* and the *hierarchical*

*clustering* induced by it. The hierarchical clustering which we discuss in more detail below, is one of many modern methods of analyzing patterns in high-dimensional data on the grounds of the so-called unsupervised statistical learning theory (cf., e.g., Hastie et al. 2001), a very dynamically developing area of modern statistics.

## 3.1 Diversity dissimilarity measures

Assume that the overall "similarity" between a pair of TCR repertoires with respective clonal abundance distributions $p$ and $q$ is quantified by some non-negative function $Q(p,q)$ referred to as the *similarity index* or *similarity measure*. Since typically the samples from the joined distribution (frequency) of abundance are not available in data collected from TCR repertoires, some of the crude similarity indices are based simply on the joined 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). Examples of such indices are the classical Jaccard index and the closely related Sørensen index, the two oldest and most widely used similarity indices in ecological biodiversity studies (Magurran, 2005). One of the advantages of the Poisson mixture model described in the previous section, is that it allows for defining meaningful indices incorporating pairwise comparisons of the TCR species based on joined distribution of abundance. As representative examples of such indices we consider here the Morisita-Horn index ($\mathcal{D}_{MH}$) (Magurran, 2005), the mutual information criterion ($\mathcal{D}_{MI}$) which is a special case of a Kullback-Leibler divergence (see, e.g., Koski, 2001), and the overlap index ($\mathcal{D}_{OV}$) introduced by Smith et al. (1996). All of these indices give rise to the corresponding measures of dissimilarity concentrated on the unit interval with the perfect correlation (or complete overlap) between the frequency distributions yielding the value zero. Indeed, they are all seen as special cases of the following general construction. For any bivariate probability distribution $p_\theta$ with corresponding marginal distributions $p_\theta^{(1)}$ and $p_\theta^{(2)}$ consider a similarity index $\mathcal{Q}(p_\theta^{(1)}, p_\theta^{(2)})$ satisfying

$$0 \le \mathcal{Q}(p_\theta^{(1)}, p_\theta^{(2)}) \le \frac{\mathcal{Q}(p_\theta^{(1)}, p_\theta^{(1)}) + \mathcal{Q}(p_\theta^{(2)}, p_\theta^{(2)})}{2} \tag{3.1}$$

with the right bound attained when $p_\theta^{(1)} = p_\theta^{(2)}$. Then the corresponding (normalized) $\mathcal{Q}$-induced measure of dissimilarity between the pair $(p_\theta^{(1)}, p_\theta^{(2)})$ may be defined as

$$\mathcal{D}(p_\theta^{(1)}, p_\theta^{(2)}) = 1 - \frac{2\,\mathcal{Q}(p_\theta^{(1)}, p_\theta^{(2)})}{\mathcal{Q}(p_\theta^{(1)}, p_\theta^{(1)}) + \mathcal{Q}(p_\theta^{(2)}, p_\theta^{(2)})}. \tag{3.2}$$

To obtain the *Morisita-Horn dissimilarity index* ($\mathcal{D}_{MH}$) we take in (3.2) $\mathcal{Q} = \mathcal{Q}_{MH}$

$$\mathcal{Q}_{MH}(p_\theta^{(1)}, p_\theta^{(2)}) = \sum_{k,l \ge 1} k\, l\, p_\theta(k,l). \tag{3.3}$$

which obviously satisfies (3.1) for any non-negative random variables. A closely related *correlation-based dissimilarity index* $\mathcal{D}_\rho$ is obtained when we take $\mathcal{Q} = \mathcal{Q}_\rho$ with

$$\mathcal{Q}_\rho(p_\theta^{(1)}, p_\theta^{(2)}) = |\sum_{k,l \ge 0} \tilde{k}\tilde{l}\, p_\theta(k,l)|. \tag{3.4}$$

where $\tilde{k} = (k - m_1)/s_1$ and $\tilde{l} = (l - m_2)/s_2$ and $m_i$ and $s_i$ are, respectively, mean and standard deviation of $p_\theta^{(i)}$, $i = 1, 2$. In this case the inequality (3.1) simply asserts that $0 \le \mathcal{Q}_\rho \le 1$.

A popular dissimilarity measure which we shall denote here by $\mathcal{D}_{OV}$ is obtained by averaging conditional probabilities of individual receptor species presence in both samples, given its presence in one. The measure was introduced by Smith et al. (1996) to quantify 'overlap' between repertoires and is obtained by taking the following similarity index.

$$\mathcal{Q}_{OV}(p_\theta^{(1)}, p_\theta^{(2)}) = \frac{\sum_{k,l>0} p_\theta(k,l)}{2} \left( \frac{1}{\sum_{k>0} p_\theta^{(1)}(k)} + \frac{1}{\sum_{l>0} p_\theta^{(2)}(l)} \right) \tag{3.5}$$

which again trivially satisfies (3.1).

Finally, in order to obtain the *mutual information dissimilarity index* ($\mathcal{D}_{MI}$) we take in (3.2) $\mathcal{Q} = \mathcal{Q}_{MI}$ where

$$\mathcal{Q}_{MI}(p_\theta^{(1)}, p_\theta^{(2)}) = \sum_{k,l\geq 0} p_\theta(k,l) \log \left( \frac{p_\theta(k,l)}{p_\theta^{(1)}(k)\, p_\theta^{(2)}(l)} \right). \tag{3.6}$$

The fact that the above function satisfies (3.1) is shown in the appendix. Note that all the above similarity indices $\mathcal{Q}_{MH}, \mathcal{Q}_\rho, \mathcal{Q}_{OV}$ and $\mathcal{Q}_{MI}$ (and thus also their corresponding dissimilarity measures) depend on the underlying mixing distribution parameters $\theta$ but not explicitly on the species number $M$. This is desirable since the quantity $M$ is typically unknown and difficult to estimate due to often very severe undersampling of the TCR repertoires (cf., e.g., Sepúlveda et al. 2009). In the parametric setting considered here, if needed, the value of $M$ may be estimated (a-posteriori) by either of the estimates

$$\hat{M}_1 = D/(1 - p_\theta(0,0)) \tag{3.7}$$

$$\hat{M}_2 = \sum_{k,l\geq 0, k+l>0} f_{k,l}/p_\theta(k,l). \tag{3.8}$$

whose close numerical agreement usually indicates a robust fit of the bivariate parametric mixture model. Note that $\hat{M}_2$ is simply a parametric version of the Horvitz-Thompson estimator (Horvitz and Thompson, 1952).

## 3.2 Hierarchical clustering

For a given pairwise dissimilarity measure $\mathcal{D}$ of TCR repertoires, it is a standard unsupervised statistical learning approach to simultaneously compare $N$ repertoires in terms of $\mathcal{D}$ by means of building hierarchical clusters which are graphically represented by *dendrogarms* or "tree diagrams". In such hierarchical clustering procedure the TCR data are not partitioned into a particular cluster in a single step, instead, as a name suggests, a hierarchical structure is produced in which the clusters at each level of the hierarchy are created by merging clusters at the next lower level. The main advantage of hierarchical clustering approach lies in the fact that no cluster number needs to be specified in advance. Hierarchical clustering is performed via either agglomerative methods, which proceed by series of fusions of the $N$ objects into groups, or divisive methods, which separate objects successively into finer groupings. Agglomerative techniques are more commonly used, and this is the method we consider below for TCR repertoires. For a general introduction to clustering and unsupervised learning, interested reader is referred to chapter 14 in the popular monograph Hastie et al. (2001).

## 3.3 Poisson-lognormal model

Whereas there are many possible models of parametric bivariate mixture, the recent studies in Engen et al. (2002) and Sepúlveda et al. (2009) seem to indicate that lognormal mixing distributions may be

often an appropriate choice for TCR repertoires modeling. In that spirit we consider herein a bivariate model based on log-binormal variates. Under the assumption of random sampling, the number of individuals sampled from a given receptor species with abundance $\lambda$ is Poisson distributed with mean $\lambda$. If we assume that $\ln \lambda$ is normally distributed with mean $\mu$ and variance $\sigma^2$ among species, then the vector of individuals sampled from all $M$ species constitutes a sample from the Poisson lognormal distribution with parameters $\theta = (\mu, \sigma^2)$, where $\mu$ and $\sigma^2$ are the mean and variance of the log abundances. The corresponding mass function is of the general form (2.1) and may be written as

$$p(k; \mu, \sigma^2) = \int_{-\infty}^{\infty} g_k(\mu, \sigma, u)\phi(u)du \tag{3.9}$$

where $\phi(\cdot)$ is a 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-parametrized Poisson distribution. Similarly, when we consider pairs of counts of individual receptors from two different repertoires we may think of them as a random sample (of size $M$) from the bivariate Poisson-lognormal distribution (BPLN) with probability mass function given as in (2.3). That is, we assume that the log abundances among species have the binormal distribution with parameters $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. For a general detailed description of the multivariate Poisson distributions, see Aitchison and Ho (1989). Let $\phi(u, v; \rho)$ denote the normal bivariate density with correlation coefficient $\rho$, zero means and unite variances. The distribution of BPLN is given in terms of the bivariate probability mass function $p_\theta(k, l) = p(k, l; \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ for $k, l \geq 0$ where

$$p(k, l; \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_k(\mu_1, \sigma_1, u)g_l(\mu_2, \sigma_2, v)\phi(u, v; \rho)du\, dv. \tag{3.10}$$

From the above formula it follows in particular that both marginals of BPLN are the univariate Poisson-lognormal distributions (3.9) with respective parameters $(\mu_i, \sigma_i^2)$ $(i = 1, 2)$. Since $M$ is usually unknown, when fitting the model we only consider the number of individuals for the observed receptor species and thus the distribution of the number of observed individual receptors follows the zero-truncated BPLN distribution with probability mass function

$$\frac{p(k, l; \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)}{1 - p(0, 0; \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)}. \tag{3.11}$$

The maximum-likelihood estimators (MLEs) of the parameters of this distribution were discussed e.g., in Bulmer (1974) and more recently in Karlis (2003) and Engen et al. (2002). The latter approach is conveniently implemented in the freely available R package *poilog* (R Development Core Team, 2009) which we have used in the current paper to perform all the necessary parameter fitting. In our setting, the model parameters were calculated from the multinomial conditional likelihood function (2.4) where the truncated probability quantity $p_\theta(k, l)/(1 - p_\theta(0, 0))$ is given by (3.11).

Under the assumed BPLN model the measures of dissimilarity may be computed either directly or by Monte-Carlo approximations. Denoting the means of the BPLN marginals by

$$\alpha_i = \exp[\mu_i + \sigma_i^2/2] \quad i = 1, 2, \tag{3.12}$$

the moment-based dissimilarity measures $\mathcal{D}_{MH}$ and $\mathcal{D}_\rho$ are given by

$$\mathcal{D}_{MH} = 1 - \frac{2\alpha_1\alpha_2 \exp[\rho\sigma_1\sigma_2]}{\alpha_1(1+\alpha_1\exp[\sigma_1^2]) + \alpha_2(1+\alpha_2\exp[\sigma_2^2])} \tag{3.13}$$

$$\mathcal{D}_{\rho} = 1 - \frac{2\alpha_1\alpha_2 \left(\exp[\rho\sigma_1\sigma_2] - 1\right)}{\sqrt{\alpha_1(1+\alpha_1(\exp[\sigma_1^2]-1))}\sqrt{\alpha_2(1+\alpha_2(\exp[\sigma_2^2])-1)}} \tag{3.14}$$

where in (3.14) the quantities

$$\alpha_i(1+\alpha_i(\exp[\sigma_i^2]-1)) \quad i=1,2 \tag{3.15}$$

are seen to be the marginal variances of the BPLN distribution. The formulae (3.13-3.14) provide for a convenient way of estimating the dissimilarities $\mathcal{D}_{MH}$ and $\mathcal{D}_{\rho}$ from the data simply by replacing the unknown distribution parameters by their sample estimates calculated via maximum likelihood, for instance, by using the numerical algorithms implemented in the "poilog" R-package.

Unfortunately, due to the fact that the mass function (3.10) is not available in a closed form, there are no similar formulae for the indices $\mathcal{D}_{MI}, \mathcal{D}_{OV}$. These indices are typically approximated by the Monte-Carlo-based bootstrap procedures which are also used to derive confidence intervals and standard errors of the parameter estimates in the model. The general justification for such derivations is provided e.g., in Gill et al. (2009) and Rempala and Szatzschneider (2004). The reader is referred to Efron and Tibshirani (1997) for a general introduction to the theory and practice of statistical bootstrap which is the computational technique we rely upon heavily in our modeling approach and the data analysis described below.

# 4 Application to TCR Repertoire Data

In this section we illustrate the parametric inference on T-cell receptor data based on BPLN model and the associated pairwise dissimilarity measures by analyzing diversity of the repertoires of T-cell receptors derived from two types of genetically-engineered (TCR-mini) mice: the "wild" type with genetically restricted TCR repertoire and unaltered repertoire of self antigens bound to class II MHC, and the "Ep" (B63VJEp) mice that in addition to restricted TCR repertoire also express only single, covalently linked to MHC Ep peptide (Pacholczyk et al., 2006).

## 4.1 Data description and processing

The description of the transgenic "TCRmini" mice modifications was already detailed in the recent work Pacholczyk et al. (2007). Herein, we mention only briefly that the "TCRmini" mice is a new generation of TCR transgenic mouse where all T-cells express one pre-specified TCR$\beta$ chain (specifically, the chain V$\beta$14D$\beta$2J$\beta$2.6), and the unique TCR$\alpha$ mini-locus. This mini-locus allows only for restricted rearrangements of a single V$\alpha$2.9 segment to one of the two J$\alpha$ (J$\alpha$26 and J$\alpha$2) segments. These mice have no other loci that encode TCR$\alpha$ chains and therefore their entire diversity of TCRs is derived from the artificially introduced TCR$\alpha$ mini locus, resulting in greatly altered TCRs abundance.

For the purpose of testing our statistical model, two subpopulations of CD4+ T-cells (i.e., T-cells expressing a surface marker CD4) were collected representing, respectively, regulatory (TR) and naive (TN) T-cells (where the TR cells are defined as those expressing the additional marker Foxp3). In addition, these two subpopulations of CD4+ T-cells were isolated either from (1) the peripheral lymph nodes or (2) from the thymus, giving us a total of *eight* TCR populations differing by the animal type,

| | Ep | | | | Wt | | | |
|---|---|---|---|---|---|---|---|---|
| | TN1 | TR1 | TN2 | TR2 | TN1 | TR1 | TN2 | TR2 |
| $\hat{\mu}$ | -4.54 | -3.90 | -3.58 | -3.80 | -3.70 | -2.97 | -3.55 | -2.90 |
| Lo | -5.37 | -4.81 | -4.59 | -4.71 | -4.86 | -3.77 | -4.46 | -3.81 |
| Up | -3.31 | -2.43 | -2.40 | -2.54 | -2.59 | -2.13 | -2.38 | -1.87 |
| $\hat{\sigma}^2$ | 2.03 | 1.93 | 1.66 | 1.70 | 1.99 | 1.48 | 1.84 | 1.31 |
| Lo | 1.60 | 1.52 | 1.17 | 1.21 | 1.61 | 1.15 | 1.46 | 0.90 |
| Up | 2.36 | 2.29 | 1.99 | 2.02 | 2.47 | 1.74 | 2.15 | 1.70 |

Table 1: Maximum likelihood estimates of the means and variances of the Ep (TCR-restricted) and Wt (wild-type) mice repertoires along with bias corrected 95% confidence bounds generated via parametric bootstrap.

T-cell type, and tissue location. In the thymus, gross of CD4+ T-cells undergo development and in the lymph nodes CD4+ T-cells are retained unless activated by specific antigen and therefore two markedly different patterns of clonal abundance in these organs are generally expected. Additionally, because in the "Ep" mice express only single class II MHC/peptide complex, the diversity of CDR3 region of TCR$\alpha$ chain is drastically reduced in comparison with TCR-mini "wild" mice. For these reasons the dataset is uniquely suitable for testing the BPLN model.

The TCR data from both types of mice was collected as follows (for more details on a similar data harvesting procedure, see also e.g., Warren et al. (2009); Freeman et al. (2009)). Using specific fluorescent reagents, populations of T-cells from different organs were separated into individual wells on 96-well plates and amplified in their unique CDR3 regions of their TCR$\alpha$ chain using single-cell RT-PCR (see, e.g., Kuczma et al. 2009). Following this amplification, the CDR3 regions were sequenced and analyzed providing the distribution of these regions in native subpopulation(s) of T-lymphocytes. This type of procedure has been widely considered to be one of the most reliable methods to harvest T-cell repertoires (Luczynski et al., 2007). Single T-cells can be separated from cell suspension or be isolated from tissue sections. Both the $\beta$ and the $\alpha$ chains can be amplified and sequenced to provide unambiguous identification of T-cell clones. This method avoids the problems of skewed PCR amplification and varying TCR mRNA expression in different cells. Its obvious drawback is the under-sampling issue alluded to already in Section 1: a very large number of cells need to be analyzed to ensure detection of rare clones and to provide a global representation of the T-cell repertoire. We note that with the availability of the next generation sequencing technology (Wong et al., 2007) the large number of single cell RT-PCR could be replaced with the high throughput PCR from heterogeneous population of T-cells. However, at its current stage the technology is not yet recommended for repertoires analysis due to difficulties of matching specific TCR$\alpha$ and TCR$\beta$ chains when amplified simultaneously. In addition, there is also a high risk of counts bias due to the skewed amplification process in high throughput data which tends to over-express the most dominant DNA sequences and under-express (or even remove) the rare ones.

## 4.2 Analysis under BPLN model

In order to assess the usefulness of a proposed parametric method of TCR data modeling, we have first generated the BPLN model-based estimates of the dissimilarity matrix for the eight TCR repertoires using separately each of the four dissimilarity measures $\mathcal{D}$ of the general form (3.2) described in Section 3.

For the moment-based dissimilarities $\mathcal{D}_{MH}$ and $\mathcal{D}_{\rho}$ we have used the explicit formulae (3.13) and (3.14) whereas for the remaining two measures $\mathcal{D}_{OV}$ and $\mathcal{D}_{MI}$ we have directly approximated the quantities $p(k, l; \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ given by (3.10) and subsequently used the parametric bootstrap procedure to produce the empirical approximations of dissimilarities. In all cases the BPLN distribution parameters were estimated by the maximum likelihood estimators (MLEs) computed by maximizing the conditional multinomial likelihood function (2.4) based on the zero-truncated probabilities (3.11). In principle, one could use the conditional likelihood of multivariate Poisson-lognormal distribution directly to estimate all of the parameters simultaneously, however, due to the complicated form of the resulting mixture probabilities, we have deemed that approach to be too unreliable numerically. On the other hand, the iterative bivariate model fitting (fitting one BPLN model at a time, conditionally on the remaining repertoires and iterating until convergence) was seen to be a reasonable fast and numerically stable procedure yielding a set of estimates consistent with marginally MLE-fitted parameters, regardless of the order in which conditioning was performed.

The results of the conditional MLE procedure are partially summarized in Table 1 where the estimates of the $\mu$ and $\sigma^2$ parameters for Poisson-lognormal abundance distributions for the eight repertoires are reported along with the corresponding confidence intervals obtained via the parametric bootstrap bias-corrected percentile method (see e.g., Rempala and Szatzschneider 2004 for details on bootstrap-based interval estimation). Note that the parameters $\mu$ and $\sigma^2$ are related to the marginal estimates of means and variances of the Poisson-lognormal variates by the formulae (3.12) and (3.15), respectively. In order to conserve space, the estimates of the BPLN correlation coefficients are not shown as they are similar in relative values to the moment-based dissimilarities summarized below in Figure 2. We note that the marginal values of the repertoire-specific parameters in both types of repertoires (wild-type and Ep) were found to be of similar magnitude (with estimated values of $\mu$ parameters between -4.5 and -2.9 and $\sigma^2$ parameters between 1.3 and 2.) Overall, the numerical values of the parameters indicated smaller Poisson-lognormal means for the restricted-repertoire mice as compared with the wild-type. Additionally, the naive T-cell repertoires generally seemed to have smaller means and larger variances of the mixing log-normal distributions then the regulatory T-cell repertoires.

The goodness of fit statistics were calculated for the bivariate marginal fit via the bootstrap distribution of the conditional likelihood statistic (2.4). In all cases the differences between the bivariate data and fitted models were not-significant (all $p$-values $<.05$) as measured by the bootstrap tests, indicating reasonable good fit of the parametric distributions to the (zero-truncated) abundance data. In addition to the goodness-of-fit testing, we have also performed qualitative comparisons of BPLN model versus data via smoothed heat-map plots (Anderes et al., 2009). One example of such comparison is provided in Figure 1 where the smoothed heat-map illustrates both true and model-generated bivariate abundance distributions of wild-type receptor repertoires: naive TCRs from lymph nodes and regulatory TCRs from thymus (i.e., Wt TN1 and Wt TR2).

One of the major advantages of the parametric BPLN model is that in many practical situations of interest it produces, across a number of commonly used dissimilarity indices of the form (3.2), hierarchical clustering models which are consistent in terms of the final clusters composition. That is, unlike in the non-parametric analysis, for many TCR datasets the choice of the dissimilarity measure $\mathcal{D}$ is largely irrelevant when modeling repertoires with BPLN model. This is not surprising since our interpretation of the dissimilarity measure is that it accounts for differences in diversity $(\boldsymbol{\theta}, M)$ between pairs of repertoires. Under the condition (which is often satisfied, see Table 1) that $\alpha_1 \approx \alpha_2$ and $\sigma_1 \approx \sigma_2$ one may show that in the BPLN model the dissimilarity is a monotone function of the parameter $\rho$ and hence measures the correlation between repertoires. This is indeed the case for all indices $\mathcal{D}$ discussed in Section 3 above.

The results of hierarchical clustering analysis of the eight mice TCR repertoires under $\mathcal{D}_{MH}$ and $\mathcal{D}_{MI}$ are presented in Figure 2. The figure's left panels (top and bottom) show the dendrograms obtained by agglomerative hierarchical clustering with a complete link function using $\mathcal{D}_{MH}$ (top) and $\mathcal{D}_{MI}$ (bottom) as the dissimilarity measures. Both dendrograms indicate a very good relative agreement of the cluster hierarchical structure and the correct final classification of the eight repertoires in terms of the experimental condition (restricted vs wild-type) as well as the repertoire type (naive or regulatory) and tissue type (thymus vs lymph nodes). Almost identical dendrograms (not shown) were also produced by applying the remaining two measures discussed, namely $\mathcal{D}_{\rho}$ and $\mathcal{D}_{OV}$. The central panels in Figure 2 illustrate the bootstrap approximations to the distribution of the Frobenius distance (see, e.g., Golub and Van Loan 1996 for more on matrix distances) of the dissimilarity matrix $[\mathcal{D}_{MH}(i,j)]$ (top) and $[\mathcal{D}_{MI}(i,j)]$ (bottom) $(1 \leq i,j \leq 8)$, with 95% confidence bound marked with vertical lines. The right panels show the dendrograms corresponding to the dissimilarity matrices at the upper bound of the corresponding 95% confidence intervals (i.e., right vertical lines of the central panels). The fact that the left and right dendrograms in top and bottom panels have identical relative hierarchies indicates strong robustness of the hierarchical clustering against the fluctuations of both $\mathcal{D}_{MH}$ and $\mathcal{D}_{MI}$. The similar robustness was also true for $\mathcal{D}_{\rho}$, and $\mathcal{D}_{OV}$. This agreement between the dissimilarity matrices entries generated under BPLN model using four different dissimilarity measures $\mathcal{D}_{MH}$, $\mathcal{D}_{\rho}$, $\mathcal{D}_{MI}$ and $\mathcal{D}_{OV}$ is further illustrated in Figure 3 where the pairwise loess regressions (Cleveland and Devlin, 1988) of the entries of the dissimilarity measures on each other are presented indicating their monotone relationship (almost a linear one between the first three measures). As already indicated, such agreement should be expected if the data closely follows the BPLN model.

As our final analysis for BPLN model, we have also computed the species richness estimates (3.7) and (3.8) for each pair of repertoires (i.e, each in the total of 28 comparisons) and averaged the result to obtain a pooled estimator of the ratio $D/M$ which was found to be .09 with the 95% confidence interval of (.06, .11). These values suggest a much more severe under-sampling of the TCR populations then the traditional non-parametric estimator of Good (1953) given by $f_1/D \approx .25$ (where now $f_1$ and $D$ are computed from pooled repertoire data).

### 4.3 Analysis under non-parametric model

In order to further examine the results of our parametric hierarchical clustering analysis of the TCR repertoires, we have also performed the more traditional, non-parametric hierarchical clustering analysis of the repertoires in which we have estimated the values of the two dissimilarity measures $\mathcal{D}_{MH}$ and $\mathcal{D}_{MI}$ with the non-parametric estimates based directly on the sample data. Note that $\mathcal{D}_{MH}$ is particularly convenient to analyze non-parametricaly as it only requires the relative estimates of the mixed and marginal moments of order two, which may be calculated directly from the observed (zero-truncated) joined abundance. For that reason, the parametric and non-parametric measures $\mathcal{D}_{MH}$ may be directly compared with each other. On the other hand, the information-based measure $\mathcal{D}_{MI}$ may be estimated by means of a recently popularized Chao-Shen estimator (Chao and Shen, 2003) which, in a manner similar to the ACE and Horwitz-Thomson estimates (see, e.g., Chao, 2006), attempts to adjust explicitly for the fact of only observing the truncated joined distribution.

The result of the direct comparison between all the pairwise estimated $\mathcal{D}_{MH}$ values under the BPLN and the non-parametric models is presented as a scatter plot with fitted loess trend-line in Figure 4. The plot follows a linear trend indicative of a very close agreement between non-parametric and parametric $\mathcal{D}_{MH}$ values for our TCR data. This apparent almost linear relationship between the dissimilarities estimated under the two measures is also confirmed by the dendrogram computed under non-parametric

$\mathcal{D}_{MH}$ (top panel in Figure 5) which gives a stable set of hierarchical clusters almost identical to those obtained under BPLN model (top panel in Figure 2).

In contrast to the parametric case, we found that the non-parametric analogue of the mutual information (MI) dissimilarity (3.6), based on the coverage adjusted Chao-Shen entropy estimator (Vu et al., 2007), did not agree with the non-parametric Morisita-Horn (MH) dissimilarity and consequently yielded a very different and biologically uninterpretable TCR clustering which lacked separation between the wild-type and Ep mice. These differences between non-parametric MH and MI measures may be clearly seen in the two top panels of Figure 5 where also the lack of stability of the MI-dissimilarity is clearly manifested by the large difference between dendrograms within the 95% confidence bound induced by the Frobenius norm of the MI-dissimilarity matrix. Note that this is not the case for the MH dissimilarity, which appears quite stable.

The lowest third panel of Figure 5 illustrates a different non-parametric dissimilarity analysis we have performed, based on the coverage adjusted estimated values of the Shannon entropy function for the eight repertoires (see e.g., Vu et al. 2007, and formula (A.1) in the appendix). The pairwise dissimilarities were computed as absolute differences between the estimated entropy values. Such "linear" comparisons of the diversity measures across repertoires are often appropriate when the repertoires are assumed to have similar abundance patterns (see e.g., Sepúlveda et al. 2009). However, as we may see from the plots, for our datasets the entropy-based clustering turned out to be only partially satisfactory, as the entropy measure was only able to clearly separate two repertoires derived from the wild-type regulatory cells but not the remaining ones. Overall, it appears that for our dataset the entropy based clustering both via MI and the Shannon entropy performed poorly whereas the MH clustering was satisfactory and comparable (in fact almost identical) with our parametric results. These large differences seems to be at least partially due to the fact that, unlike the MH dissimilarity estimate, the non-parametric entropy estimates are adjusting for the zero-truncation of the observed distribution and these adjustments in our particular dataset seem to fluctuate widely from repertoire to repertoire.

# 5 Summary and Discussion

We have presented a simple bivariate-Poisson-lognormal parametric model for fitting and analyzing TCR repertoires data which may be regarded as a natural multivariate extension of a Poisson abundance model with lognormal mixing distribution which has been applied in TCR modeling perviously. As seen from our example of analyzing TCR naive and regulatory repertoires in the wild-type and TCR-restricted Ep mice, the model provides for a robust and consistent fit to the TCR data allowing for a very detailed yet relatively simple comparison of multiple repertoires, for instance by means of dissimilarity analysis and hierarchical clustering. We show that in our TCR dataset under the parametric model the four popular dissimilarity indices: (i) Morista-Horn and (ii) correlation indices, (iii) mutual information and (iv) the overlap dissimilarities are monotone functions of each other, leading therefore to the same, biologically meaningful, clustering of the repertoires. In contrast, when applying to the same data the non-parametric estimates of dissimilarities, the clusterings are seen to behave erratically and are highly dependent on the particular choices of measures with some of them (e.g.,the mutual information) yielding biologically implausible clustering.

Our parametric analysis suggests that in a typical experiment based on harvesting sequences from single-cell TCRs, the overall under-sampling of the TCR population is much higher than in the macroscopic biodiversity studies, for which many of the statistical tools of species abundance comparison were originally developed. This and the apparent lack of agreement between the non-parametric dissimilarity

measures when applied to our, relatively simple, dataset, seem to indicate that many commonly used non-parametric biodiversity statistics may perform poorly when applied to severely under-sampled TCR repertoires. The advantage of the parametric approach and in particular the model proposed here is that even with very severe data under-sampling it allows for the proper adjustment for the missing abundance information and estimation of the full set of repertoires features. The statistical package *poilog* available from CRAN archive (`http://cran.r-project.org`) makes the fitting of our model particularly convenient by providing numerical algorithms for the parameters estimation via the maximum likelihood. Further studies and a larger number of TCR datasets with more sequences are needed in order to more comprehensively evaluate the BPLN model and test its ability to discriminate between TCR repertoires in a biologically meaningful way.

# A Appendix: Mutual Information Bounds

The fact that the bounds (3.1) hold for the dissimilarity index (3.6) follows from the general properties of the Shannon entropy function which is defined (see e.g., Koski 2001) for any discrete random vector $X$ with probability distribution $p(x)$ as

$$H(X) = -\sum_x p(x)\log p(x), \tag{A.1}$$

with the summation is taken over $x$ values for which $p(x) > 0$. Extending the definition of the index (3.6) to any pair of discrete real random variables $X, Y$ with joined distribution $p(x, y)$ and marginals $p(x), p(y)$, we define their *mutual information* as

$$\begin{aligned} MI(X,Y) &= \sum_{x,y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \\ &= -\sum_x p(x)\log p(x) - \sum_y p(y)\log p(y) + \sum_{x,y} p(x,y)\log p(x,y) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{A.2}$$

Due to the elementary inequality $\log(x) \le x - 1$ valid for any $x > 0$ we have that

$$-MI(X,Y) = \sum_{x,y} p(x,y)\log\left(\frac{p(x)p(y)}{p(x,y)}\right) \le \sum_{x,y} p(x,y)\left(\frac{p(x)p(y)}{p(x,y)} - 1\right) = -\sum_{x,y} p(x)p(y) + 1 = 0$$

and therefore

$$MI(X,Y) \ge 0 \tag{A.3}$$

Note that $MI(X,X) = H(X)$ and therefore (due to symmetry and (A.2)) to argue upper bound in (3.1) it suffices to show that

$$H(X,Y) \ge H(X). \tag{A.4}$$

This follows easily, since

$$H(X,Y) - H(X) = \sum_{x,y} p(x,y)\log p(x,y) - \sum_x p(x)\log p(x) = -\sum_{x,y} p(x,y)\log\left(\frac{p(x,y)}{p(x)}\right) \ge 0.$$

The bounds (3.1) for $MI(X,Y)$ follow now from (A.3)–(A.4) and (A.2). The measure and (3.6) is, of course a special case of $MI(X,Y)$.

# References


Aitchison, J. and Ho, C. (1989). The multivariate poisson-log normal distribution. *Biometrika*, 76(4):643.

Anderes, E., Stein, M., Minin, V., O'Brien, J., Seregin, A., Marron, B., Tablar, A., Fujita, A., Sato, J., Kojima, K., et al. (2009). Local likelihood estimation of local parameters for nonstationary random fields. *Arxiv preprint arXiv:0911.0047*.

Arstila, T. P., Casrouge, A., Baron, V., Even, J., Kanellopoulos, J., and Kourilsky, P. (1999). A direct estimate of the human alphabeta t cell receptor diversity. *Science*, 286(5441):958–61.

Barger, K. and Bunge, J. (2008). Bayesian estimation of the number of species using noninformative priors. *Biometrical Journal*, 50(6):1064–1076.

Barth, R. K., Kim, B. S., Lan, N. C., Hunkapiller, T., Sobieck, N., Winoto, A., Gershenfeld, H., Okada, C., Hansburg, D., and Weissman, I. L. (1985). The murine t-cell receptor uses a limited repertoire of expressed v beta gene segments. *Nature*, 316(6028):517–23.

Behlke, M. A., Spinella, D. G., Chou, H. S., Sha, W., Hartl, D. L., and Loh, D. Y. (1985). T-cell receptor beta-chain expression: dependence on relatively few variable region genes. *Science*, 229(4713):566–70.

Bulmer, M. (1974). On fitting the poisson lognormal distribution to species abundance data. *Biometrics*, 30:101–110.

Busch, D. H. and Pamer, E. G. (1999). T cell affinity maturation by selective expansion during infection. *J Exp Med*, 189(4):701–10.

Butz, E. A. and Bevan, M. J. (1998). Massive expansion of antigen-specific cd8+ t cells during an acute virus infection. *Immunity*, 8(2):167–75.

Casrouge, A., Beaudoing, E., Dalle, S., Pannetier, C., Kanellopoulos, J., and Kourilsky, P. (2000). Size estimate of the alpha beta tcr repertoire of naive mouse splenocytes. *J Immunol*, 164(11):5782–7.

Chao, A. (2006). Species richness estimation. In Balakrishnan, N., Read, C., and Vidakovic, B., editors, *Encyclopedia of Statistical Sciences*. Wiley, New York.

Chao, A. and Shen, T. (2003). Nonparametric estimation of shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics*, 10(4):429–443.

Chen, W., Jin, W., Hardegen, N., Lei, K.-J., Li, L., Marinos, N., McGrady, G., and Wahl, S. M. (2003). Conversion of peripheral cd4+cd25- naive t cells to cd4+cd25+ regulatory t cells by tgf-beta induction of transcription factor foxp3. *J Exp Med*, 198(12):1875–86.

Cleveland, W. and Devlin, S. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610.

Davis, M. M. and Bjorkman, P. J. (1988). T-cell antigen receptor genes and T-cell recognition. *Nature*, 334(6181):395–402.

Efron, B. and Tibshirani, R. (1997). *An introduction to the bootstrap*. Chapman & Hall.

Engen, S., Lande, R., Walla, T., and DeVries, P. (2002). Analyzing spatial structure of communities using the two-dimensional poisson lognormal species abundance model. *The American Naturalist*, 160(1):60–73.

Ferreira, C., Singh, Y., Furmanski, A. L., Wong, F. S., Garden, O. A., and Dyson, J. (2009). Non-obese diabetic mice select a low-diversity repertoire of natural regulatory t cells. *Proc Natl Acad Sci U S A*, 106(20):8320–5.

Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, 12:42–58.

Freeman, J. D., Warren, R. L., Webb, J. R., Nelson, B. H., and Holt, R. A. (2009). Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res.*

Gill, R., Rempala, G., and Czajkowski, M. (2009). Confidence estimation via the parametric bootstrap in logistic joinpoint regression. *Journal of Statistical Planning and Inference.*

Golub, G. and Van Loan, C. (1996). *Matrix computations.* Johns Hopkins Univ Pr.

Good, I. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3):237–264.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *Elements of Statistical Learning: Data Mining,Inference, and Prediction.* Springer-Verlag, New York.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Hsieh, C.-S., Liang, Y., Tyznik, A. J., Self, S. G., Liggitt, D., and Rudensky, A. Y. (2004). Recognition of the peripheral self by naturally arising cd25+ cd4+ t cell receptors. *Immunity*, 21(2):267 – 277.

Hsieh, C.-S., Zheng, Y., Liang, Y., Fontenot, J. D., and Rudensky, A. Y. (2006). An intersection between the self-reactive regulatory and nonregulatory t cell receptor repertoires. *Nat Immunol*, 7(4):401–10.

Janeway, C. (2005). *Immunobiology: The Immune System in Health And Disease.* Garland Science, New York, 6th edition.

Karlis, D. (2003). An em algorithm for multivariate poisson distribution and related models. *Journal of Applied Statistics*, 30(1):63–77.

Kepler, T. B., He, M., Tomfohr, J. K., Devlin, B. H., Sarzotti, M., and Markert, M. L. (2005). Statistical analysis of antigen receptor spectratype data. *Bioinformatics*, 21(16):3394–400.

Komatsu, N., Mariotti-Ferrandiz, M. E., Wang, Y., Malissen, B., Waldmann, H., and Hori, S. (2009). Heterogeneity of natural foxp3+ t cells: a committed regulatory T-cell lineage and an uncommitted minor population retaining plasticity. *Proc Natl Acad Sci U S A*, 106(6):1903–8.

Koski, T. (2001). *Hidden Markov models for bioinformatics.* Springer.

Kuczma, M., Pawlikowska, I., Kopij, M., Podolsky, R., Rempala, G. A., and Kraj, P. (2009). Tcr repertoire and foxp3 expression define functionally distinct subsets of cd4+ regulatory t cells. *J Immunol*, 183(5):3118–29.

Legendre, P. and Legendre, L. (1998). *Numerical ecology*, volume 20 of *Developments in Environmental Modelling*. Elsevier Science B.V., Amsterdam, english edition. Translated and revised from the second French (1984) edition.

Lewins, W. and Joanes, D. (1984). Bayesian-estimation of the number of species. *Biometrics*, 40(2):323–328.

Luczynski, W., Stasiak-Barmuta, A., Piszcz, J., Ilendo, E., Kowalczuk, O., and Krawczuk-Rybak, M. (2007). B-cell chronic lymphocytic leukemia-derived dendritic cells stimulate allogeneic t-cell response and express chemokines involved in t-cell migration. *Neoplasma*, 54(6):527–535.

Magurran, A. E. (2005). Biological diversity. *Current Biology*, 15(4):R116–8.

McHeyzer-Williams, L. J., Panus, J. F., Mikszta, J. A., and McHeyzer-Williams, M. G. (1999). Evolution of antigen-specific t cell receptors in vivo: preimmune and antigen-driven selection of preferred complementarity-determining region 3 (cdr3) motifs. *J Exp Med*, 189(11):1823–38.

Naumov, Y. N., Naumova, E. N., Hogan, K. T., Selin, L. K., and Gorski, J. (2003). A fractal clonotype distribution in the cd8+ memory t cell repertoire could optimize potential for immune responses. *J Immunol*, 170(8):3994–4001.

Nayak, T. (1991). Estimating the number of component processes of a superimposed process. *Biometrika*, 78(1):75–81.

Naylor, K., Li, G., Vallejo, A. N., Lee, W.-W., Koetz, K., Bryl, E., Witkowski, J., Fulbright, J., Weyand, C. M., and Goronzy, J. J. (2005). The Influence of Age on T Cell Generation and TCR Diversity. *J Immunol*, 174(11):7446–7452.

Ord, J. K. and Whitmore, G. A. (1986). The poisson-inverse gaussian distribution as a model for species abundance. *Commun. Statist.-Theory Methods*, 15:853–871.

Pacholczyk, R., Ignatowicz, H., Kraj, P., and Ignatowicz, L. (2006). Origin and t cell receptor diversity of foxp3+cd4+cd25+ t cells. *Immunity*, 25(2):249–59.

Pacholczyk, R., Kern, J., Singh, N., Iwashima, M., Kraj, P., and Ignatowicz, L. (2007). Nonself-antigens are the cognate specificities of foxp3+ regulatory t cells. *Immunity*, 27(3):493–504.

Pewe, L. L., Netland, J. M., Heard, S. B., and Perlman, S. (2004). Very diverse cd8 t cell clonotypic responses after virus infections. *J Immunol*, 172(5):3151–6.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rempala, G. A. and Szatzschneider, K. (2004). Bootstrapping parametric models of mortality. *Scand. Actuar. J.*, (1):53–78.

Rodrigues, J., Milan, L., and Leite, J. (2001). Hierarchical bayesian estimation for the number of species. *Biometrical Journal*, 43(6):737–746.

Sallusto, F., Geginat, J., and Lanzavecchia, A. (2004). Central memory and effector memory t cell subsets: function, generation, and maintenance. *Annu Rev Immunol*, 22:745–63.

Sallusto, F., Lenig, D., Förster, R., Lipp, M., and Lanzavecchia, A. (1999). Two subsets of memory t lymphocytes with distinct homing potentials and effector functions. *Nature*, 401(6754):708–12.

Savage, P. A., Boniface, J. J., and Davis, M. M. (1999). A kinetic basis for t cell receptor repertoire selection during an immune response. *Immunity*, 10(4):485–92.

Sebzda, E., Mariathasan, S., Ohteki, T., Jones, R., Bachmann, M. F., and Ohashi, P. S. (1999). Selection of the t cell repertoire. *Annu Rev Immunol*, 17:829–74.

Sepúlveda, N., Paulino, C. D., and Carneiro, J. (2009). Estimation of t-cell repertoire diversity and clonal size distribution by poisson abundance models. *J Immunol Methods*.

Sheldon, A. (1969). Equitability indices: dependence on the species count. *Ecology*, 50(3):466–467.

Sichel, H. (1997). Modelling species-abundance frequencies and species-individual functions with the generalized inverse gaussian-poisson distribution. *South African Statistical Journal*, 31(1):13–37.

Smith, W., Solow, A., and Preston, P. (1996). An estimator of species overlap using a modified beta-binomial model. *Biometrics*, 52(4):1472–1477.

Solow, A. (1994). On the bayesian-estimation of the number of species in a community. *Ecology*, 75(7):2139–2142.

Venturi, V., Kedzierska, K., Tanaka, M. M., Turner, S. J., Doherty, P. C., and Davenport, M. P. (2008). Method for assessing the similarity between subsets of the t cell receptor repertoire. *J Immunol Methods*, 329(1-2):67–80.

Vu, V. Q., Yu, B., and Kass, R. E. (2007). Coverage-adjusted entropy estimation. *Statistics In Medicine*, 26(21):4039–4060.

Warren, R. L., Nelson, B. H., and Holt, R. A. (2009). Profiling model T-cell metagenomes with short reads. *Sequence analysis Profiling model T-cell metagenomes with short reads*, 25(4).

Wong, J., Mathis, D., and Benoist, C. (2007). Tcr-based lineage tracing: no evidence for conversion of conventional into regulatory t cells in response to a natural self-antigen in pancreatic islets. *J Exp Med*, 204(9):2039–2045.

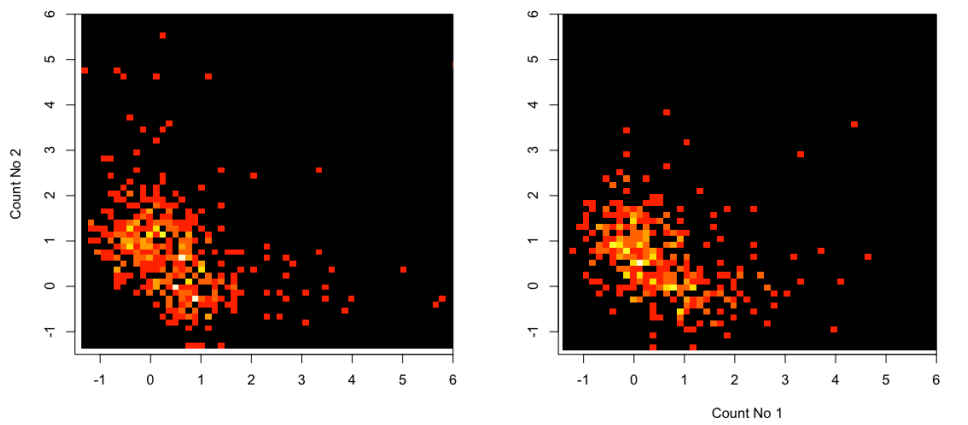


Figure 1: **Model generated vs observed pairs of frequency data.** *Left panel:* kernel-density-smoothed heatmap of joined frequency data of lymph nodes naive ($x$-axis) and thymus regulatory ($y$-axis) TCR repertoires in wild mice (Wt.TN1 and Wt.TR2 ). *Right panel:* kernel-density-smoothed heatmap for the same-size sample simulated from the distribution of BPLN random variable fitted to the data. Increased brightness indicates higher frequency.

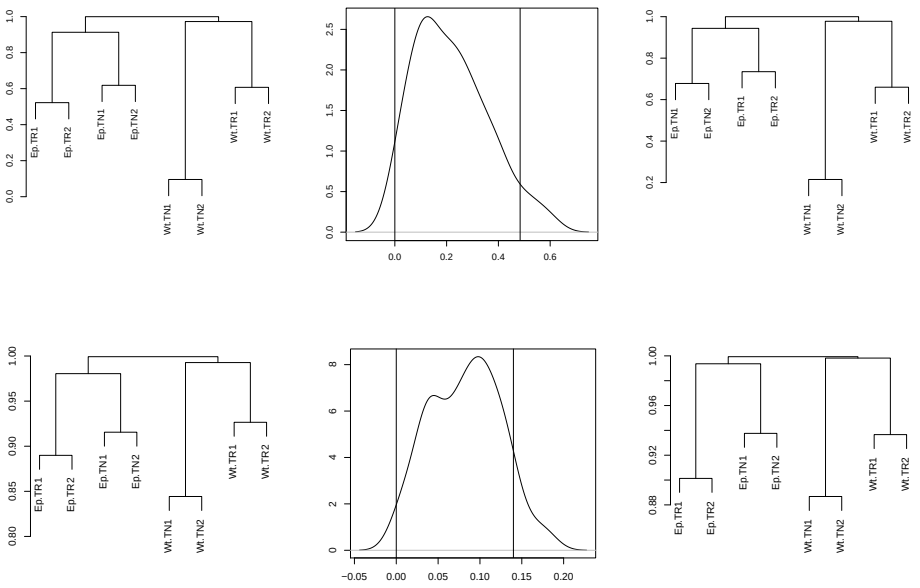


Figure 2: **Repertoire dendrograms and their confidence bounds obtained under BPLN model.** Dendrograms for hierarchical clustering and their corresponding confidence intervals obtained using agglomerative clustering and a complete link for eight repertoires of naive and regulatory TCRs derived from (1) the lymph nodes and (2) thymus in wild type and TCR mini mice. *Top panel:* (left) clustering using Morisita-Horn dissimilarity measure $\mathcal{D}_{MH}$ given by (3.3); (middle) bootstrap estimate of the 95% confidence interval of the Phrobenius norm of the $\mathcal{D}_{MH}$ dissimilarity matrix; (right) dendrogram corresponding to the upper bound of the 95% CI $\mathcal{D}_{MH}$ dissimilarity matrix. *Bottom panel:* hierarchical clustering according to the parametric mutual information dissimilarity measure $\mathcal{D}_{MI}$ given by (3.6).

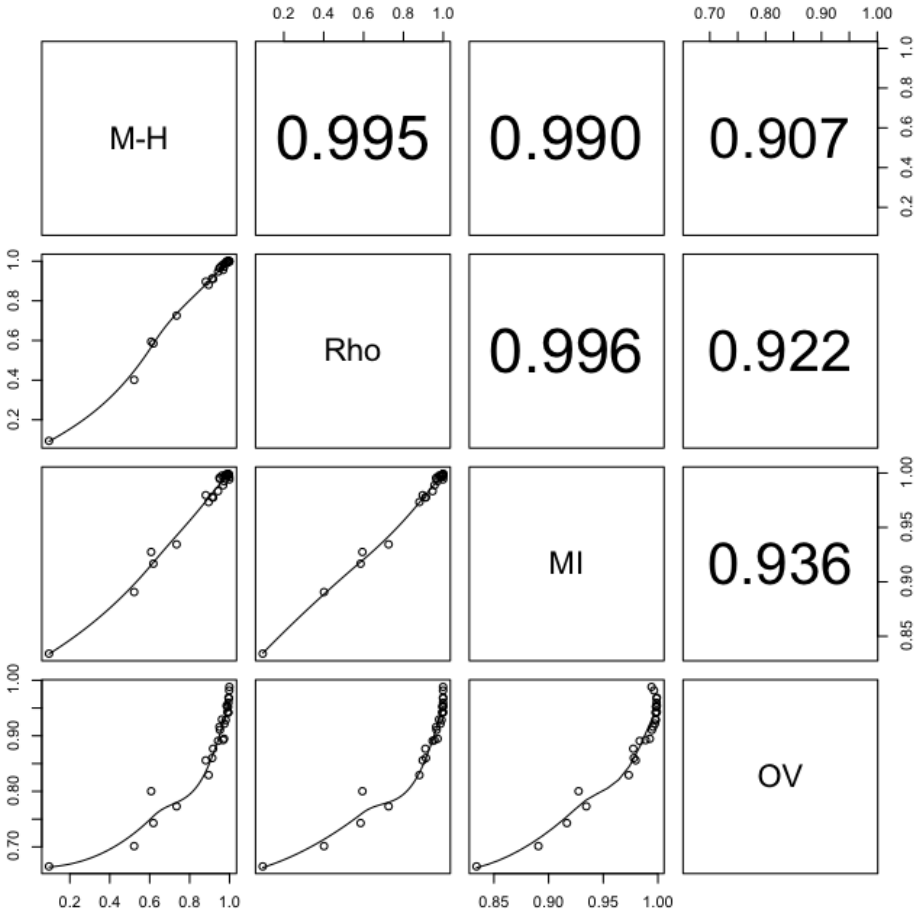


Figure 3: **Pairwise comparison of dissimilarities under BPLN model.** *Lower triangular panels:* pairwise dissimilarities plots obtained under BPLN models fitted to mice data for the four different $\mathcal{D}$ measures discussed in Section 3.1. Local (loess) regression curves were added to the plots for better readability. *Upper triangular panels:* $R^2$ values for the corresponding linear regressions.

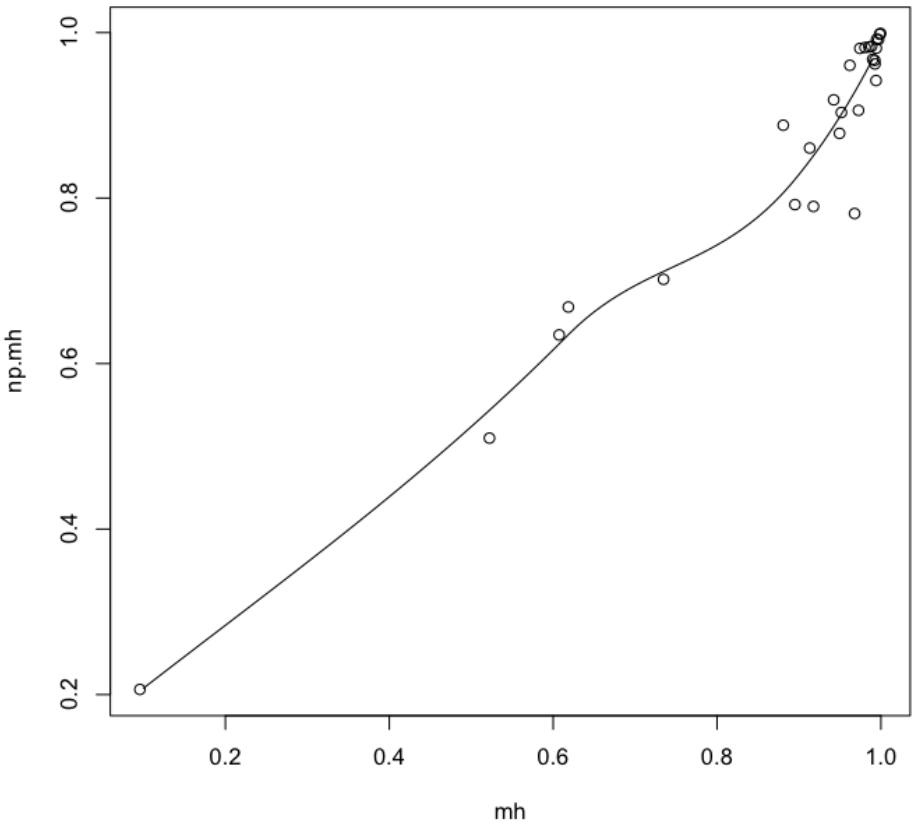


Figure 4: **Morisita-Horn ($\mathcal{D}_{MH}$) dissimilarities under parametric and non-parametric models.** Scatter plot along with the local (loess) regression curve for pairwise dissimilarities between eight repertoires computed under parametric (x-axis) and non-parametric (y-axis) models using Morisita-Horn index as given by (3.3). In the non-parametric case the joined probabilities $p_\theta(k, l)$ are estimated by the corresponding joined frequencies.

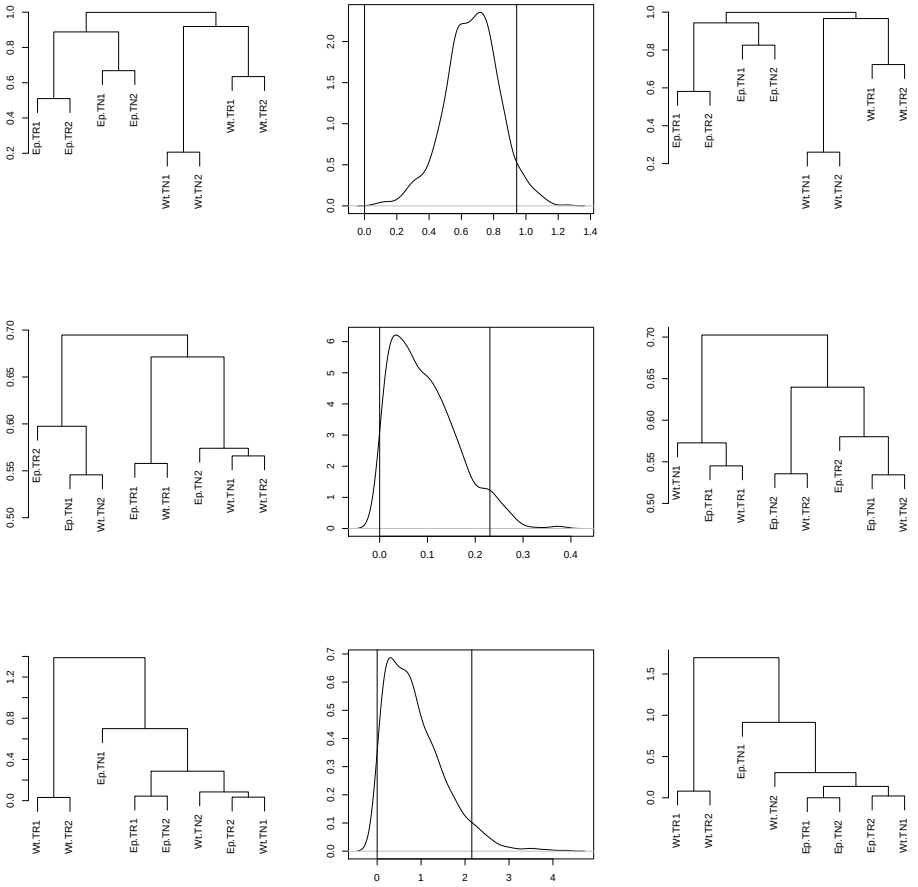


Figure 5: **Repertoire dendrograms and their confidence bounds under non-parametric measures of dissimilarity.** Dendrograms under various dissimilarity measures obtained from the non-parametric analogue of the model (3.11) using agglomerative clustering and the complete link. In all three panel rows the most left dendrograms were obtained using the point estimates of the dissimilarity matrix calculated from the data, whereas the most right ones were obtained using upper 95% confidence bound on the Frobenius norm distribution of the dissimilarity matrix. The corresponding bootstrap estimate of the entire norm distribution is provided in the middle plot as a density estimator, with 95% bounds marked with vertical lines. *Top panels*: hierarchical clusters based on the nonparametric version of the Morisita-Horn dissimilarity index (3.3). *Middle panels:* clusters based on the non-parametric version of the mutual information dissimilarity (3.6). *Bottom panels:* clusters based on the values of the Shannon entropy function with no direct pairwise comparisons.