Journal of the Royal Statistical Society
SERIES C Applied Statistics

Appl. Statist. (2014)
63, *Part* 3, *pp.* 385–404

A Bayesian semiparametric approach for the differential analysis of sequence counts data

Michele Guindani,

University of Texas M. D. Anderson Cancer Center, Houston, USA

Nuno Sepúlveda,

London School of Hygiene and Tropical Medicine, UK, and University of Lisbon, Portugal

Carlos Daniel Paulino

Instituto Superior Técnico, Lisbon, and University of Lisbon, Portugal

and Peter Müller

University of Texas at Austin, USA

[Received December 2011. Final revision June 2013]

Summary. Data obtained by using modern sequencing technologies are often summarized by recording the frequencies of observed sequences. Examples include the analysis of T-cell counts in immunological research and studies of gene expression based on counts of RNA fragments. In both cases the items being counted are sequences, of proteins and base pairs respectively. The resulting sequence abundance distribution is usually characterized by overdispersion. We propose a Bayesian semiparametric approach to implement inference for such data. Besides modelling the overdispersion, the approach takes also into account two related sources of bias that are usually associated with sequence counts data: some sequence types may not be recorded during the experiment and the total count may differ from one experiment to another. We illustrate our methodology with two data sets: one regarding the analysis of CD4+ T-cell counts in healthy and diabetic mice and another data set concerning the comparison of messenger RNA fragments recorded in a serial analysis of gene expression experiment with gastrointestinal tissue of healthy and cancer patients.

Keywords: Bayesian non-parametrics; Differential abundance; Sequence counts data; T-cell repertoire

1. Introduction

Many problems in biomedical research require inference for frequencies of some biological entity, such as gene transcripts or molecular receptors. For example, in immunology, investigators may want to compare the number of distinct T-cell receptors (TCRs) and their respective abundances in autoimmune patients and healthy individuals (Hsieh *et al.*, 2006; Ferreira *et al.*, 2009). TCRs are molecules at the surface of T-cells (a white blood cell) that bind to antigens and trigger the immune response itself. Diversity of TCRs, i.e. the number of distinct molecules,

Address for correspondence: Michele Guindani, Department of Biostatistics, University of Texas M. D. Anderson Cancer Center, Unit 1411, PO Box 301402, Houston, TX 77025, USA.
E-mail: mguindani@mdanderson.org

© 2013 Royal Statistical Society 0035–9254/14/63385

is an important characteristic of the immune system. In molecular biology, the comparison of different tissue samples may be based on a comparison of counts of different types of messenger ribonucleic acid (RNA), which is an intermediate product in protein synthesis. The abundance of different types of messenger RNA characterizes gene expression. These two examples serve as the motivating applications in this discussion. In both applications, the data are sequence counts, of proteins and of base pairs respectively. We shall therefore generically refer to the biological entities that are being counted as sequences. Table 1 shows the first few lines of a typical data set. Section (a) shows the raw data of counts for each unique sequence. The number of rows is the number of unique sequences. Section (b) shows the summary as frequencies of counts (the clonal size distribution or sequence abundance distribution). The corresponding number of rows is the number of different counts in section (a).

By the nature of the underlying biological processes or by the way that the data are collected or processed, the distribution of the observed data in such studies shows several general characteristics. First, samples may be characterized by a large number of sequences recorded at low frequencies together with just a few sequences recorded very frequently. This leads to overdispersion relative to a Poisson model where variance is tied with the mean (see examples in Bentley *et al.* (2008) and Yoon *et al.* (2009)). In addition, the sequencing experiment may fail to record many rare sequence types of the original population. Therefore, a single sample may not capture all distinct sequences that are present in a population. Zero counts are not included in the count data, i.e. zero counts are censored. As a consequence, the observed frequencies are biased estimates of true abundances. Some correction is needed when analysing the data (Morris *et al.*, 2003; Sepúlveda *et al.*, 2010).

Several modelling approaches have been proposed to analyse sequence abundance distributions under different inferential assumptions. Common modelling strategies include the zero-inflated Poisson regression model (Nie *et al.*, 2006; Dhavala *et al.*, 2010) and the negative binomial distribution where the variance is modelled explicitly as a function of the mean and an additional dispersion parameter (see Robinson and Smyth (2007), Hardcastle and Kelly (2010) and Anders and Huber (2010)). Indeed, the negative binomial distribution can be characterized as the marginal distribution in a Poisson–gamma mixture model (see, for example, Cameron and Trivedi (1998)). The method by Robinson and Smyth (2007) is implemented in the widely used EdgeR package and relies on the estimation of a common overdispersion parameter across samples. Other proposals include the Poisson–log-normal distribution (Sepúlveda *et al.*, 2010; Rempala *et al.*, 2011) or the truncated Poisson–gamma distribution (Thygesen and Zwinderman, 2006). Alternatively, finite mixtures of Poisson distributions have been proposed as a way to

Table 1. Typical data of sequence abundances as counts of unique amino acid sequences (section (a)) and summarized as frequencies of counts (section (b))

Unique sequence	*Count*	*Clonal size (count)*	*Frequency*
(a) Sequence counts		*(b) Clonal size distribution*	
CAARGGLSGKLTF	40	1	22
CAAPRGGLSGKLTF	39	...	...
CAARTGGLSGKLTF	39	39	2
...	...	40	1
CAARGADDNYQLIW	1		
CAARGAKDNYQLIW	1		

provide a better description of this type of data, often with an assumed known number of mixing components (Zuyderduyn, 2007).

We build on these approaches and propose Bayesian inference in a semiparametric mixture of Poisson distributions model to estimate the distribution of the observed counts in the presence of overdispersion and uncertainty on the true number of unique sequences in a population (e.g. tissue or cell type). More specifically, we assume a Dirichlet process prior on the mean of the Poisson components. The Dirichlet process prior has been extensively used as an automatic and adaptive method for density estimation (see, for example, Ferguson (1983), Escobar and West (1995) and Gasparini (1996)). In addition, we explicitly model the experimentally induced censoring of zero counts. Inference includes estimating the underlying population sequence diversity, i.e. the number of different unique sequences that can be found in the population. We show that, for small and moderate size data sets with overdispersed data, such a correction is non-negligible. Reliable estimates of sequence abundances are particularly important if one aims to compare the sequence abundance distribution under different biological or experimental conditions. An extension of the semiparametric Poisson mixture model to carry out such comparisons is another contribution of this paper. The inference proposed is valid for both small and large data sets and allows for different degrees of overdispersion across samples.

Recent related non-parametric Bayesian literature includes work by Trippa and Parmigiani (2011) who used semiparametric Poisson mixture models as a tool to generate realistic simulation scenarios for the evaluation of false discovery procedures, and a sequence of papers by Lijoi *et al.* (2007a, b, 2008) and Favaro *et al.* (2009, 2012), where they used non-parametric Bayesian priors for a sampling model on species diversity in experiments similar to the serial analysis of gene expression (SAGE) experiments that are discussed in this paper. They focused on inference for the expected number of species under a given additional sampling effort. Similarly to the use of semiparametric Poisson mixtures to model overdispersion, Canale and Dunson (2012) developed models with underdispersion, i.e. distributions where the variance is less than the mean, by using kernel mixture models with kernels induced through rounding of continuous kernels.

We illustrate our methodology with two applications: one related to the study of T-cell receptors in healthy and diabetic mice (Ferreira *et al.*, 2009) and another one measuring gene expression using SAGE technology. The analysis of the T-cell receptor data has two objectives: to quantify the receptor diversity and to compare the frequencies of the receptors across different T-cell populations. In the analysis of the SAGE data, the main objective is to compare the frequency of messenger RNA transcripts across different libraries of healthy and tumour colon tissues.

The structure of the paper is as follows. Section 2 introduces the model for the single-sample analysis. The section includes a simulation study to compare inference with similar parametric models as well as inference for the two motivating applications. Section 3 extends the model to the multivariate case to compare several sequence abundance distributions. Sections 3.3 and 3.4 discuss related inference in the two motivating applications. Finally, Section 4 finishes with some concluding remarks. Supporting information on the journal's Web site contains implementation details for the posterior simulation. The computer code that was used to analyse the data can be obtained from the authors on request.

2. Modelling relative sequence abundances

2.1. A semiparametric model

2.1.1. Sampling model

We start by considering a model for inference with a single sample or library of sequence counts.

Let y_i, $i=1,\ldots,k$, denote the counts of the k distinct sequences in a sample. However, counts $y_i=0$ are not observed, leaving only $k' \leqslant k$ observed counts $y_i>0$. We refer to k and k' as the population and the sample diversity respectively. The balance $k_0=k-k'$ is a measure of the sampling error, i.e. the undercounting of sequences with low copy numbers, due to censoring of zero counts.

We first summarize the proposed model in words. Let $G(y_i)$ denote the unknown distribution of counts y_i. We estimate $G(\cdot)$ on the basis of observed data y_i, using a Bayesian non-parametric approach, i.e. with minimal assumptions on the data-generating mechanism. In turn, this allows us to make inference about the sample diversity k. Let $\hat{G}(y)$ denote the empirical frequency of counts $y>0$. Fig. 1 shows $\hat{G}(y)$ as a pin plot. The frequency of zero counts can be estimated by extrapolating the observed trend of $\ldots,\hat{G}(3),\hat{G}(2),\hat{G}(1)$ to $y=0$. This could be done by inferring *ad hoc* the behaviour of the $\hat{G}(\cdot)$ curve at $y=0$ and adequately renormalizing the weights to ensure that the total mass is 1. The model-based approach proposed makes that 'eyeballing' more formal and attempts to describe the related uncertainties. In Fig. 1 the full curve shows the estimated $E(G|y)$ under the model proposed. The many thin grey dotted curves illustrate uncertainty as draws from the posterior distribution of G, i.e. $G \sim p(G|y)$. Information on $G(\cdot)$ does not yet allow inference about true abundances of the sequences, for lack of any parameter that could be readily interpreted as sequence abundance. We achieve such inference with a simple trick. We represent $G(\cdot)$ as a mixture of Poisson distributions. The mixture naturally introduces a latent variable λ_i that can be interpreted as true abundance of the ith sequence.

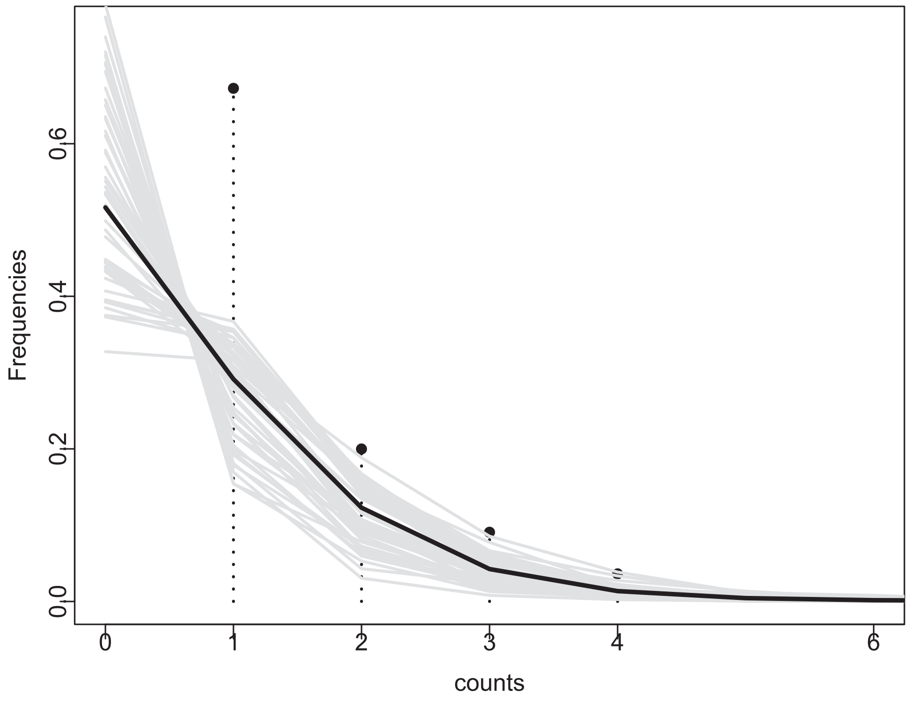


Fig. 1. Rationale for a model-based approach to inference on zero counts for sequence counts data: ⋮, data y_i listed in Table 2, healthy mouse Tconv 2, summarized as empirical frequencies $\hat{G}(y)$, $y=1,2,\ldots$ (*ad hoc* inference for $G(0)$ could be based on extrapolating the trend of $\hat{G}(y)$, $y \geqslant 1$, to $y=0$; the approach proposed formalizes this extrapolation as model-based inference); ——, posterior expectation $E(G\,|\,\text{data})$ under the proposed semiparametric mixture of Poisson distributions model; ——, draws from the posterior distribution $G \sim p(G\,|\,y)$

We start the formal model construction with a binomial distribution for the sample diversity k':

$$p(k'|k,G)=\binom{k}{k'}G(0)^{k-k'}G(+)^{k'},$$

where $G(0)=\Pr(y_i=0)$ is the probability of a zero count, and $G(+)=\Pr(y_i>0)$ denotes the complement probability of a non-zero count.

Let $\mathbf{y}=\{y_i, i=1,\ldots,k'\}$ denote the observed sequence counts. Conditional on the sample diversity k', the likelihood for G and k is $p(\mathbf{y}|k',k,G)=\Pi_{i=1}^{k'}G(y_i)/G(+)$. Thus,

$$p(\mathbf{y},k'|k,G)=\binom{k}{k'}\prod_{i=1}^{k'}G(y_i)\prod_{i=k'+1}^{k}G(0). \tag{1}$$

Note that the sampling model (1) includes positive probability for $k'=0$, although by the nature of the experiment at least some sequences are always counted, i.e. $k'>0$. This is not a problem for posterior inference, since the observed data always include $k'>0$.

2.1.2. Non-parametric prior

We continue the model construction with a non-parametric Bayesian prior for $G(\cdot)$. A parametric model would be too restrictive and inappropriately determine the extrapolation to $G(0)$ by the particular assumed parametric form. For example, in a Poisson sampling model the mean determines the Poisson rate, and thus $\Pr(y_i=0)$. From a data analysis perspective this is undesirable. Instead, we assume that the counts y_i are independently sampled from a mixture of Poisson distributions. Let $\mathrm{Poi}(x;\lambda)$ denote a Poisson distribution for the random variable x with expectation λ. In our setting, the parameter λ is a measure of the true abundance of a sequence in the biological sample. We assume that

$$G(y_i)=\int \mathrm{Poi}(y_i;\lambda_i)\,P(\mathrm{d}\lambda_i), \tag{2}$$

$i=1,\ldots,k$, independently conditional on P. Besides the increased flexibility in modelling $G(\cdot)$, the introduction of the mixture model is important for the desired inference. One of the inference goals is to estimate the abundance of each sequence. Consider an equivalent expression of model (2) as a hierarchical model:

$$y_i|\lambda_i \overset{\text{ind}}{\sim} \mathrm{Poi}(\lambda_i) \qquad \lambda_i \overset{\text{IID}}{\sim} P. \tag{3}$$

The variables λ_i can be interpreted as the true abundance of the ith sequence. Also, since P features in the model as the prior for the latent unobservable variables λ_i, it suffices to assume a discrete probability measure P. In fact, a discrete mixing measure P has the additional advantage of introducing a clustering of the observed sequences into groups of comparable abundance. We discuss details below.

We specify a prior distribution for the mixture model (2) by assuming a non-parametric prior on the mixing measure P. One of the most commonly used non-parametric Bayes priors is the Dirichlet process prior. The Dirichlet process is characterized by two parameters: the prior mean G^* and the mass parameter ν. The mass parameter determines, among other important properties, the variation in the random measure around the prior mean. We write $P\sim\mathrm{DP}(G^*,\nu)$. We refer to Ferguson (1973) and Walker *et al.* (1999) for a definition and important properties of the Dirichlet process model. The main reasons that motivate us to consider the proposed semiparametric mixture of Poisson distributions model are the increased robustness of inference with respect to modelling assumptions, the lack of a good biologically justified parametric model and the interpretation of the latent λ_i as sequence abundance. We note that the non-parametric prior cannot add information without exploiting additional expert judgement. In general, the semiparametric model will only allow for a more honest description of uncertainties in the extrapolation to $G(0)$. The random G is infinite dimensional, in contrast with, for example, a Poisson model that is determined by a single parameter.

2.1.3. Random partition

Among other implications, the Dirichlet process prior on the mixing measure allows for inference about clustering of observations, in the following sense. A key feature is the almost sure discreteness of P; hence, a sample from P has a positive probability of ties. In model (3), let λ_j^*, $j=1,\ldots,L$, denote the $L\leqslant k$ unique values of λ_i. If we use the ties to define clusters, then mixture of Dirichlet process models such as model (2) can be used for detecting clusters of observations (Green and Richardson, 2001; Quintana and Iglesias, 2003). It is often convenient to restate the model in terms of latent cluster indicators s_i, such that $s_i=j$ if and only if observation i belongs to cluster j, i.e. $\lambda_i=\lambda_j^*$. Let L denote the number of clusters and let $n_j=|\{i:s_i=j\}|$ denote the size of the jth cluster. The prior probability of a given clustering structure, say $\mathbf{s}=\{s_1,\ldots,s_k\}$, is

$$p(\mathbf{s}|k)=\frac{\nu^L\,\Gamma(\nu)\prod_{j=1}^{L}\Gamma(n_j)}{\Gamma(\nu+k)}. \qquad (4)$$

Assume that the goal is to group sequences with similar frequency patterns, according to the values of the true abundances λ_i. The posterior distribution $p(\mathbf{s}|\mathbf{y})$ provides a full probabilistic description of such partitions.

Dirichlet process mixture models like model (3) have been extensively studied in the literature. Posterior inference can be implemented by Markov chain Monte Carlo posterior simulation to obtain desired posterior and posterior predictive summaries (Escobar and West, 1995; MacEachern and Müller, 1998; Neal, 2000; Papaspiliopoulos and Roberts, 2008; Dahl, 2003). To achieve faster mixing Markov chains it is possible to integrate out analytically the random probability measure P and the abundances λ_i, leaving a model in $\mathbf{s}$ only.

Model (2) is completed with a base measure G^* and a prior on k. Using $G^*(\lambda)\equiv\text{Ga}(\alpha,\beta)$, i.e. a gamma distribution with mean α/β, defines a conjugate Dirichlet process mixture. This greatly simplifies posterior simulation. Integrating out the λ_is, we are left with a Poisson–gamma random-mixture model. In the analysis of the examples in the following section, we considered fixed hyperparameters α and β, chosen to provide large support to the prior distribution. Furthermore, since $E(G)=G^*$ *a priori*, the implied marginal for one observation, $p(y_i)=E\{G(y_i)\}$, is a negative binomial. However, the data will inform posterior inference and $E(G|\text{data})$ can be very different *a posteriori*. Consider, for example, the data for diabetic mouse Treg 1 in Table 2. The data show the skewed distribution that is typical of the data sets that we consider. We can recognize a peak for low counts around 1, and some evidence for a secondary peak for high counts around 36 and 40. Under the negative binomial model, posterior inference must balance between the very few sequences with high counts and the several sequences with low counts. The resulting posterior shows a single peak around 6. In contrast, the Dirichlet process mixture model can capture the observed imbalance in the data, allowing for a peak at count 0, and a secondary, much smaller, peak around 38 (which is not shown).

Similar parametric (finite) Poisson mixture models for the analysis of SAGE data have been considered also by Zuyderduyn (2007) as a way to describe the overdispersion that is typical of these data and to identify sets of coexpressed genes. Although similar in motivation, our

Table 2. Clonal size (counts) distributions of regulatory, Treg, and conventional T-cells, Tconv, across samples of three healthy and two diabetic mice†

Count	*Results for healthy mice 1 and 2*				*Results for diabetic mice 1–3*					
	Tconv		*Treg*		*Tconv*			*Treg*		
	1	*2*	*1*	*2*	*1*	*2*	*3*	*1*	*2*	*3*
1	78	37	40	36	46	14	29	8	3	11
2	5	11	5	18	17	6	16	1	0	0
3	1	5	5	3	0	6	4	2	0	0
4	1	2	2	1	0	4	3	0	0	0
5	0	0	3	0	0	4	0	1	0	0
6	0	0	0	0	0	5	0	0	0	0
7	0	0	0	0	0	0	2	0	0	0
8	0	0	0	0	0	0	0	0	1	0
10	0	0	0	0	0	1	0	0	0	0
12	0	0	0	0	0	0	0	0	1	1
16	0	0	0	0	0	0	0	0	1	0
17	0	0	0	0	0	0	0	0	1	1
20	0	0	0	0	0	0	0	0	0	1
23	0	0	0	0	0	0	0	0	1	0
36	0	0	0	0	0	0	0	1	0	0
39	0	0	0	0	0	0	0	0	1	1
40	0	0	0	0	0	0	0	1	0	0
k'	85	55	55	58	63	40	54	14	9	15
n	95	82	88	85	80	120	99	97	118	99

†Here, k' is the total number of distinct sequences (clonotypes) in the samples, whereas n is the total number of TCR sequences per sample. The total number of distinct TCR sequences across all samples is 234 for the healthy mice and 172 for the diabetic mice.

approach does not require us to fix in advance the number of mixture components and it allows for simultaneous estimation of the number of missing sequences.

Finally, the choice of the prior on k requires careful consideration. Inference on k can be very sensitive to the choice of prior, and a poor prior may result in an overshrinkage of the frequency estimates of the scarce sequences. The prior should ideally reflect specific expert information. In lack of such information, we suggest proceeding with a conservative approach. For example, a prior centred around the number of observed sequences, k', provides a reasonable default choice.

Implementing posterior simulation includes a Metropolis transition probability to update k, using a simple discrete proposal (move up or move down). Since the number of parameters of the model is implicitly determined by the non-parametric prior, there is no need for a reversible jump algorithm (Green, 1995). In the on-line supporting information, we discuss full details of the conditional distributions that are used to define the transition probabilities of a posterior simulation scheme.

2.2. Estimation of T-cell receptor diversity

In immunology, T-cell diversity is a key quantity to understand how an efficient immune system can react to a virtually infinite set of micro-organisms without responding against body components at the same time (Nikolich-Zugich *et al.*, 2004). T-cell diversity is usually defined as the number of distinct TCRs collectively presented by this cell type. Distinction between these molecular receptors is often made through the corresponding nucleotide or amino acid sequence encoding them, as illustrated in Table 1, section (a). These receptors are generated during T-cell development in the thymus through a recombination mechanism where different gene segments are randomly assembled. Therefore, TCR diversity might be extremely high in the body and can only be ascertained through sampling. In this scenario, one usually collects a sample of T-cells and determines their corresponding TCR sequences. Data are then summarized as sequence counts as shown in Table 1.

We analyse data that report the number of distinct TCR sequences and their respective abundances obtained from three NOD and two B6 mice as summarized in Table 2 (Ferreira *et al.*, 2009). NOD and B6 are two laboratory strains that are commonly used in immunology and hereafter are referred to as diabetic and healthy mice respectively, since the former strain spontaneously develops type I diabetes whereas the latter maintains stably healthy under strict laboratory conditions. The goal of the analysis is to estimate TCR diversity of two important T-cell types, the so-called regulatory and conventional CD4+ T-cells, which have been implicated in the pathogenesis of type I diabetes. We fit the semiparametric hierarchical model (3) to each sample, i.e. the mice in Table 2. Here the index i in model (2) denotes the ith unique sequence and y_i denotes the respective observed abundance. The parameter k is the unknown TCR diversity. We use the following hyperparameters for the non-parametric prior. The mass parameter of the Dirichlet process is fixed at $\nu=1$. This implies a prior expectation of $E(L)=\nu\log\{(\nu+k')/\nu)\}\approx\log(k')$ clusters in the population (Antoniak, 1974). The centring measure $G^*=\text{Ga}(1.0,0.05)$ (mean $=20$; $\sigma=400$) was chosen to provide substantial probability for a large range of y_i-values, as it is common to observe a long-tailed sequence abundance distribution in this type of data. The prior distribution for k, the unknown TCR diversity of each T-cell population, is assumed to be a Poisson distribution with mean $\Xi=172$ and $\Xi=234$ for diabetic and healthy mice respectively, corresponding to the overall number of distinct sequences observed across all samples of the same laboratory strain. Centring near the observed diversity is a conservative choice to avoid an overestimation of k by shrinkage to the prior mean. Indeed, we also considered alternative values for Ξ but found no remarkable variations on the resulting inference over k. Finally, we speed up posterior simulation by analytically integrating out the latent sequence abundance λ_i, leaving us with the marginal model of the cluster membership indicators s_i only. For the Markov chain Monte Carlo implementation we took advantage of the split–merge algorithms proposed in Jain and Neal (2004) and Dahl (2003). See the on-line supporting information for details.

Fig. 2(a) shows the box plots of the posterior distributions for the TCR diversity k. The posterior distributions for TCR diversity are essentially the same across all samples of healthy mice (Fig. 2(a)). In other words, in healthy mice, regulatory and conventional T-cells seem equally diverse in terms of TCR diversity. In the case of the diabetic mice data, the posterior distributions for TCR diversity suggest that regulatory T-cells are less diverse than conventional T-cells (Fig. 2(b)). The result is in agreement with previous findings (Ferreira *et al.*, 2009).

For these results we carried out separate inference for each sample in Table 2. Later we shall extend model (2) to allow joint inference across samples in a single hierarchical model and compare abundance of each sequence across cell types and laboratory strains.

2.3. Simulation study

We carried out a simulation study to confirm that the semiparametric model proposed can indeed address some of the shortcomings of a traditional parametric model. We created a simulation

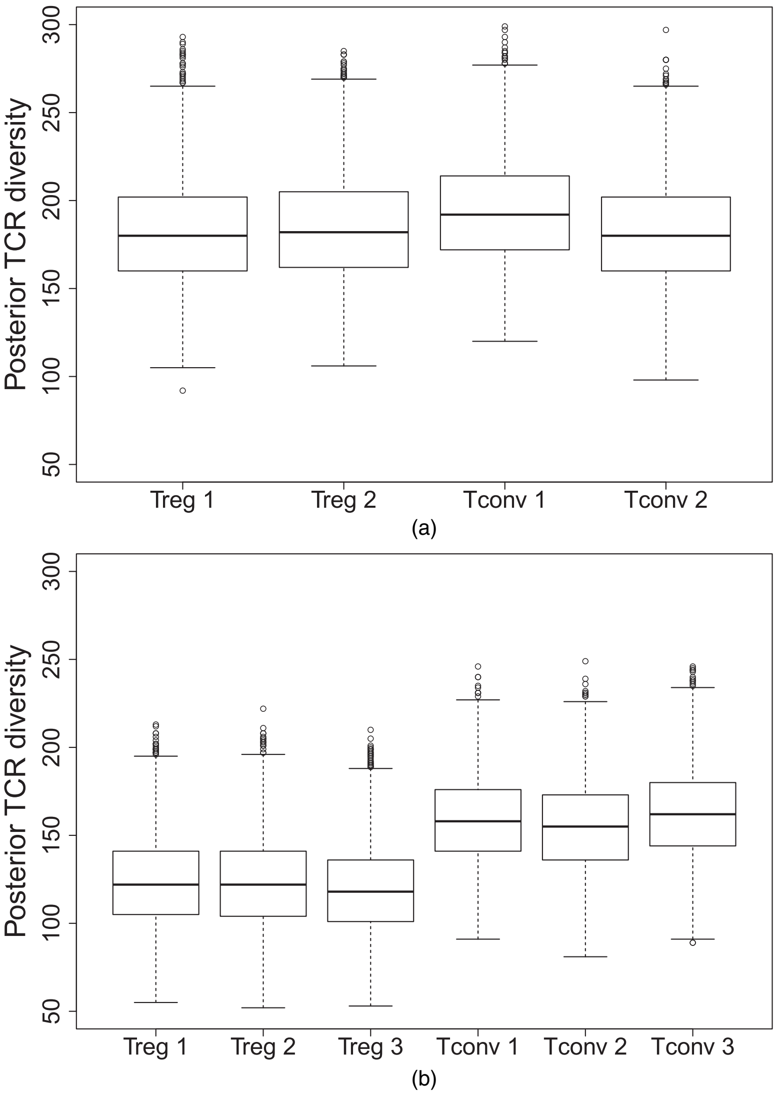


Fig. 2. Box plots of the posterior distribution of TCR diversity *k* for regulatory, Treg, and conventional T-cells, Tconv, across the various samples of (a) healthy and (b) diabetic mice

setting with 172 unique TCR sequences by pooling all samples from the experiments with diabetic mouse strains described in Section 2.2. We decided on the use of such pooled data in the absence of a gold standard data set for TCR diversity estimation and because similar pooled data are commonly used in the immunological literature, presumably to obtain larger sample sizes. Experimental constraints make it difficult to obtain large individual samples (Wong *et al.*, 2007; Hsieh *et al.*, 2004). Also, pooling different data sets has the effect of producing intricate sequence abundance distributions (Sepúlveda, 2009).

Fig. 3 shows the proportions of the unique sequences in the pooled data. Data simulation was implemented as a multinomial sampling of 250 sequences from the population of 172 unique

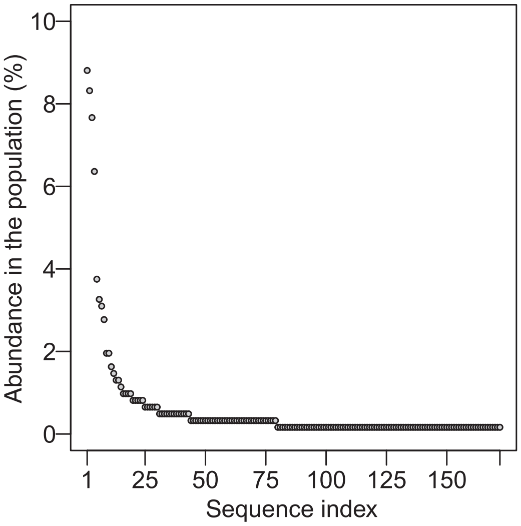


Fig. 3. Relative abundance of 172 unique TCR sequences resulting from pooling all samples of the diabetic mouse strain together

sequences. The probabilities of each hypothetical sequence were defined by the corresponding relative abundances in the pooled data for the diabetic mouse strains.

We generated $M=100$ different data sets. In each simulated data set we carried out inference under the proposed semiparametric approach using a uniform prior with support between k' and 5000 for the unknown true diversity k. Results were compared with those obtained under previously proposed parametric models for TCR diversity estimation, namely the homogeneous Poisson, the geometric, the Poisson–gamma, the Poisson–log-normal and the Yule model (Sepúlveda *et al.*, 2010). The homogeneous Poisson model assumes a single Poisson distribution for all sequence counts. The other four models assume a mixture of Poisson sampling models as in model (2) with exponential, gamma and log-normal prior densities in the geometric, Poisson–gamma and Poisson–log-normal models respectively. The Yule model uses a mixture of exponential distributions (Sepúlveda *et al.*, 2010). The five parametric models were estimated by using the same uniform prior on k and appropriate uniform priors for the remaining parameters.

For each simulation and each model we computed bias and coverage of highest posterior density credible intervals. Bias was measured as the average difference (over the M repeated simulations) between a Bayesian point estimate (posterior mean or median) for k and the simulation truth ($k=172$). Coverage is reported for 95% and 99% highest posterior density credible intervals calculated according to the method proposed by Chen and Shao (1999). Results are summarized in Table 3.

Inference about the unknown true diversity relies on the assumed model and its ability to recover the salient features of the true sampling distribution from the observed counts. Inference under none of the models attains both a precise and a close estimate of true diversity. However, inference under the semiparametric approach proposed compares favourably with the other models in terms of small bias and high coverage. We believe that this is mostly due to the flexibility of the Bayesian semiparametric specification, which does not constrain the sampling distribution to a specific shape. Also, compared with the parametric approaches, the lengths of the credible intervals under the Bayesian semiparametric model are neither too narrow nor too

Table 3. Summary results of the simulation study based on $M = 100$ data sets from a population of 172 unique TCR sequences whose relative abundances are shown in Fig. 3†

Model	*Bias*		*Credible interval length*		*Coverage*	
	Mean	*Median*	*95%*	*99%*	*95%*	*99%*
Homogeneous Poisson	−77.2	−77.4	9	11	0.00	0.00
Geometric	−35.4	−36.0	34	44	0.11	0.21
Poisson–gamma	2287.1	2001.9	4490	4669	0.00	0.00
Poisson–log-normal	1015.0	678.5	3588	4630	0.56	0.58
Yule	109.8	107.4	123	160	0.00	0.03
Semiparametric	56.75	53.24	173	245	1.00	1.00

†The reported credible interval lengths are median lengths across simulations.

wide to make inference meaningless. With respect to the 95% and 99% highest posterior density credible intervals and the observed overcoverage, we note that the (frequentist) coverage need not exactly match the nominal (posterior) probability of the credible interval. In particular, in non-parametric Bayesian inference there is no equivalent to the Bernstein–von Mises theorem, which states that credible sets for parametric models are asymptotically equivalent to confidence regions based on maximum likelihood estimators (see Freedman (1999)). We refer the interested reader to recent work that attempts to provide a better understanding of frequentist properties of non-parametric Bayes procedures, e.g. Knapik *et al.* (2011) and Castillo (2010). In summary, our comparison confirms some recent theoretical studies that have demonstrated that the sequence abundance distribution might be too complex to be captured by simple parametric models (Sepúlveda, 2009). The Bayesian semiparametric approach provides a natural and flexible modelling choice to tackle TCR diversity and gives a reasonable account of all uncertainties in the specific estimation problem.

2.4. *Estimation of serial analysis of gene expression tag abundances*

We consider a SAGE library of normal colon epithelium tissue with a total number of 49610 recorded tags. Here, tags refer to the RNA fragments being counted in the experiment. The library is publicly available on SAGE Genie (`http://cgap.nci.nih.gov/SAGE`, NC1 library) and has been analysed by many researchers (Zhang *et al.*, 1997; Stollberg *et al.*, 2000; Morris *et al.*, 2003). A total of 17703 distinct tags were observed from a healthy colon tissue of a single individual. Most distinct tags were recorded with low counts: 75% of the tags were observed once and 92% of the tags show fewer than five copies. However, tags with low counts represent only 26.4% of the total messenger RNA mass expressed in the sample. This skewed distribution with many scarce tags and few abundant tags is typical of SAGE experiments (Morris *et al.*, 2003). We fit model (2)–(3) to this data set, with the following choices of the hyperparameters. The base measure of the Dirichlet process is $G^* = \text{Ga}(1.0, 0.05)$ (mean $= 20$; $\sigma = 400$), and the mass parameter is $\nu = 5.0$, corresponding to $E(L) \approx 40$. This allows for a large range of λ_i-values as well as flexibility in the estimation of the density of the observed tag counts. We also investigated alternative choices of ν (which are not shown). As expected, different ν did not affect the estimate of $E(k|\text{data})$ but only the number of clusters in equation (2). We considered three values for the hyperparameter Ξ in a $\text{Poi}(\Xi)$ prior for k: $\Xi = 17703$, which corresponds to centring around the observed number of distinct tags, k', and is the most conservative estimate; $\Xi = 25536$, which is the number of unique tags estimated by Stollberg

Table 4. Posterior quantiles of tag diversity k according to different prior mean choices

Prior expectation $E(k)$	*Posterior quantiles*		
	0.05	*0.50*	*0.95*
17703	19934	21371	23124
25536	21279	23437	25541
50000	24962	28218	31763

et al. (2000) in this normal colon tissue; $\Xi = 50\,000$, as a realistic upper bound. The results are reported in Table 4 and suggest that the estimate that was reported by Stollberg *et al.* (2000) might be slightly overestimating the true diversity of tags, assuming similar bias as we observed in the simulation study.

Also, we note that the influence of prior assumptions is tempered by the evidence in the data, as desired. Fig. 4 summarizes posterior inference under $\Xi = 25\,536$. Fig. 4(a) plots the semiparametric Bayes estimates *versus* the observed counts. The plot demonstrates the shrinkage profiles relative to the empirical frequencies (the diagonal line), which are the maximum likelihood estimates under a Poisson model. In particular, for censored sequences (with $y_i = 0$) the posterior mean estimate inflates the maximum likelihood estimate and reports $E(\lambda_i|\text{data}) \approx 0.9$. The shrinkage profile that is induced by the semiparametric model is non-linear and affects the rare sequences more than the abundant sequences. For abundant sequences, posterior inference is driven only by the observed counts, and $E(\lambda_i|\text{data}) \approx y_i$. Fig. 4(b) shows the estimated distribution of sequence abundances λ_i. It is highly skewed, with high probability mass on low abundances and small mass on extreme values.

3. Model for comparison across biological conditions

3.1. Multiple samples

When the data include samples across different biological conditions, then the desired inference extends beyond the estimation of the within-sample abundances. The semiparametric Bayesian approach that was introduced in the previous sections can be extended to accommodate the general case of sequence counts observed over several samples and different biological conditions. Let T and C denote respectively the total number of samples and conditions included in the data set.

As in Section 2, the total number of distinct sequences, k, is unknown owing to sampling variability. Data consist of counts y_{it} of a sequence i in sample $t = 1, \ldots, T$. Not all samples may include the same set of sequences. This is similar to the censoring of zero counts in the earlier set-up, but now with many possible censoring patterns. The characterization of censoring patterns requires some additional notation. We introduce binary indicators γ_{it} with $\gamma_{it} = 1$ if a sequence i is present in sample t; otherwise $\gamma_{it} = 0$ and $y_{it} = 0$. Then $\gamma_i = (\gamma_{i1}, \ldots, \gamma_{iT})'$ is a binary vector that records the presence of sequence i. The set of k sequences can be partitioned into subsets of units appearing in the same samples according to γ_i. We denote the partitioning subsets by $\mathcal{K}_j$, indexed by a running integer $j = 0, \ldots, 2^T - 1$. The mapping from γ_i to j is determined by $j = \Sigma_{t=1}^T 2^{t-1} \gamma_{it}$. Vice versa, the digits of j in a binary expansion are the γ_{it}. Let γ_j^* denote the common value of γ_i for all units in $\mathcal{K}_j$, i.e. $\gamma_i = \gamma_j^*$ for all $i \in \mathcal{K}_j$. For example, if $T = 2$, then $\gamma_0^* = (0, 0)$, $\gamma_1^* = (1, 0)$, $\gamma_2^* = (0, 1)$ and $\gamma_3^* = (1, 1)$.

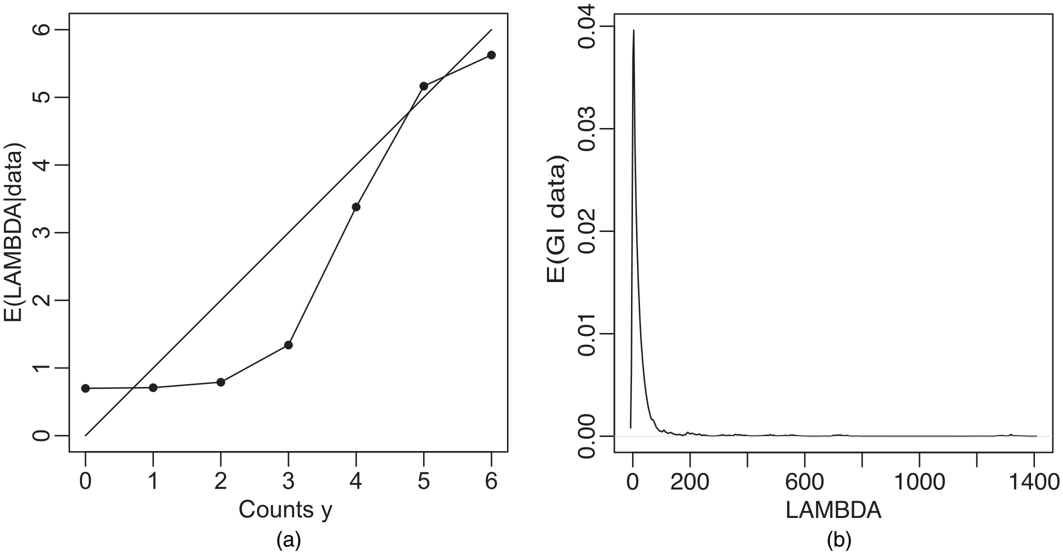


Fig. 4. Analysis of the SAGE library of normal colon epithelium tissue: (a) posterior means $E(\lambda_i|\text{data})$ on the vertical axis against observed counts y_i on the horizontal axis (zooming in to small counts, where the shrinkage with respect to the maximum likelihood estimates is most noticeable); (b) estimated mixing measure $E(G|\text{data})$

Also denote by $k_j = |\mathcal{K}_j|$ the size of each of these sets, which corresponds to the number of sequences with the same pattern of being observed and censored across samples. Here, k_0 is the number of sequences with zero counts across all samples, i.e. the number of sequences that are zero censored in all samples. The model construction proceeds similarly to before. We start with a model for censoring patterns and then use a semiparametric mixture model to specify $G(\cdot)$. Specifically, we extend the model that was discussed in Section 2 by assuming that the distribution of k_j is multinomial:

$$p(k_j, j=0,\ldots,2^T-1|k,G) = \binom{k}{k_0 \ldots k_{2^T-1}} \prod_{j=0}^{2^T-1} G_+(j)^{k_j},$$

where $G_+(j) = \Pr(\cap_{t:\gamma^*_{jt}=0}\{y_{it}=0\} \cap \cap_{t:\gamma^*_{jt}=1}\{y_{it}>0\})$ denotes the probability that a sequence belongs to the partition set $\mathcal{K}_j$. Similarly to the discussion in Section 2 and in accordance with the interpretation of the set $\mathcal{K}_0$, $G_+(0)$ denotes the probability of all zero counts. Denote the set of recorded observations by $\mathbf{y}_+ = \{y_{it} : y_{it} > 0, i=1,\ldots,k, t=1,\ldots,T\}$ and let $\mathbf{y}_i = \{y_{it}, t=0,\ldots,T\}$. Also, let $\mathbf{k} = (k, k_j, j=0,\ldots,2^T-1)^{\mathrm{T}}$. Then, we can write

$$p(\mathbf{y}_+|\mathbf{k}) = \prod_{j=0}^{2^T-1} \prod_{i\in\mathcal{K}_j} G(\mathbf{y}_i)/G_+(j),$$

where $G(\mathbf{y}_i) = \Pr(Y_{it} = y_{it}, i \in \mathcal{K}_j)$ is the sampling distribution for the observations in the partition set $\mathcal{K}_j$. Hence, the likelihood of the observed counts can be expressed as

$$p(\mathbf{y}, k_j, j=0,\ldots,2^T-1|k,G) = \binom{k}{k_0 \ldots k_{2^T-1}} \prod_{j=0}^{2^T-1} \prod_{i\in\mathcal{K}_j} G(\mathbf{y}_i). \tag{5}$$

The framework is sufficiently general to accommodate a wide range of sampling distributions. In the following examples we shall again assume the random mixture of Poisson distributions models as in expression (2).

3.2. Multiple conditions

We extend the model to study the abundance of sequences across different conditions. We denote by $x_t = x$, $x \in \{1, \ldots, C\}$, the type of tissue (or condition) collected in a sample t. Without loss of generality, we assume that $x = 1$ refers to samples from a reference tissue (e.g. a healthy cell), whereas $x > 1$ indexes treatment samples (e.g. tumour).

We replace the hierarchical model (3) by the more general

$$y_{it} | \lambda_{ix_t}, x_t, n_t \sim \text{Poi}(\lambda_{ix_t} n_t) \qquad i = 1, \ldots, k, \quad t = 1, \ldots, T, \tag{6}$$

where n_t is the size of sample t. Here, the abundance parameter λ_{ix_t} is normalized with respect to the total level of abundance n_t in each sample t, $t = 1, \ldots, T$. This accounts for between-library variability (Baggerly *et al.*, 2003). Also, we assume that the abundance rate is condition specific, i.e. $\lambda_{ix_t} = \lambda_{ix}$ for all samples with $x_t = x$. When t is a library from the reference tissue, we assume

$$\lambda_{ix_t} | P \overset{\text{IID}}{\sim} P, \qquad \text{if } x_t = 1, \tag{7}$$

as in model (2), $i = 1, \ldots, k$. For other types of tissue, $x \neq 1$, we allow for the possibility of different abundance rates $\lambda_{ix_t} \neq \lambda_{i1}$. We augment the hierarchical model (3) by an additional layer that models differential abundance across tissue types. We assume that

$$\lambda_{ix_t} | \pi, P, \lambda_{i1} \sim \pi I(\lambda_{ix_t} = \lambda_{i1}) + (1 - \pi) P \qquad \text{if } x_t = 2, \ldots, C. \tag{8}$$

For each condition, x_t, the prior probability of non-differential abundance with respect to the reference tissue is $\pi + (1 - \pi)/(\nu + 1)$, where the second term arises from the event that two independent draws from P are tied and ν is the mass parameter of the Dirichlet process as in Section 2.1. The marginal distribution of λ_{ix} is P for any given x, i.e.

$$p(\lambda_{ix}, i = 1, \ldots, k | P) = \prod_{i=1}^{k} P(\lambda_{ix}).$$

We introduce latent variables $w_{ix} \sim \text{Bern}(\pi)$ and reformulate expression (8) as

$$\lambda_{ix} | w_{ix}, \lambda_{i1} \sim \begin{cases} I(\lambda_{ix} = \lambda_{i1}) & \text{if } w_{ix} = 1, \\ P & \text{if } w_{ix} = 0. \end{cases}$$

The use of the latent variables w_{ix} results in an augmentation scheme that simplifies posterior simulation. In particular, it simplifies inference on the λ_{ix}, and hence the probabilities of differential abundance, and hence also inference on k. Finally, note that we could always modify equation (8) to obtain more detailed inference across conditions. In particular, if three or more conditions were examined, it would be possible to include in equation (8) some terms that allow for non-differential abundance among any pairs of conditions.

3.3. Comparison of T-cell receptor abundances

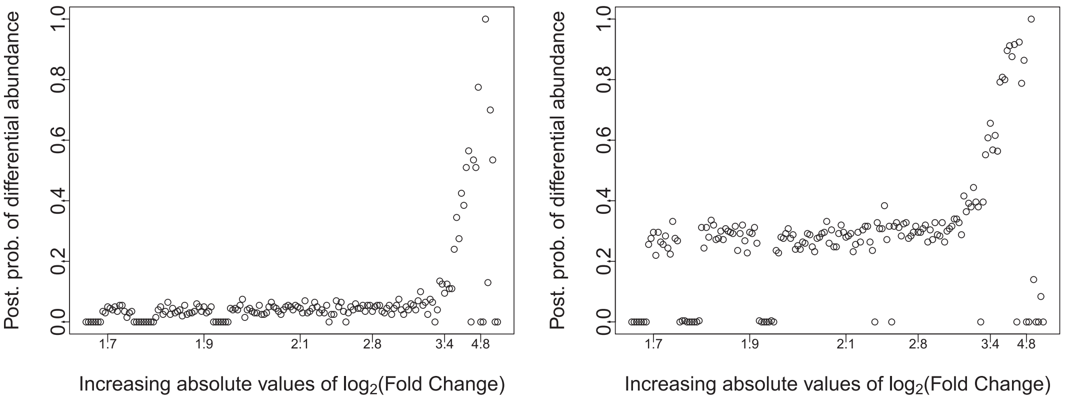
Many recent studies have tried to understand better the protective role of regulatory T-cells against several autoimmune diseases, such as type I diabetes or lupus. Some of these studies were designed to ascertain core differences between the TCR frequency presented by these cells and their conventional T-cell counterparts (Hsieh *et al.*, 2004, 2006; Pacholczyk *et al.*, 2006, 2007; Wong *et al.*, 2007; Ferreira *et al.*, 2009; Sepúlveda *et al.*, 2010). In this section we extend the analysis shown in Section 2.2 to compare the abundances of each TCR sequence across different conditions.

We analysed the data under model (6)–(8) by using the same prior specification as in Section 2.2. Additionally, we chose $\pi = 0.9$. This choice limits findings to a small set of scientifically interesting prospects (Efron, 2008). For comparison we also considered $\pi = 0.5$. We shall describe several comparisons. Each comparison is across $C = 2$ conditions. See below for details. Let $v_i = \Pr(\lambda_{i1} \neq \lambda_{i2}|\text{data})$ be the posterior probability of differential abundance of TCR sequence i across the two conditions. When multiple comparison is set as a Bayesian decision problem, the optimal decision rule for many common loss functions can be defined as identifying a TCR sequence with differential abundance across cell type or mouse strain whenever $v_i > \delta$ for some threshold δ (see Müller *et al.* (2007) and Bogdan *et al.* (2008)). It is worth noting that the choice of π generally affects inferences over v_i but not the resulting ordering of the TCR sequences.

We start by comparing the abundances of each TCR sequence in the diabetic mice data. After pooling samples of the same cell type, we compare counts in regulatory T-cells *versus* conventional T-cells. Fig. 5 plots the posterior probability v_i under the model proposed *versus* the difference on a scale of base 2 logarithm ($\log_2$-fold change) as calculated by the R package EdgeR (Robinson and Smyth, 2007). According to Fig. 5 TCR sequences with differential abundances are usually associated with the highest $\log_2$-fold change difference with few exceptions. The number of differentially abundant TCR sequences, i.e. $v_i > \delta$, decreases rapidly with increasing threshold δ. But, for a fixed threshold δ, the number of detections may depend on the choice of π.

Next we compared counts across healthy *versus* diabetic mice. Fig. 6 shows the comparison, for the regulatory T-cells (Fig. 6(a)) and conventional T-cells (Fig. 6(b)). Fig. 6 plots the posterior probability v_i *versus* the $\log_2$-fold change. Fig. 6 suggests that there are no TCR sequences with differential abundance across mouse strains. For this analysis we used $\pi = 0.9$. A qualitatively similar pattern was observed by using $\pi = 0.5$ (the results are not shown), but the corresponding posterior values were inflated.

For both comparisons, the R package EdgeR did not report any significantly different abundances. The smallest p-value was greater than 0.25 in all cases. The method of Robinson and Smyth (2007), which is implemented in the EdgeR package, relies on the estimation of a common overdispersion parameter across samples. The estimation of this overdispersion parameter improves with the sample size. However, in some cases, including the T-cell sequences that are considered in this work, the number of sequences available in each sample is limited by



(a) (b)

Fig. 5. Comparison of TCR sequence abundances in regulatory and conventional T-cells-collected diabetic mice: v_i *versus* the corresponding ranking based on the absolute values of $\log_2$-fold changes as computed in EdgeR, with the prior probability of differential expression in assumption (8) set to (a) $\pi = 0.9$ and (b) $\pi = 0.5$

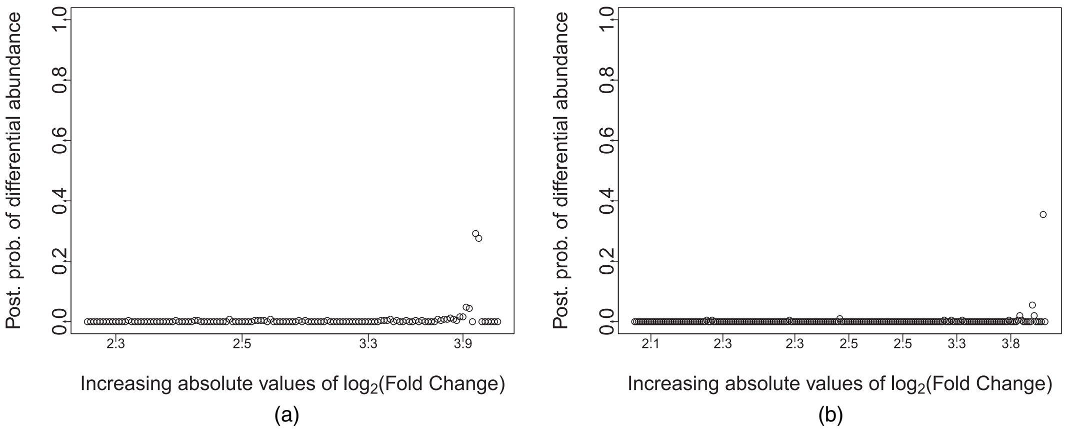


Fig. 6. Comparison of TCR sequence abundances between diabetic and healthy mice: posterior probability of differential abundances across distinct TCR sequences as a function of their ranking *versus* the corresponding ranking based on the absolute values of $\log_2$-fold changes as computed in EdgeR for (a) regulatory T-cells and (b) conventional T-cells

the nature of the experiment. Hence, the estimation of a single overdispersion parameter can be particularly challenging and it might affect the overall inferential results.

3.4. Comparison of serial analysis of gene expression libraries

The previous analysis is now extended to compare four libraries of normal and primary cancer colon epithelium tissues. The same data have been analysed by several researchers before (Zhang *et al.*, 1997; Baggerly *et al.*, 2004; Robinson and Smyth, 2007).

The number of unique tags recorded across the four libraries is 55 209. We implement model (6)–(8), with $\pi = 0.9$, and the specifications of the prior hyperparameters as described in Section 2.4. As a prior for the parameter k, we assumed a Poisson distribution Poi(Ξ), with $\Xi = 55209$. This reflects the prior belief that most tags should already be present in the libraries collected. This prior belief is in close agreement with the data as the mean posterior of the overall k is 56 125 tags. In Fig. 7, we plot the percentage of tags assigned to the alternative hypothesis, i.e. $(1/n)\,\Sigma_i^n I(v_i > \delta)$, against the threshold δ.

We compare our findings with those reported in Baggerly *et al.* (2003) and Robinson and Smyth (2007). Baggerly *et al.* (2003) developed a beta–binomial sampling model to account for within- and between-library variability and discussed the use of test statistics $t_i^w = (\hat{p}_{i1} - \hat{p}_{i2})/\surd(\hat{V}_{i1} + \hat{V}_{i2})$ for group comparisons. Here, $\hat{p}_{ij}$, $j = 1, 2$, is a weighted sum of the proportions of tag i in each group and $\hat{V}_{ij}$ is an unbiased estimator of the variance of $\hat{p}_{ij}$. For example, $\hat{p}_{ij} = \Sigma_{t:x_t=j}\, w_t \hat{p}_{it}$, with $p_{it} = y_{it}/n_t$ and $w_t = n_t/\Sigma_{t:x_t=j}\, n_t$. Fig. 8 compares the differences $\hat{p}_{i1} - \hat{p}_{i2}$, as computed in Baggerly *et al.* (2003), against the estimated differential expressions from our model, computed as $E(\lambda_{i1}|\text{data}) - E(\lambda_{i2}|\text{data})$. Although the two quantities seem to be overall comparable, we observe many tags for which the difference $\hat{p}_{i1} - \hat{p}_{i2}$ is significantly different from 0 according to Baggerly *et al.* (2003) but are still assigned to the null hypothesis under the non-parametric prior. One of the reasons for the discrepancy is that our approach allows for ‘borrowing strength’ across tags, whereas the test in Baggerly *et al.* (2003) works with one tag at a time.

Finally, we also analysed the same data with the R package EdgeR (Robinson and Smyth, 2007) and found substantial agreement between that method and ours. More specifically, the package EdgeR identifies 243 tags as differentially expressed with p-values less than 0.01. Out of the first 250 tags that our method flags as differentially expressed, we identify the same list as

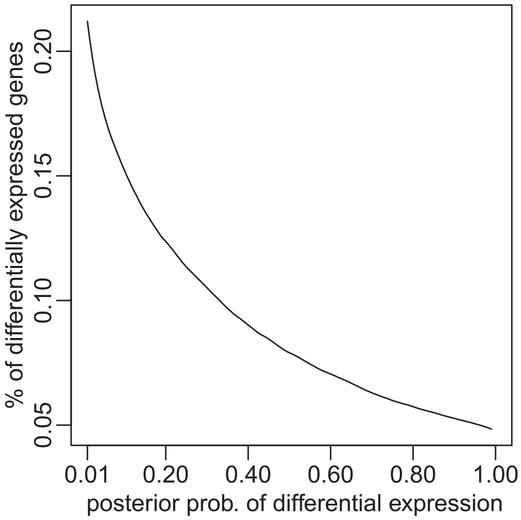


Fig. 7. Percentage of tags that are declared differentially expressed *versus* the threshold δ in the rule $v_i > \delta$ to determine differential expression: see Section 3.3 for details

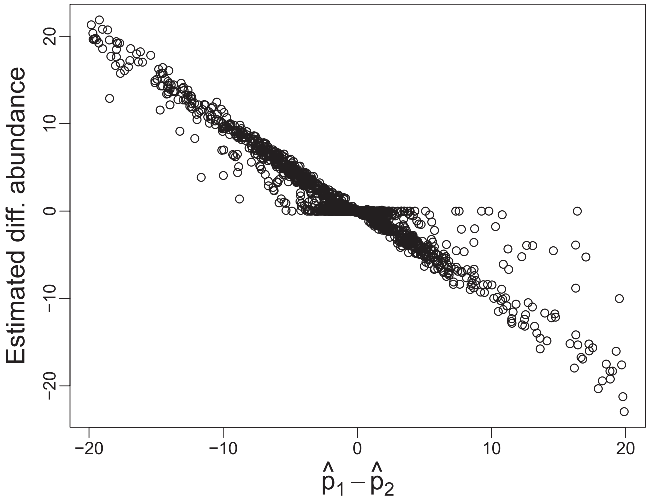


Fig. 8. Differential abundance estimated according to model (6)–(8) *versus* the difference of weighted proportions $\hat{p}_1 - \hat{p}_2$ as in Baggerly *et al.* (2003) for the SAGE libraries considered in Section 3.3: the zoom to absolute differences less than 20 highlights the shrinkage implied by the non-parametric prior

that from EdgeR, except for five tags which are assigned probability of differential expression equal to 0 by our model. The difference in absolute counts in those five cases may be explained by the slightly larger size of the two tumour libraries.

4. Summarizing remarks

We have discussed a coherent Bayesian semiparametric approach for the analysis of sequence counts data. The strength of our modelling framework is that it adapts easily to analyse data sets of different dimensions. In particular, we can apply it to small but still overdispersed data sets like those described in Sections 2.2 and 3.3, which may pose a challenge for alternative methods that strongly rely on an accurate estimate of some overdispersion parameter (e.g. Robinson and Smyth (2007)). Many experimental techniques imply data censoring in the sense that many sequences might not be observed. We account for unobserved sequences by explicitly modelling the discrepancy between population and sample diversity. We showed how the shrinkage properties that are implied by the non-parametric prior allow estimation of the true abundance of scarcely represented sequences without affecting the estimation of true expression for more abundant tags. Finally, we showed how our modelling framework can be extended to tackle the general multicomparison problem across samples and conditions.

The implementation of the proposed method is computationally intensive. For example, for the application in Section 3, it took a C++ program 1 day to conclude approximately 1000 iterations on an Intel Quad-Core processor with 4 Gbytes memory. The use of fast mixing samplers, such as the sequentially allocated merge–split sampler of Dahl (2003), is therefore particularly recommended. In particular, the application of this methodology to modern sequencing technologies may require the use of fast approximation algorithms such as those described in Blei and Jordan (2006), Daumé (2007) and Wang and Dunson (2011). In addition, models of sequence formation like that proposed by Gilchrist *et al.* (2007) could be incorporated in the analysis to account explicitly for a varying probability of generating an observable unit across sequences and experiments.

Finally, the proposed corrections for censoring are not needed for all types of sequence count data. For example, for high throughput data the correction is less meaningful, because censoring at zero counts is not an issue in that context with large total counts.

Acknowledgements

The authors gratefully acknowledge the Joint Editor, the Associate Editor and two referees for their constructive comments on earlier versions of the paper. Nuno Sepúlveda and Carlos Daniel Paulino were partially funded by Fundação para a Ciência e Tecnologia, Portugal, through the project Pest-OE/MAT/UI0006/2011. Nuno Sepúlveda was also financed by a grant from the Foundation for the National Institutes of Health (grant 566) and the Wellcome Trust (grant 077383/Z/05/Z) through the 'Grand challenges in global health initiative'. Peter Müller was partially supported by grant NIH/R01 CA075981.

References

Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, article R106.

Antoniak, C. E. (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.*, **2**, 1152–1174.

Baggerly, K. A., Deng, L., Morris, J. S. and Aldaz, C. M. (2003) Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics*, **19**, 1477–1483.

Baggerly, K. A., Deng, L., Morris, J. S. and Aldaz, C. M. (2004) Overdispersed logistic regression in SAGE. *BMC Bioinform.*, **5**, article 144.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. U., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M., Reed, M. I., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, O., Aniebo, I. C., Bailey, D. M., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N.

P., Castillo, N., Chiara, E., Catenazzi, M., Chang, S., Cooley, N. R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnishik, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G. D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriquez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin. R. and Smith, A. J. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

Blei, D. and Jordan, M. (2006) Variational inference for dirichlet process mixture models. *Baysn Anal.*, **1**, 121–144.

Bogdan, M., Gosh, J. and Tokdar, S. (2008) A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen* (eds N. Balakrishnan, E. Peña and M. Silvapulle), pp. 211–230. Beachwood: Institute of Mathematical Statistics.

Cameron, A. and Trivedi, P. (1998) *Regression Analysis of Counts Data*. Cambridge: Cambridge University Press.

Canale, A. and Dunson, D. (2012) Bayesian kernel mixtures for counts. *J. Am. Statist. Ass.*, **106**, 1528–1539.

Castillo, I. (2010) A semiparametric Bernstein-von Mises theorem for Gaussian process priors. *Probab. Theor. Reltd Flds*, **152**, 53–99.

Chen, M. H. and Shao, Q. M. (1999) Monte Carlo estimation of Bayesian credible and HPD intervals. *J. Computnl Graph. Statist.*, **8**, 69–92.

Dahl, D. (2003) An improved merge-split sampler for conjugate Dirichlet process mixture models. *Technical Report 1086*. Department of Statistics, University of Wisconsin, Madison.

Daumé III, H. (2007) Fast search for dirichlet process mixture models. In *Proc. 11th Int. Conf. Artificial Intelligence and Statistics, San Juan*.

Dhavala, S. S., Datta, S., Mallick, B. K., Carroll, R. J., Khare, S., Lawhon, S. D. and Adams, L. G. (2010) Bayesian modeling of MPSS data: gene expression analysis of bovine salmonella infection. *J. Am. Statist. Ass.*, **105**, 956–967.

Efron, B. (2008) Microarrays, empirical bayes and the two-groups model. *Statist. Sci.*, **23**, 1–22.

Escobar, M. D. and West, M. (1995) Bayesian density estimation and inference using mixtures. *J. Am. Statist. Ass.*, **90**, 577–588.

Favaro, S., Lijoi, A., Mena, R. H. and Prünster, I. (2009) Bayesian non-parametric inference for species variety with two-parameter Poisson–Dirichlet process prior. *J. R. Statistic. Soc.* B, **71**, 993–1008.

Favaro, S., Lijoi, A., and Prünster, I. (2012) Conditional formulae for Gibbs-type exchangeable random partitions. *Ann. Appl. Probab.*, to be published.

Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, **1**, 209–230.

Ferguson, T. (1983) Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics* (eds H. Rizvi and J. Rustagi), pp. 287–302. New York: Academic Press.

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. Natn. Acad. Sci. USA*, **106**, 8320–8325.

Freedman, D. (1999) On the Bernstein-von Mises theorem with infinite dimensional parameters. *Ann. Statist.*, **27**, 1119–1140.

Gasparini, M. (1996) Bayesian density estimation via dirichlet density processes. *J. Nonparam. Statist.*, **6**, 355–366.

Gilchrist, M., Qin, H. and Zaretzi, R. (2007) Modelling SAGE tag formation and its effects on data interpretation within a Bayesian framework. *BMC Bioinform.*, **8**, article 403.

Green, P. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Green, P. and Richardson, S. (2001) Modelling heterogeneity with and without the Dirichlet process. *Scand. J. Statist.*, **28**, 355–377.

Hardcastle, T. and Kelly, K. (2010) bayseq: empirical bayesian methods for identifying differential expression in sequence count data. *BMC Bioinform.*, **11**, article 422.

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**, 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**, 401–410.

Jain, S. and Neal, R. M. (2004) A split-merge Markov chain Monte Carlo procedure for the Dirichlet Process mixture model. *J. Computnl Graph. Statist.*, **13**, 158–182.

Knapik, B., van der Vaart, A. and van Zanten, J. (2011) Bayesian inverse problems with gaussian priors. *Ann. Statist.*, **39**, 2626–2657.

Lijoi, A., Mena, R. H. and Prünster, I. (2007a) Bayesian Nonparametric estimation of the probability of discovering new species. *Biometrika*, **94**, 769–786.

Lijoi, A., Mena, R. H. and Prünster, I. (2007b) A Bayesian Nonparametric method for prediction in EST analysis. *BMC Bioinform.*, **8**, article 339.

Lijoi, A., Mena, R. H. and Prünster, I. (2008) A Bayesian Nonparametric approach for comparing clustering structures in EST libraries. *J. Computnl Biol.*, **15**, 1315–1327.

MacEachern, S. N. and Müller, P. (1998) Estimating mixtures of Dirichlet process models. *J. Computnl Graph. Statist.*, **7**, 223–238.

Morris, J., Baggerly, K. and Coombes, K. (2003) Bayesian shrinkage estimators of the relative abundance of mRNA transcripts using SAGE. *Biometrics*, **59**, 476–486.

Müller, P., Parmigiani, G. and Rice, K. (2007) FDR and Bayesian multiple comparisons rules. In *Bayesian Statistics 8* (eds J. M. Bernardo, M. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West). Oxford: Oxford University Press.

Neal, R. M. (2000) Markov chain sampling methods for Dirichlet process mixture models. *J. Computnl Graph. Statist.*, **9**, 249–265.

Nie, L., Wu, G., Brockman, F. J. and Zhang, W. (2006) Integrated analysis of transcriptomic and proteomic data of desulfovibrio vulgaris: zero-inflated Poisson regression models to predict abundance of undetected proteins. *Bioinformatics*, **22**, 1641–1647.

Nikolich-Zugich, J., Slifka, M. K. and Messaoudi, I. (2004) The many important facets of t-cell repertoire diversity. *Nat. Rev. Immunol.*, **2**, 123–132.

Pacholczyk, R., Ignatowicz, H., Kraj, P. and 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. and Ignatowicz, L. (2007) Nonself-antigens are the cognate specificities of Foxp3+ regulatory T cells. *Immunity*, **27**, 493–504.

Papaspiliopoulos, O. and Roberts, G. O. (2008) Retrospective Markov chain Monte Carlo methods for Dirichlet Process hierarchical models. *Biometrika*, **95**, 169–186.

Quintana, F. A. and Iglesias, P. L. (2003) Bayesian clustering and product partition models. *J. R. Statist. Soc.* B, **65**, 557–574.

Rempala, G. A., Seweryn, M. and Ignatowicz, L. (2011) Model for comparative analysis of antigen receptor repertoires. *J. Theor. Biol.*, **269**, 1–15.

Robinson, M. D. and Smyth, G. K. (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**, 2881–2887.

Sepúlveda, N. (2009) How is the T-cell repertoire shaped? *PhD Thesis*. University of Oporto, Oporto.

Sepúlveda, N., Paulino, C. D. and Carneiro, J. (2010) Estimation of T-cell repertoire diversity and clonal size distribution by Poisson abundance models. *J. Immunol. Meth.*, **35**, 124–137.

Stollberg, J., Urschitz, J., Urban, Z. and Boyd, C. (2000) A quantitative evaluation of SAGE. *Genome Res.*, **10**, 1241–1248.

Thygesen, H. and Zwinderman, A. (2006) Modeling SAGE data with a truncated Gamma-Poisson model. *BMC Bioinform.*, **7**, article 157.

Trippa, L. and Parmigiani, G. (2011) False discovery rate in somatic mutation studies of cancer. *Ann. Appl. Statist.*, **5**, 1360–1378.

Walker, S. G., Damien, P., Laud, P. and Smith, A. F. M. (1999) Bayesian nonparametric inference for random distributions and related functions. *J. R. Statist. Soc.* B, **61**, 485–527.

Wang, L. and Dunson, D. (2011) Fast bayesian inference in Dirichlet process mixture models. *J. Computnl Graph. Statist.*, **20**, 196–216.

Wong, J., Obst, R., Correia-Neves, M., Losyev, G., Mathis, D. and Benoist, C. (2007) Adaptation of TCR repertoires to self-peptides in regulatory and nonregulatory CD4+ T cells. *J. Immunol.*, **178**, 7032–7041.

Yoon, S., Xuan, Z., Makarov, V., Ye, K. and Sebat, J. (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.*, **19**, 1586–1592.

Zhang, L., Zhou, W., Velculescu, V., Kern, S., Hruban, R., Hamilton, S., Vogelstein, B. and Kinzler, K. (1997) Gene expression profiles in normal and cancer cells. *Science*, **276**, 1268–1272.

Zuyderduyn, S. (2007) Statistical analysis and significance testing of serial analysis of gene expression data using a Poisson mixture model. *BMC Bioinform.*, **8**, article 282.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'"A Bayesian semi-parametric approach for the differential analysis of sequence counts data" by Guidani *et. al.*—supporting information'.