Bioinformatics, 31(9), 2015, 1420–1427
doi: 10.1093/bioinformatics/btu845
Advance Access Publication Date: 5 January 2015
Original Paper

Gene expression

Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models

Andrea Rau[1,2,*], Cathy Maugis-Rabusseau[3], Marie-Laure Martin-Magniette[4,5,6,7] and Gilles Celeux[8]

[1]INRA, UMR1313 Génétique animale et biologie intégrative, Jouy-en-Josas, France, [2]AgroParisTech, UMR1313 Génétique animale et biologie intégrative, Paris 05, France, [3]Institut de Mathématiques de Toulouse, INSA de Toulouse, Université de Toulouse, Toulouse, France, [4]UMR AgroParisTech/INRA MIA 518, Paris, France, [5]INRA, UMR 1165 URGV, Saclay Plant Sciences, Evry, France, [6]UEVE, UMR URGV, Saclay Plant Sciences, Evry, France, [7]CNRS, ERL 8196, URGV, Saclay Plant Sciences, Evry, France and [8]Inria Saclay - Île-de-France, Orsay, France

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on October 30, 2013; revised on October 23, 2014; accepted on December 19, 2014

Abstract

Motivation: In recent years, gene expression studies have increasingly made use of high-throughput sequencing technology. In turn, research concerning the appropriate statistical methods for the analysis of digital gene expression (DGE) has flourished, primarily in the context of normalization and differential analysis.

Results: In this work, we focus on the question of clustering DGE profiles as a means to discover groups of co-expressed genes. We propose a Poisson mixture model using a rigorous framework for parameter estimation as well as the choice of the appropriate number of clusters. We illustrate co-expression analyses using our approach on two real RNA-seq datasets. A set of simulation studies also compares the performance of the proposed model with that of several related approaches developed to cluster RNA-seq or serial analysis of gene expression data.

Availability and and implementation: The proposed method is implemented in the open-source R package `HTSCluster`, available on CRAN.

Contact: andrea.rau@jouy.inra.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The application of high-throughput sequencing (HTS) to the study of gene expression has revolutionized the scope and depth of understanding of the genome, epigenome and transcriptome of dozens of organisms. In particular, the recent use of HTS technologies to sequence ribonucleic acid content (RNA-seq), has rivaled the use of microarrays for transcriptomic studies as it offers a way to quantify gene expression without prior knowledge of the genome sequence by providing counts of transcripts. Although both technologies seem to be complementary (Naghavachari *et al.*, 2012; SEQC/MAQC-III Consortium, 2014; Wang *et al.*, 2014), RNA-seq can provide information about the transcriptome at a level of detail not possible with microarrays, including allele-specific expression and transcript discovery.

Although a variety of different protocols exist for HTS studies, the same broad preprocessing steps are followed. Namely, if an appropriate genome sequence reference is available, reads are mapped to the genome or transcriptome; otherwise, *de novo* assembly may be used. After alignment or assembly, read coverage for a given biological entity (e.g. a gene) is subsequently calculated. The quantification of gene expression in RNA-seq data remains an active area of research (Trapnell *et al.*, 2010), and in this work, we focus on measures of digital gene expression (DGE) (counts). These count-based measures of gene expression differ substantially from data produced with microarrays. For example, RNA-seq data are discrete, positive, and highly skewed, with a very large dynamic range. In addition, due to the sampling nature of sequencing, low precision tends to be observed for weakly to moderately expressed genes (McIntyre *et al.*, 2011). Finally, sequencing depth (i.e. the library size) and coverage vary between experiments, and read counts are known to be correlated with gene length (Łabaj *et al.*, 2011; Oshlack and Wakefield, 2009).

To date, most developments concerning the statistical analysis of RNA-seq data have dealt with the issues of experimental design (Auer and Doerge, 2010), normalization (Robinson and Oshlack, 2010) and the analysis of differential expression (Anders and Huber, 2010; Auer *et al.*, 2012; Law *et al.*, 2014; McCarthy *et al.*, 2012; Zhou *et al.*, 2014). In this work, we focus on the question of co-expression analyses for RNA-seq data. Identifying biological entities that share similar profiles across several treatment conditions, such as co-expressed genes, may help identify groups of genes that are involved in the same biological processes (Eisen *et al.*, 1998; Jiang *et al.*, 2004). Clustering analyses based on metric criteria, such as the K-means algorithm (MacQueen, 1967) and hierarchical clustering (Ward, 1963), have been used to cluster microarray-based measures of gene expression as they are rapid, simple and stable. However, such methods require both the choice of metric and criterion to be optimized, as well as the selection of the number of clusters. An alternative to such methods are probabilistic clustering models, where the objects to be classified (genes) are considered to be a sample of a random vector and a clustering of the data is obtained by analyzing the density of this vector (McLachlan *et al.*, 2004; Yeung *et al.*, 2001).

Presently, most proposals for clustering RNA-seq data have focused on grouping together biological samples rather than biological entities (e.g. genes). For example, Anders and Huber (2010) perform a hierarchical clustering with a Euclidean distance of samples following a variance-stabilizing transformation, and Severin *et al.* (2010) cluster fourteen diverse tissues of soybean using hierarchical clustering with Pearson correlation after normalizing the data using a variation of the Reads Per Kilobase per Million mapped reads (RPKM) measure. Witten (2011) discussed the clustering of samples using hierarchical clustering with a modified loglikelihood ratio statistic as distance measure based on a Poisson loglinear model; this model is similar to that of Cai *et al.* (2004) for the clustering of Serial Analysis of Gene Expression (SAGE) gene profiles using a K-means algorithm and a Poisson loglinear model. More recently, Si *et al.* (2014) considered Poisson and negative binomial mixture models to develop a model-based hybrid-hierarchical clustering algorithm.

In this work, like Cai *et al.* (2004) and Si *et al.* (2014), we focus on the use of Poisson loglinear models for the clustering of count-based HTS expression profiles; however, rather than using such a model to define a distance metric to be used in a K-means or hierarchical clustering algorithm, we make use of finite mixtures of Poisson loglinear models. This framework has the advantage of providing a straightforward procedure for parameter estimation and model selection, as well as a per-gene conditional probability of belonging to each cluster.

2 Methods

Let Y_{ijl} be the random variable corresponding to the DGE measure for biological entity i ($i = 1, \ldots, n$) of condition j ($j = 1, \ldots, d$) in biological replicate l ($l = 1, \ldots, r_j$), with y_{ijl} being the corresponding observed value of Y_{ijl}. Let $q = \sum_{j=1}^{d} r_j$ be the total number of variables (all replicates in all conditions) in the data, such that $\mathbf{y} = (y_{ijl})$ is the $n \times q$ matrix of the DGE for all observations and variables, and $\mathbf{y}_i$ is the q-dimensional vector of DGE for all variables of observation i. We use dot notation to indicate summations in various directions, e.g. $y_{\cdot jl} = \sum_i y_{ijl}$, $y_{i\cdot\cdot} = \sum_j \sum_l y_{ijl}$, and so on.

2.1 Poisson mixture model

To cluster RNA-seq data, we consider a model-based clustering procedure based on a mixture of Poisson distributions. The data $\mathbf{y}$ are assumed to come from K distinct subpopulations (clusters), each of which is modeled separately (McLachlan and Peel, 2000). The overall population is thus distributed under the following mixture:

$$f(\mathbf{y}; K, \boldsymbol{\Psi}_{\mathrm{K}}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i; \boldsymbol{\theta}_{ik}) \qquad (1)$$

where $\boldsymbol{\Psi}_{\mathrm{K}} = (\pi_1, \ldots, \pi_{K-1}, \boldsymbol{\theta}')'$, $\boldsymbol{\theta}'$ contains all of the parameters in $\{\boldsymbol{\theta}_{ik}\}_{i,k}$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$ are the mixing proportions, with $\pi_k \in (0, 1)$ for all k and $\sum_{k=1}^{K} \pi_k = 1$.

Although a multivariate version of the Poisson distribution does exist (Karlis, 2003), it is difficult to implement, particularly for data with high dimensionality. For this reason, the samples are assumed to be independent conditionally on the components:

$$f_k(\mathbf{y}_i; \boldsymbol{\theta}_{ik}) = \prod_{j=1}^{d} \prod_{l=1}^{r_j} \mathcal{P}(y_{ijl}; \mu_{ijlk}),$$

where $\mathcal{P}(\cdot)$ denotes the standard Poisson probability mass function and $\boldsymbol{\theta}_{ik} = \{\mu_{ijlk}\}_{j,l}$. We note that the conditional independence of components could be considered to be a rather strong assumption. However, this assumption appears to be quite reasonable and is often employed to analyze multivariate categorical data; for instance, the latent class model is a reference model in model-based cluster analysis of categorical data (McCutcheon, 1987). When this conditional independence assumption is not expected to hold, in practice it leads to a larger number of clusters and a more complex mixture model that is still able to adequately fit the data. Moreover, attempts to avoid this assumption are definitively inefficient in high dimension settings.

Each mean μ_{ijlk} is parameterized by

$$\mu_{ijlk} = w_i s_{jl} \lambda_{jk} \qquad (2)$$

where $w_i = y_{i\cdot\cdot}$ corresponds to the overall expression level of observation i (e.g. weakly to strongly expressed) as well as a proxy for gene length, and s_{jl} represents the normalized library size for replicate l of condition j, such that $\sum_{j,l} s_{jl} = 1$. These normalization factors take into account the fact that the number of reads expected to map to a particular gene depends not only on its expression level, but also on the library size (overall number of mapped reads) and the overall composition of the RNA population being sampled (Dillies *et al.*, 2012). In particular, larger library sizes result in higher counts for the entire sample. We note that $\{s_{jl}\}_{j,l}$ are estimated from the data prior to fitting the model (see Section 2.3 for more details), and like the overall expression levels w_i, they are subsequently considered to be fixed in the Poisson mixture model. Note that under these conditions, as the marginal sums are fixed for each gene, the model in Equation (1) is in fact a multinomial mixture model. Finally, the unknown parameter vector $\boldsymbol{\lambda}_k = (\lambda_{1k}, \ldots, \lambda_{dk})$ corresponds to the clustering parameters that define the profiles of the genes in cluster k across all biological conditions.

2.2 Inference

To estimate mixture parameters $\mathbf{\Psi}_{\mathrm{K}} = (\boldsymbol{\pi}, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K)$ by computing the maximum likelihood estimate (MLE), an Expectation-Maximization (EM) algorithm is considered (Dempster *et al.*, 1977). The mixture model in Equation (1) is thought of as an incomplete data structure model, with complete data

$$(\mathbf{y}, \mathbf{z}) = ((\mathbf{y}_1, \mathbf{z}_1), \ldots, (\mathbf{y}_n, \mathbf{z}_n))$$

where the missing data are $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n) = (z_{ik}; i = 1, \ldots, n, k = 1, \ldots, K)$, such that $z_{ik} = 1$ if observation i arises from group k and 0 otherwise. The latent variable $\mathbf{z}$ thus defines a partition $P = (P_1, \ldots, P_K)$ of the observed data $\mathbf{y}$ with $P_k = \{i; z_{ik} = 1\}$.

After initializing the parameters $\mathbf{\Psi}_{\mathrm{K}}^{(0)}$ and $\mathbf{z}^{(0)}$ by a so-called small-EM strategy (Biernacki *et al.*, 2003) (see Section 2.3 for more details), the E-step at iteration b corresponds to computing the conditional probability that an observation i arises from the kth component for the current value of the mixture parameters:

$$t_{ik}^{(b)} = \frac{\pi_k^{(b)} f_k(\mathbf{y}_i; \boldsymbol{\theta}_{ik}^{(b)})}{\sum_{m=1}^{K} \pi_m^{(b)} f_m(\mathbf{y}_i; \boldsymbol{\theta}_{im}^{(b)})} \qquad (3)$$

where $\boldsymbol{\theta}_{ik}^{(b)} = \{w_i s_{jl} \lambda_{jk}^{(b)}\}_{jl}$. Then, in the M-step the mixture parameter estimates are updated to maximize the expected value of the completed likelihood, which leads to weighting the observation i for group k with the conditional probability $t_{ik}^{(b)}$. Thus,

$$\pi_k^{(b+1)} = \frac{1}{n}\sum_{i=1}^{n} t_{ik}^{(b)} \text{ and } \lambda_{jk}^{(b+1)} = \frac{\sum_{i=1}^{n} t_{ik}^{(b)} y_{ij\cdot}}{s_{j\cdot} \sum_{i=1}^{n} t_{ik}^{(b)} y_{i\cdot\cdot}},$$

since $w_i = y_{i\cdot\cdot}$. Note that at each iteration of the EM algorithm, we obtain that $\sum_{j=1}^{d} \lambda_{jk}^{(b)} s_{j\cdot} = 1$. Thus $\lambda_{jk}^{(b)} s_{j\cdot}$ can be interpreted as the proportion of reads that are attributed to condition j in cluster k, after accounting for differences due to library size; this proportion is shared among the replicates of condition j according to their respective library sizes s_{jl}.

For model selection (i.e. the choice of the number of clusters K), a reference penalized likelihood criterion with a fixed penalty for mixture models is the Bayesian Information Criterion (BIC) (Schwarz, 1978):

$$\mathrm{BIC}(K) = -\log f(\mathbf{y}; K, \hat{\mathbf{\Psi}}_K) + \frac{\nu_K}{2}\log(n)$$

where $\hat{\mathbf{\Psi}}_K$ are the ML parameter estimates and $\nu_K = (K-1) + K \times d$ is the number of free parameters in the model with K clusters; the use of the BIC is primarily motivated by asymptotic properties that may not hold in practice. An alternative approach to model selection is the use of the so-called *slope heuristics* (Birgé and Massart, 2001, 2006), which is a data-driven method to calibrate a penalized criterion that is known up to a multiplicative constant. Briefly, in our context the penalty is assumed to be proportional to the number of free parameters ν_K (i.e. the model dimension), such that pen(K) $\propto \kappa\nu_K$; we note that this assumption may be verified in practice. The penalty is calibrated using the *data-driven slope estimation* (DDSE) procedure available in the `capushe` R package (Baudry *et al.*, 2012). This procedure directly estimates the slope of the expected linear relationship of the loglikelihood with respect to the model dimension for the most complex models (here, models with large K). Denoting the estimated slope $\hat{\kappa}$, in our context the slope heuristics consists of setting the penalty to be $2\hat{\kappa}\nu_K$. The number of selected clusters $\hat{K}$ then corresponds to the value of K minimizing the penalized criterion:

$$\mathrm{crit}(K) = -\log f(\mathbf{y}; K, \hat{\mathbf{\Psi}}_K) + 2\hat{\kappa}\nu_K.$$

For more details, see Baudry et al. (2012).

Finally, based on $\hat{\mathbf{\Psi}}_{\hat{K}}$, each observation i is assigned to the component maximizing the conditional probability $\hat{t}_{ik}$ (i.e. using the so-called MAP rule).

2.3 `HTSCluster` R package

Our proposed clustering procedure based on a Poisson mixture model is implemented in the R package `HTSCluster`, freely available on CRAN; in this section, we describe some of the options available in this package.

2.3.1 Normalization factors

In the model described in Equation (2), the cluster-specific parameters λ_{jk} are assumed to be shared among replicates within the same condition j; as such, our model assumes that differences in mean expression for a given gene among replicates within the same condition may be explained by differences in library sizes. As described in Section 2.1, these library size normalization factors s_{jl} are estimated from the data and considered to be fixed parameters in the Poisson mixture model. Several options are available in `HTSCluster` to provide estimates for these normalization factors, including the Trimmed Mean of M-values (TMM) normalization (Robinson and Oshlack, 2010) in the `edgeR` Bioconductor package (Robinson *et al.*, 2010) and the median ratio normalization developed in the `DESeq` Bioconductor package (Anders and Huber, 2010). Although Dillies *et al.* (2012) performed an evaluation of the impact of these normalization factors in the context of differential analyses, such a comparison remains an open research question for co-expression analyses.

2.3.2 Parameter estimation and initialization

For parameter estimation in `HTSCluster`, in addition to the EM algorithm described above, it is also possible to use the so-called CEM (Clustering EM) algorithm (Celeux and Govaert, 1992). The CEM algorithm estimates both the mixture parameters and the cluster labels by maximizing the completed likelihood. In the E-step of the algorithm, the conditional probabilities $t_{ik}^{(b)}$ are computed as in Equation (3). In the C-step, the MAP rule is used to assign each observation to a component. Finally, in the M-step, the mixture parameter estimates are updated by maximizing the completed likelihood. Contrary to the EM algorithm, the CEM algorithm converges in a finite number of iterations, although it does provide biased estimates of the mixture parameters [see for instance McLachlan and Peel (2000), Section 2.21]. Because it aims to maximize the completed likelihood, where the component label of each sample point is included in the data, the CEM may be seen as a K-means-like algorithm.

To initialize parameter values for the EM algorithm, the default option is a so-called small-EM strategy (Biernacki *et al.*, 2003), where the following procedure is used to obtain initial parameter values: first, the data are partitioned into K clusters ($\hat{\mathbf{z}}^{(0)}$) using either a K-means algorithm (MacQueen, 1967) or the splitting initialization strategy of Papastamoulis *et al.* (2014), where the cluster from the model with $K-1$ clusters with the largest entropy is chosen

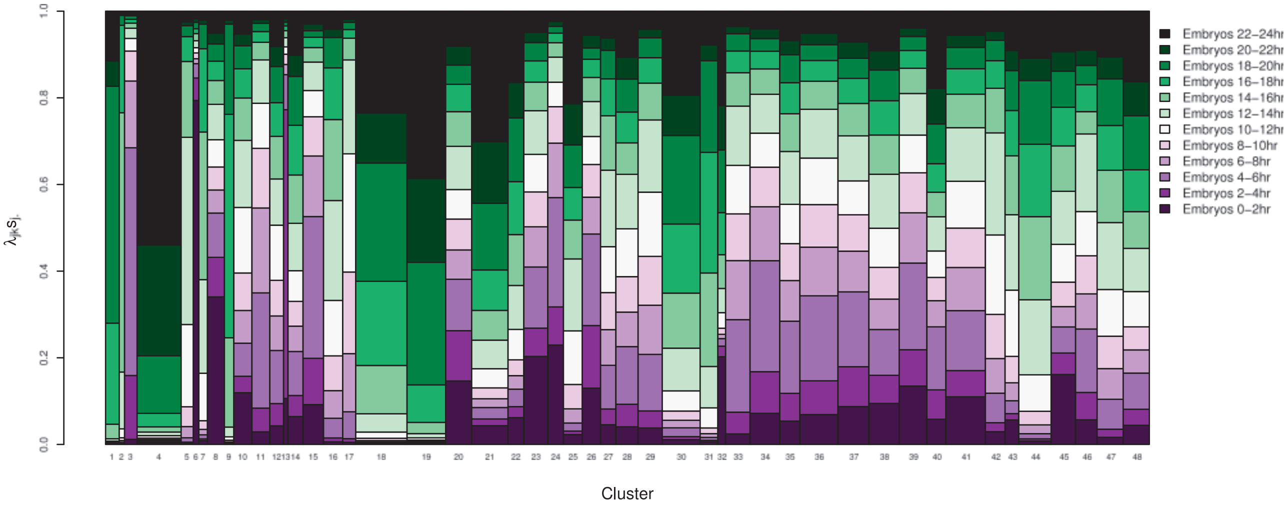

Fig. 1. Visualization of overall cluster behavior for the *Drosophila melanogaster* developmental data. For each cluster, bar plots of $\hat{\lambda}_{jk}s_{j.}$ are drawn for each developmental time point, where the width of each bar corresponds to the estimated proportion $\hat{\pi}_k$

to be split into two new clusters. Second, initial parameter values $\boldsymbol{\pi}^{(0)}$ and $\boldsymbol{\lambda}^{(0)}$ are calculated as follows:

$$\pi_k^{(0)} = \frac{1}{n}\sum_{i=1}^{n}\hat{z}_{ik}^{(0)} \text{ and } \lambda_{jk}^{(0)} = \frac{\sum_i y_{ij\cdot}\hat{z}_{ik}^{(0)}}{s_{j\cdot}\sum_i w_i\hat{z}_{ik}^{(0)}}.$$

Third, 10 iterations of the EM algorithm are run. Finally, the parameter estimates $\hat{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\pi}}$ are used initialize the subsequent full EM algorithm.

2.3.3 Additional options

Finally, HTSCluster provides flexibility to the user through a variety of graphical representations as well as a set of additional options, including the following: (i) cluster proportions $\boldsymbol{\pi}$ can be variable (the default option) or fixed to be equal for all clusters; and (ii) one or more clusters may be included with a fixed value for $\boldsymbol{\lambda}_k$. The latter option may be particularly useful in the context of differential analyses, where a group of genes may be assumed to have identical expression across experimental conditions. We recently proposed an approach and an associated R package HTSDiff that make use of this particular functionality.

3 Results

3.1 Real data analysis

In the following, we illustrate the use of the HTSCluster package for a co-expression analysis of two real RNA-seq datasets. We stress that in both cases, it is not possible to compare the co-expression results obtained using HTSCluster to a 'true' clustering of the data, as such a classification does not generally exist. However, in order to identify whether co-expressed genes appear to be implicated in similar biological processes, we conduct functional enrichment analyses of gene ontology (GO) terms for the clusters identified by HTSCluster.

3.1.1 Dynamic expression of the transcriptome in embryonic flies

As part of the modENCODE project, which aims to provide functional annotation of the *Drosophila melanogaster* genome, Graveley *et al.* (2011) characterized the expression dynamics over 27 distinct stages of development during the life cycle of the fly using RNA-seq. In this work, we focus on a subset of these data from 12 embryonic samples that were collected at 2-h intervals for 24 h, with one biological replicate for each time point. The phenotype tables and raw read counts for the 13 164 genes with at least one non-zero count among the 12 time points were obtained from the ReCount online resource (Frazee *et al.*, 2011).

Over three independent runs, we used the HTSCluster package with default settings and the splitting small-EM initialization strategy (described in Section 2.3.2) to fit a sequence of Poisson mixture models with $K = 1, \ldots, 60$ clusters; for each number of clusters, the model corresponding to the largest loglikelihood among the three runs was retained. To ensure that the collection of models considered is large enough to apply the slope heuristics model selection, one additional set of Poisson mixture models was fit for $K = 65, \ldots, 95$ (in steps of 5) and $K = 100, \ldots, 130$ (in steps of 10). Using the slope heuristics, the number of clusters was determined to be $\hat{K} = 48$; see the Supplementary Materials for more detail.

Visualizing the results of a co-expression analysis for RNA-seq data can be somewhat complicated by the extremely large dynamic range of DGE and the fact that more highly expressed genes tend to exhibit greater variability (though much smaller coefficients of variation) than weakly expressed genes. Two possibilities to avoid this issue are to apply a logarithmic transformation to obtain pseudo-counts (Robinson *et al.*, 2010) or a variance-stabilizing transformation (Anders and Huber, 2010) prior to graphical representation; however, the choice of the data to be visualized (e.g. raw, scaled, or transformed data) as well as the most appropriate manner in which they should be graphically displayed are still an open matter of research. For the purposes of co-expression, rather than directly visualizing the data themselves, we propose an alternative visualization of the overall behavior of each cluster, as shown in Figure 1. In this plot, bar widths correspond to the estimated proportion of genes in each cluster ($\hat{\pi}_k$), and the proportion of reads attributed to each developmental time point in each cluster $\hat{\lambda}_{jk}s_{j.}$ are represented by the colored segments within each bar. The advantage of such a visualization is that it enables a straightforward comparison of typical gene profiles among clusters. For instance, it can be seen that clusters characterized by higher relative expression in the early embryonic stages, such as Clusters 6 and 13 (composed of 70 and

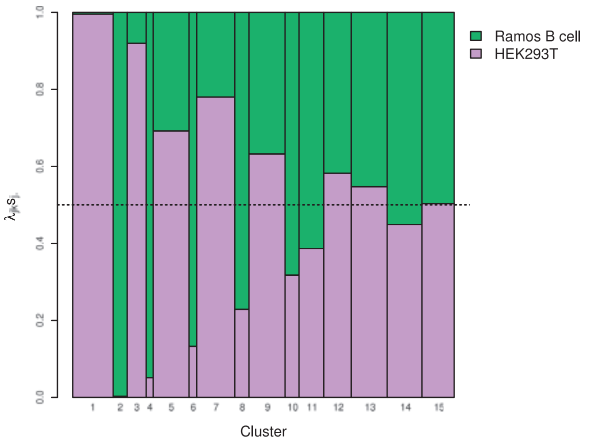

Fig. 2. Visualization of overall cluster behavior for the human liver RNA-seq data. For each cluster, bar plots of $\hat{\lambda}_{jk}s_{j.}$ are drawn for each experimental condition, where the width of each bar corresponds to the estimated proportion $\hat{\pi}_k$

60 genes, respectively) tend to be much smaller than those with higher relative expression in later stages, e.g. Clusters 4, 18, 19 and 21 (composed of 567, 680, 485 and 475 genes, respectively).

A functional enrichment analysis of GO biological processes revealed that of the 48 clusters identified by `HTSCluster`, 33 were associated with at least one GO term. For example, cluster 39 was found to be associated with terms pertaining to morphogenesis (GO:0009653, GO:0048858) and cell development (GO:0048468), while cluster 6 is associated with muscle attachment (GO:0016203). As a comparison, we also fit the closely related Poisson and negative binomial mixture models proposed by Si *et al.* (2014) for $K = 48$ clusters; for these models, a total of 22 and 25 clusters, respectively, were associated with at least one GO term. Additional details may be found in the Supplementary Materials.

3.1.2 Sex-specific expression using RNA-seq in human liver cells

High-throughput transcriptome RNA-seq data were obtained (Sultan *et al.*, 2008) from a human embryonic kidney (HEK293T) and a Ramos B cell line, with two biological replicates in each experimental condition. The raw read counts for 9010 genes and phenotype tables were obtained from the ReCount online resource (Frazee *et al.*, 2011). Following filtering using the HTSFilter package (Rau *et al.*, 2013) to remove weakly expressed genes across the two conditions, 4956 were retained for the subsequent coexpression analysis.

As for the previous dataset, we applied the `HTSCluster` package with default parameters and the splitting small-EM initialization strategy over three independent runs to fit a sequence of Poisson mixture models with $K = 1, \ldots, 50$, where only the model with the highest loglikelihood for each number of clusters was retained. One additional set of models was fit for $K = 55, \ldots, 75$ (in steps of 5) to ensure the applicability of the slope heuristics for model selection. Using the slope heuristics, the number of clusters was determined to be $\hat{K} = 15$; additional details may be found in the Supplementary Materials.

As before, a visualization of the overall behavior of genes belonging to each cluster may be found in Figure 2. It can be noted that Cluster 1 represents a fairly large number of genes (540) with dominant expression in the HEK293T cell line, whereas the much smaller Cluster 2 (192 genes) represents those genes primarily expressed in the Ramos B cell line; on the other hand, Clusters 12–15 represent groups of genes (212, 546, 511 and 365 genes, respectively) with largely balanced expression in the two cell lines. We note that such a visualization may be useful as a complement to a full differential expression analysis, as it enables a global characterization of the differences that are present between the two conditions.

A functional enrichment analysis of GO biological processes revealed that, of the 15 clusters identified by `HTSCluster`, 11 were associated with at least one GO term. For example, Cluster 4 is associated with terms related to morphogenesis (GO:0048644, GO:0055008) and mesenchyme tissue development (GO:0060485), while Cluster 10 is associated with the negative regulation of action potential, nitric oxide synthase activity, oxidoreductase activity and leukocyte activity (GO:0045759, GO:0051001, GO:0051354, GO:0002695). In comparison, the closely related Poisson and negative binomial mixture models proposed by Si *et al.* (2014) were also estimated for $K = 15$ clusters; for these models, a total of 10 and 5 clusters, respectively, were associated with at least one GO term. Additional details may be found in the Supplementary Materials.

3.2 Simulation study

In this section, we perform a set of simulation experiments in order to compare the performance of the proposed Poisson mixture model to that of several alternative related approaches, described below.

3.2.1 Description of alternative clustering approaches

- **PoisL**: Originally proposed for SAGE data, the PoisL approach (Cai *et al.*, 2004) assumes that, given the cluster k, genes follow a Poisson distribution with mean $\mu_{ijk} = w_i\lambda_{jk}$, under the constraint that $\sum_j \lambda_{jk} = 1$ for all k; the existence of replicates within each condition is not taken into account in the original method. Using this model, a K-means algorithm is proposed, where each gene i is assigned to the cluster k at iteration b if $k = \text{argmin}_{k'} - \log f_{k'}(\mathbf{y}_i; \{\mu_{ijk}^{(b)}\}_j)$. This procedure is exactly equivalent to the Poisson mixture model implemented in `HTSCluster` with equiprobable Poisson mixtures (i.e. $\pi_k = \pi$ for all k), parameter estimation via the CEM algorithm, and unreplicated data. For comparison with the other methods described here, we also include normalization factors s_{jl} in the model as replicates are present.
- **Witten**: Witten (2011) recently considered the issue of clustering samples, rather than genes, using RNA-seq data. After fitting a Poisson loglinear model to the power-transformed data (Li *et al.*, 2012), complete linkage hierarchical clustering is applied to the dissimilarity matrix calculated using a modified loglikelihood ratio statistic to compare Poisson distributions. Although this method was originally proposed to cluster samples, Witten (2011) claims that it may also be used to cluster genes in RNA-seq data. This procedure is available in the R package PoiClaClu.
- **Si-Pois and Si-NB**: Si *et al.* (2014) consider Poisson mixture models (Si-Pois) where $\log(\mu_{ijlk}) = a_{ijl} + \alpha_i + \beta_{jk}$ with the constraint $\sum_j \beta_{jk} = 0$ for all k, where a_{ijl} is a normalization factor that simultaneously accounts for the length of gene i and the library size of replicate l in condition j. An EM algorithm and two stochastic versions are proposed to estimate the remaining parameters. Following parameter estimation, a model-based hybrid-hierarchical clustering algorithm is developed to build a hierarchical tree. In addition, Si *et al.* (2014) also consider negative binomial mixture models (Si-NB) parameterized by the same mean as the Si-Pois method described above. A per-gene dispersion parameter is estimated by quasi-likelihood prior to fitting

Table 1. Mean (SD) ARI for simulations with parameters based on the fly and human liver

Method	Model selection	Fly	Human
`HTSCluster`	`capushe`	**0.93** (0.05)	**0.61** (0.02)
	True *K*	0.84 (0.09)	**0.60** (0.02)
PoisL	`capushe`	0.79 (0.15)	0.53 (0.05)
	True *K*	0.82 (0.05)	0.53 (0.04)
Witten	CH index	0.15 (0.07)	0.11 (0.03)
	True *K*	0.67 (0.09)	0.39 (0.04)
Si-Pois	CH index	0.26 (0.17)	0.48 (0.04)
	True *K*	**0.95** (0.02)	**0.61** (0.02)
Si-NB	CH index	0.23 (0.16)	0.47 (0.04)
	True *K*	**0.94** (0.02)	**0.60** (0.02)
K-means	True *K*	0.79 (0.08)	0.42 (0.02)
Oracle	True *K*	**0.95** (0.01)	**0.63** (0.01)

Note: The largest values in each simulation setting are highlighted in bold font.

Table 2. Mean (SD) estimated number of clusters determined by the slope heuristics (`HTSCluster` and PoisL) or the CH index (Witten, Si-Pois, Si-NB) for simulations with parameters based on the fly and human liver data

	`HTSCluster`	PoisL	Witten	Si-Pois	Si-NB
Criterion	`capushe`	`capushe`	CH	CH	CH
Fly	19.9 (5.3)	14.1 (5.1)	2.3 (0.6)	3.9 (2.4)	3.3 (1.9)
Human	21.2 (5.5)	15.9 (4.0)	2.8 (0.4)	8.4 (2.1)	8.4 (2.4)

Note: The true number of clusters for all simulations was fixed to $K = 15$.

the model and considered to be fixed. The remainder of the Si-NB procedure is similar to Si-Pois, and both may be implemented using the R package `MBCluster.Seq`.

- **K-means:** As a comparison, we also consider a classic K-means algorithm (MacQueen, 1967) with the usual Euclidian distance, which is applied to the gene expression profiles $\tilde{\mathbf{y}} = (\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_n)$ such that $\tilde{\mathbf{y}}_i = \{y_{ijl}/w_i\}_{j,l}$.

We note that model selection is not addressed by any of the methods described above; for this purpose, we use the index proposed by Caliński and Harabasz (1974) (Witten, Si-Pois, and Si-NB), which is a pseudo F-statistic that compares between and within-cluster dispersion, or the slope heuristics (PoisL) to select the appropriate number of clusters and deduce a data clustering. Although the slope heuristics could potentially be used for the Si-Pois and Si-NB methods, the output provided by their implementation in the `MBCluster.Seq` package renders this calculation difficult in practice. In addition, we note that all methods (with the exception of the Si-NB and classic K-means) are based on the use of an underlying Poisson model.

3.2.2 Simulation strategy

Using the parameter estimates obtained by `HTSCluster` in the fly and human liver RNA-seq datasets (described in Section 3.1), we simulated 50 datasets for each setting under a Poisson mixture model as in Equation (2) in the following manner.

For the simulations based on each real dataset (fly or human liver), the numbers of conditions and replicates per condition were fixed to be equivalent to the experimental design of each real dataset. The number of clusters K was fixed to be equivalent to 15; for the human liver data, this corresponds to the model selected via the slope heuristics, while for the fly data, 15 clusters were randomly chosen among the 48 estimated in the selected model. In addition, normalization factors s_{jl}, cluster proportions π_k (renormalized to sum to 1, in the case of the fly data), and cluster parameters λ_{jk} were fixed to be their estimated values for each dataset, and overall expression levels w_i were fixed to be equal to the observed $y_{i\cdot\cdot}$ values. A total of $n = 3000$ genes were randomly sampled from the fly or human liver data, weighted by their maximum conditional probability from the selected `HTSCluster` model. For each sampled gene i, we sampled from the appropriate Poisson distribution $Y_{ijl} \sim \mathcal{P}(\mu_{ijlk})$, where $\mu_{ijlk} = w_i s_{jl} \lambda_{jk}$ if $\hat{z}_{ik} = 1$. In all datasets, we verified that simulated data were indeed represented by $K = 15$ clusters.

3.2.3 Results

For each simulated dataset in the two settings (fly and human liver), `HTSCluster` and the methods described in Section 3.2.1 were fit over a range of possible numbers of clusters ($K = 1, \ldots, 40$), with model selection performed using the slope heuristics (`HTSCluster`, PoisL) or the CH index (Witten, Si-NB, Si-Pois). In addition, for all competing approaches, the model with the true number of clusters ($K = 15$) was also included. Models were subsequently compared using the adjusted Rand index (ARI) (Hubert and Arabie, 1985) and the estimated number of clusters $\hat{K}$, shown in Tables 1 and 2, respectively. The oracle ARI is also included for comparison, based on the assignment of observations to components maximizing the conditional probability using the true parameter values in the Poisson mixture model.

In both simulation settings considered here, we note that a major difficulty for the alternative methods is the choice of the number of clusters to be included; the Witten, Si-Pois, and Si-NB methods all exhibit significantly lower ARI values when model selection is performed using the CH index as compared to when the true number of clusters is fixed. This difficulty appears to be especially pronounced for the Witten approach, where ARI values for the model selected using the CH index is less than 0.2 in both simulation settings and the selected number of clusters is significantly underestimated. For the Si-Pois and Si-NB methods, the CH index also tends to underestimate the number of clusters present in the data, although this trend is less marked than for the Witten approach, particularly in the human simulated data. In the case of the PoisL approach, although the slope heuristics approach appears to generally yield an appropriate estimate of K in both simulation settings, the corresponding ARI values tend to be lower than those attained by the `HTSCluster` approach.

Even when the number of clusters is fixed to the true value, the competiting methods tend to have equivalent or smaller ARI values than the models selected via slope heuristics for the proposed `HTSCluster` approach, in spite of the fact that all approaches (with the exception of Si-NB and K-means) also make use of an underlying Poisson model. In other words, if the true number of clusters is known, the performance of the Si-Pois and Si-NB approaches is quite good and very nearly attains that of the oracle model; however, when the number of clusters must be estimated from the data (as is typically the case in real applications), `HTSCluster` has much stronger performance than the competing methods on these simulated data. Model selection via the slope heuristics for the `HTSCluster` approach leads to a slight overestimation of the number of clusters, but these slightly more complex models have ARI values close to those found using the oracle Poisson mixture model. Finally, we note that the clustering task in the human liver setting (where only two conditions are present) appears to be much more difficult for all methods considered here than in the fly setting (where 12 unreplicated conditions are present), as evidenced by the smaller oracle ARI value. This is perhaps unsurprising, as it is more difficult to discern differing cluster profiles for only two conditions than when multiple conditions are available; this can be seen in the overall cluster behavior in the two real data analyses (Figs 1 and 2).

4 Discussion

In this work, we have proposed a method and associated R package HTSCluster to cluster count-based DGE profiles based on a Poisson mixture model that enables the use of a rigorous framework for parameter estimation (through the EM algorithm) and model selection (through the slope heuristics). The model is parameterized to account for several characteristics of RNA-seq data, including: (i) a set of normalization factors (s_{jl}) to account for systematic differences in library size among biological replicates, (ii) a per-gene offset parameter (w_i) to account for differences among genes due to overall expression level and (iii) a condition-specific cluster effect (λ_{jk}). As the marginal sums of each gene are fixed in the model, variations in expression among experimental conditions may be modeled throughout the extremely large dynamic range of DGE typical of RNA-seq data. In particular, this parameterization enables a straightforward interpretation of the model, as $\lambda_{jk}s_{j\cdot}$ corresponds to the proportion of reads attributed to condition j in cluster k. A co-expression analysis on two sets of real RNA-seq data highlighted the functionality of HTSCluster in practice, in particular with respect to model selection and visualization of overall cluster behavior. Finally, the processing time and memory requirements of HTSCluster reflect the fact that parameter estimation must be performed over a large set of models to enable model selection; one run of HTSCluster (version 2.0.4) took about 50 minutes and used about 450 MB of memory for the human liver data ($K = 1, \ldots, 50$), and about 2 h with 1800 MB of memory for the fly developmental data ($K = 1, \ldots, 60$). (All analyses were run on a Dell Latitude E6530 quad-core 2.70 GHz Intel(R) Core(TM) with **10GB** RAM, running a 64-bit version of Windows 7 Professional.)

As previously mentioned, HTSCluster shares some similarities with other related approaches, although there are several key differences. First, we note that both PoisL (Cai *et al.*, 2004) and Witten (2011) also make use of an underlying Poisson model; however, rather than using a finite mixture model, the former uses a K-means algorithm based on the loglikelihood and the latter applies a hierarchical clustering procedure based on a pairwise dissimilarity matrix of dimension ($n \times n$). On the other hand, Si *et al.* (2014) suggest the construction of a hierarchical tree of either Poisson (Si-Pois) or negative binomial (Si-NB) mixture models with an alternative parameterization to that proposed here. Contrary to all of these alternative related approaches, the HTSCluster approach provides a straightforward and robust way to choose the number of clusters present in a given dataset.

A set of simulation studies, with parameters selected based on two real datasets, allowed a comparison of HTSCluster with the aforementioned related approaches in a controlled scenario. These simulations highlighted the importance of an appropriate procedure to perform model selection, as well as the satisfactory performance of HTSCluster in the objective of clustering and estimating the number of clusters. In addition, even when the number of clusters was fixed to the true value, we found that the alternative methods were generally observed to have similar or lower ARI values than HTSCluster. However, conclusions from these simulations should be drawn with some caution, particularly as the data were simulated based on a mixture of Poisson distributions. A great deal of discussion has focused on the most appropriate way to simulate RNA-seq data in the context of differential expression (Soneson and Delorenzi, 2013), and for the time being this remains an open question for co-expression analyses.

Finally, we note that in the context of differential expression analyses, the scientific community has generally focused on the use of negative binomial models due to the large variability typically observed among replicates for a fixed gene. This so-called overdispersion is modeled via the inclusion of a common dispersion parameter ϕ or a per-gene dispersion parameter ϕ_i, typically estimated using a shrinkage approach (Robinson and Smyth, 2007) or a parametric regression fit across all genes (Anders and Huber, 2010). The Si-NB approach (Si *et al.*, 2014) recently attempted to apply a similar approach to the task of co-expression analysis through a finite mixture of negative binomial models, where ϕ_i is estimated from the data using a quasi-likelihood approach and treated as fixed in the mixture. However, for co-expression analyses it is difficult to estimate these per-gene dispersion parameters in practice due to the small number of replicates typically available in experiments concerning multiple conditions. A useful direction for future research may be to define a mixture of negative binomial models in which information about this dispersion parameter is shared among genes belonging to the same cluster.

Acknowledgements

The authors thank the members of the Statistics for Systems Biology (SSB) working group for their helpful and insightful comments, as well as the three anonymous reviewers for their helpful comments. The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grant MixStatSeq (ANR-13-JS01-0001-01).

Funding

A.R. was funded as a postdoctoral researcher at Inria Saclay - Île-de-France for a portion of this work.

Conflict of Interest: none declared.

References

Anders,S. and Huber,W. (2010). Differential expression analysis for sequence count data. *Genome Biol.,* **11**, 1–28.

Auer,P.L. and Doerge,R.W. (2010). Statistical design and analysis of RNA-Seq data. *Genetics,* **185**, 1–12.

Auer,P.L. *et al.* (2012). Differential expression–the next generation and beyond. *Brief. Funct. Genomics,* **11**, 57–62.

Baudry,J.-P. *et al.* (2012). Slope heuristics: overview and implementation. *Stat. Comp.*, **22**, 455–470.

Biernacki,C. *et al.* (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Comp. Stat. Data Anal.*, **41**, 561–575.

Birgé,L. and Massart,P. (2001). Gaussian model selection. *J. Eur. Math. Soc.*, **3**, 203–268.

Birgé,L. and Massart,P. (2006). Minimal penalties for Gaussian model selection. *Probab. Thoery Relat. Fields,* **138**, 33–73.

Cai,L. *et al.* (2004). Clustering analysis of SAGE data using a Poisson approach. *Genome Biol.,* **5**, R51.

Caliński,T. and Harabasz,J. (1974). A dendrite method for cluster analysis. *Commun. Stat. Theory Methods,* **3**, 1–27.

Celeux,G. and Govaert,G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Comp. Stat. Data Anal.*, **14**, 315–332.

Dempster,A. P. *et al.* (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc. Series B (Methodological)*, **39**, 1–38.

Dillies,M.-A. *et al.* (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.*, **14**, 671–83.

Eisen,M. B. *et al.* (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS,* **95**, 14863–14868.

Frazee,A.C. *et al.* (2011). ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics,* **12**.

Graveley,B.R. *et al.* (2011). The development transcriptome of *Drosophila melanogaster*. *Nature,* **471**, 473–479.

Hubert,L. and Arabie,P. (1985). Comparing partitions. *J. Classif.,* **2**, 193–218.

Jiang,D. *et al.* (2004). Cluster analysis for gene expression data: a survey. *IEEE Trans. Knowl. Data Eng.,* **16**, 1370–1386.

Karlis,D. (2003). An EM algorithm for multivariate Poisson distribution and related models. *J. Appl. Stat.*, **30**, 63–77.

Łabaj,P. P. *et al.* (2011). Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics*, **27**(ISMB), i383–i391.

Law,C. *et al.* (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.,* **15**.

Li,J. *et al.* (2012). Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics,* **13**, 523–538.

MacQueen,J.B. (1967). Some methods for classification and analysis of multivariate observations. In: *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability,* p. 281–297. University of California Press, Berkeley.

McCarthy,D. *et al.* (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.,* **40**, 4288–4297.

McCutcheon,A.C. (1987). *Latent Class Analysis.* Sage Publications, Beverly Hills.

McIntyre,L.M. *et al.* (2011). RNA-seq: technical variability and sampling. *BMC Genomics,* **12**.

McLachlan,G. *et al.* (2004). *Analyzing Microarray Gene Expression Data.* Wiley-Interscience, Hoboken.

McLachlan,G. and Peel,D. (2000). *Finite Mixture Models.* Wiley-Interscience, New York.

Naghavachari,N. *et al.* (2012). A systematic comparison and evaluation of high density exon arrays and RNA-seq technology used to unravel the peripheral blood transcriptome of sickle cell disease. *BMC Medical Genomics,* **5**,28.

Oshlack,A. and Wakefield,M.J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct,* **4**.

Papastamoulis, P. *et al.* (2014). On the estimation of mixtures of Poisson regression models with large numbers of components. *Comp. Stat. Data Anal.*, doi:10.1016/j.csda.2014.07.005.

Rau, A. *et al.* (2013). Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics*, **29**, 2146–2152.

Robinson, M. D. *et al.* (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.,* **11**,R25.

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

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.,* **6**, 461–464.

SEQC/MAQC-III Consortium (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.*, **32**, 903–914.

Severin, A. J. *et al.* (2010). RNA-Seq Atlas of *Glycine max*: a guide to the soybean transcriptome. *BMC Plant Biol.,* **10**,160.

Si, Y. *et al.* (2014). Model-based clustering for RNA-seq data. *Bioinformatics*, **30**, 197–205.

Soneson, C. and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, **14**,91.

Sultan, M. *et al.* (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science,* **15**, 956–60.

Trapnell, C. *et al.* (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.,* **28**, 511–518.

Wang, C. *et al.* (2014). The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nat. Biotechnol.,* **32**, 926–932.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.,* **58**, 236–244.

Witten, D. M. (2011). Classification and clustering of sequencing data using a Poisson model. *Ann. Appl. Stat.,* **5**, 2493–2518.

Yeung, K. Y. *et al.* (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977–987.

Zhou, X. *et al.* (2014). Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Res.,* **42**, e91.