Transformation and model choice for RNA-seq co-expression analysis

Andrea Rau[1]* and Cathy Maugis-Rabusseau[2]

[1]GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France

[2]Institut de Mathématiques de Toulouse, INSA de Toulouse, Université de Toulouse, 31400 Toulouse, France

*To whom correspondence should be addressed; E-mail: andrea.rau@jouy.inra.fr.

Abstract

Although a large number of clustering algorithms have been proposed to identify groups of co-expressed genes from microarray data, the question of if and how such methods may be applied to RNA-seq data remains unaddressed. In this work, we investigate the use of data transformations in conjunction with Gaussian mixture models for RNA-seq co-expression analyses, as well as a penalized model selection criterion to select both an appropriate transformation and number of clusters present in the data. This approach has the advantage of accounting for per-cluster correlation structures among samples, which can be quite strong in real RNA-seq data. In addition, it provides a rigorous statistical framework for parameter estimation, an objective assessment of data transformations and number of clusters, and the possibility of performing diagnostic checks on the quality and homogeneity of the identified clusters. We analyze four varied RNA-seq datasets to illustrate the use of transformations and model selection in conjunction with Gaussian mixture models. Finally, we propose an R package `coseq` (**co**-expression of RNA-**seq** data) to facilitate implementation and visualization of the recommended RNA-seq co-expression analyses.

Keywords: RNA-seq, co-expression, mixture models, data transformation

Introduction

Increasingly complex studies of transcriptome dynamics are now routinely carried out using high-throughput sequencing of RNA molecules, called RNA sequencing (RNA-seq). By quantifying and comparing transcriptomes among different types of tissues, developmental stages, or experimental conditions, researchers have gained a deeper understanding of how changes in transcriptional activity reflect specific cell types and contribute to phenotypic differences. Identifying groups of co-expressed genes may help target gene modules that are involved in similar biological processes [1, 2] or that are candidates for co-regulation. Thus, by identifying clusters of co-expressed genes, we aim both to identify co-regulated genes and to characterize potential biological functions for orphan genes (namely, those whose biological function is unknown).

A great deal of clustering algorithms have been proposed for microarray data, raising the question of their applicability to RNA-seq data. In particular, after normalization, background correction, and $\log_2$-transformation of microarray data, hybridization intensities are typically modeled by Gaussian distributions [3]. RNA-seq data, on the other hand, are made up of read counts [4, 5] or pseudocounts [6, 7] for each biological entity or feature (e.g., a gene) after either alignment to a genome reference sequence or *de novo* asssembly. These data are characterized by 1) highly skewed values with a very large dynamic range, often covering several orders of magnitude; 2) positive correlation between feature size (e.g., gene length) and read counts [8]; and 3) variable sequencing depth (i.e., library size) and coverage among experiments [9]. The presence of overdispersion (i.e., variance larger than the mean) among biological replicates for a given feature is also a typical feature of these data, leading to the use of negative binomial models [10, 11] for RNA-seq differential analyses.

Statistically speaking, the goal of clustering approaches is to discover structures (clusters) within data. Many clustering methods exist and roughly fall into two categories: 1) methods based on dissimilarity distances, including tree-based hierarchical clustering [12] as well as methods like the K-means algorithm [13]; and 2) model-based methods [14], which consist of defining a clustering model and optimizing the fit between the data and the model. For the latter class of models, each cluster is represented by a distinct parametric distribution, and the entire dataset is thus modeled as a mixture of these distributions; a notable advantage of model-based clustering is that it provides a rigourous framework to assess the appropriate number of clusters and the quality of clusters obtained. Presently, most proposals for clustering RNA-seq data have focused on the question of grouping biological samples rather than features, for example using hierarchical clustering with a modified loglikelihood ratio statistic based on a Poisson loglinear model as a distance measure [15] or the Euclidean distance of samples following a variance-stabilizing transformation [16].

In recent work [17], we proposed the use of Poisson mixture models to cluster RNA-seq expression profiles. This method has the advantage of directly modeling the count nature of RNA-seq data, accounting for variable library sizes among experiments, and providing easily interpretable clusterings based on the profiles of variation around average expression of each gene. However, there are several serious limitations to this approach: 1) the assumption of conditional independence among samples, given the clustering group, is likely to be unrealistic for the vast majority of RNA-seq datasets; 2) per-cluster correlation structures cannot be included in the model; and 3) the Poisson distribution is likely overly restrictive, as it imposes an assumption of equal means and variances. In addition, classical asymptotic model selection criteria, such as the Bayesian Information Criterion (BIC) [18] and Integrated Completed

Likelihood (ICL) criterion [19], were observed to have poor behavior for the Poisson mixture model in many cases. As such, Rau et al. [17] proposed the use of a non-asymptotic penalized model selection criterion calibrated by the slope heuristics [20, 21], requiring a collection of mixture models to be fit for a very wide range of cluster numbers K; for large K, this can imply significant computational time as well as practical difficulties for parameter initalization and estimation. We note that a related approach based on a hybrid-hierarchical clustering of negative binomial mixtures was proposed by Si et al. [22]; as with the work of Rau et al. [17], this method cannot account for correlation structures among samples.

To address the aforementioned limitations of the Poisson mixture model, in this work we investigate appropriate transformations to facilitate the use of Gaussian mixture models for RNA-seq co-expression analysis. This strategy has the notable advantage of enabling the estimation of per-cluster correlation structures, as well as drawing on the extensive theoretical justifications of Gaussian mixture models [14]. We note that Law *et al.* [23] employed a related strategy for the differential analyses of RNA-seq data by transforming data, estimating precision weights for each feature, and using the `limma` empirical Bayes analysis pipeline [24]. The identification of an “appropriate" transformation for RNA-seq co-expression is not necessarily straightforward, and depends strongly on the desired interpretability of the resulting clusters as well as the model assumptions. Several transformations of read counts or pseudocounts have been proposed in the context of exploratory or differential analyses, but most largely seek to render the data homoskedastic and do not facilitate clustering together features with similar *patterns* of expression across experiments. In order to retain the latter interpretation of clusters, rather than grouping together genes with similar absolute (transformed) read abundances, we propose the use of normalized expression *profiles* for each feature, that is, the proportion of normalized counts observed for a given feature. Due to the compositional nature of these profiles (i.e., the sum for each feature equals 1), an additional transformation is required prior to fitting the Gaussian mixture model, as discussed below.

The remainder of the article is organized as follows. In the Methods section, we introduce some notation, discuss appropriate data transformation for RNA-seq co-expression analyses, and briefly review Gaussian mixture models, including parameter estimation and model selection. In the Results section, we describe several RNA-seq datasets and illustrate co-expression analyses on each using Gaussian mixture models on transformed data using the `coseq` R package. Finally, in the Discussion we provide some concluding remarks and recommendations for RNA-seq co-expression analyses in practice, as well as some opportunities for future work.

Methods

For the remainder of this work, let Y_{ij} be a random variable, with corresponding observed value y_{ij}, representing the raw read count (or pseudocount) for biological entity i ($i = 1, \dots, n$) of biological sample j ($j = 1, \dots, q$). For simplicity, in this work we typically refer to the entities i as genes, although the generality of the following discussion holds for other entities of interest (exons, etc). Each sample is typically associated with one or more experimental conditions (e.g., tissue, treatment, time); to reflect this, let $\mathcal{C}(j)$ correspond to the experimental group for sample j. Finally, let $\mathbf{y}$ be the $(n \times q)$ matrix of read counts for all genes and samples, and let $\mathbf{y}_i$ be the q-dimensional vector of raw count values across all biological samples for gene i. In the following, we use dot notation to indicate summations over a particular index, e.g. $y_{i\cdot} = \sum_j y_{ij}$.

Data transformations for RNA-seq co-expression

A feature common to many RNA-seq data transformations is the incorporation of sample-specific normalization factors, often referred to as *library size* normalization. These normalization factors account for the fact that the number of reads expected to map to a particular gene depends not only on its own expression level, but also 1) on the total number of mapped reads (also referred to as library size) in the sample, and 2) on the overall composition of the RNA population being sampled. Although several library size normalization factors have been proposed since the introduction of RNA-seq, the median ratio [11] and trimmed mean of M-values [TMM; 25] methods have been found to be robust and effective, and are now widely used [26] in the context of differential analysis. Without loss of generality, we note $\mathbf{t} = (t_j)$ as the scaling normalization factors for raw library sizes calculated using the TMM normalization method; $\ell_j = y_{\cdot j} t_j$ is then the corresponding normalized library size for sample j, and

$$s_j = \frac{\ell_j}{\sum_{j=1}^{q} \ell_j / q}$$

is the associated scaling factor with which raw counts y_{ij} may be normalized.

Several data transformations have been suggested for RNA-seq data, most often in the context of exploratory or differential analyses. These include a $\log_2$ transformation (where a small constant is typically added to read counts to avoid 0's), a variance-stabilizing transformation [27, 28, 16], moderated log counts per million [CPM; 23], and a regularized log-transformation [11]; see the Supplementary Materials for more details about each. As previously noted, each of these transformations seeks to render the data homoskedastic but does not facilitate clustering together features with similar patterns of expression across experiments. As such, rather than making use of these transformations, we propose calculating the normalized expression *profiles* for each feature, that is, the proportion of normalized reads observed for gene i with respect to the total observed for gene i across all samples:

$$p_{ij} = \frac{y_{ij}/s_j}{\sum_j y_{ij}/s_j},$$

where s_j are the scaling normalization factors for raw library sizes as indicated above. To illustrate the interest of using these normalized expression profiles for co-expression analysis, we plot y_{ij}/s_j, $\log(y_{ij}/s_j + 1)$, and the normalized expression profiles profiles p_{ij} in Figure 1 for a subset of genes from the RNA-seq data studied by Graveley et al. [29] (corresponding to

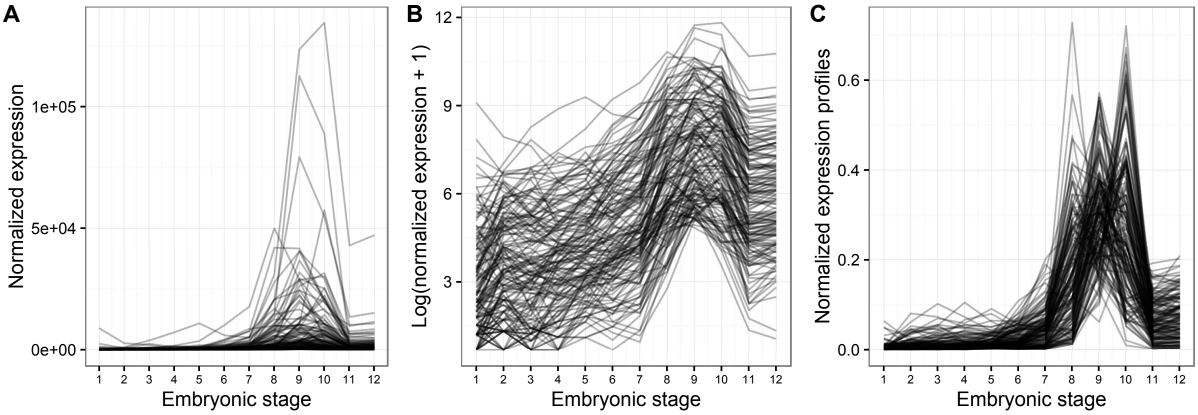

Figure 1: Normalized counts (A), log normalized counts + 1 (B), and normalized expression profiles y_{ij} for a subset of the Graveley et al. [29] fly RNA-seq data, corresponding to the genes assigned to Cluster 1 (see the Results section).

Cluster 1; see the Results section for a complete description of these data). It may clearly be seen that although large differences in magnitude among genes are dominant for normalized counts (Figure 1A), a log-transformation reveals a similarity in expression profiles (Figure 1B) that becomes even more apparent when considering the normalized expression profiles p_{ij} (Figure 1C).

It is important to note that the profile for gene i, $\mathbf{p}_i = (p_{ij})$, represents compositional data [30], as it is a q-tuple of nonnegative numbers whose sum is 1. This means that the vector of values $\mathbf{p}_i$ are linearly dependent, which imposes constraints on the covariance matrices Σ_k that are problematic for the general Gaussian mixture model (and indeed for most standard statistical approaches). For this reason, we consider two separate transformations of the profiles p_{ij} to break the sum constraint, the logit and the arcsin (also referred to as the arcsin square root, or angular) transformations:

$$g_{\text{arcsin}}(p_{ij}) = \text{arcsin}\left(\sqrt{p_{ij}}\right) \in [0, \pi/2], \text{ and} \tag{1}$$

$$g_{\text{logit}}(p_{ij}) = \log_2\left(\frac{p_{ij}}{1-p_{ij}}\right) \in (-\infty, \infty). \tag{2}$$

Over a broad range of intermediate values of the proportions, the logit and arcsin transformations are roughly linearly related to one another. However, although both transformations tend to pull out the ends of the distribution of p_{ij} values, this effect is more marked for the logit transformation, meaning that it is more affected by smaller differences at the ends of the scale.

Gaussian mixture models

Model-based clustering consists of assuming that the expression data come from several separately modeled subpopulations, where the full population of genes is a mixture of these subpopulations. Thus, observations are assumed to be a sample from an unknown probability distribution with density f, which is estimated by a finite mixture

$$f(.|\theta_K) = \sum_{k=1}^{K} \pi_k f_k(.|\alpha_k),$$

where $\theta_K = (\boldsymbol{\pi}, \alpha_1, \ldots, \alpha_K)$, and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ are the mixing proportions, with $\pi_k \in (0,1)$ for all k and $\sum_{k=1}^{K} \pi_k = 1$. The density $f_k(.|\alpha_k)$ of the k^{th} subpopulation must be chosen according to the nature of the gene expression measures; in the following, we consider the special case of Gaussian mixture models.

A collection of Gaussian mixture models can be defined as $(\mathcal{S}_m)_{m\in\mathcal{M}} = (\mathcal{S}_{(K,v)})_{(K,v)\in\mathcal{M}}$, where

$$\mathcal{S}_{(K,v)} = \left\{ f\left(.|\theta_{(K,v)}\right) = \sum_{k=1}^{K} \pi_{k,v} \phi\left(.|\mu_k, \Sigma_{k,v}\right) \right\}, \tag{3}$$

with $\phi\left(.|\mu_k, \Sigma_{k,v}\right)$ denoting the q-dimensional Gaussian density with mean μ_k and covariance matrix $\Sigma_{k,v}$. The index v denotes one of the Gaussian mixture shapes obtained by constraining one or more of the parameters in the following decomposition of each mixture component variance matrix:

$$\Sigma_k = \lambda_k D_k' A_k D_k, \tag{4}$$

where $\lambda_k = |\Sigma_k|^{1/q}$, D_k is the eigenvector matrix of Σ_k, and A_k is the diagonal matrix of normalized eigenvalues of Σ_k. Various constraints on these parameters respectively control the volume, orientation, and shape of the k^{th} cluster [31]; by additionally allowing the proportions π_k to vary according to cluster or be equal for all clusters, we may define a collection of 28 parsimonious and interpretable mixture models, available in the `Rmixmod` R package [32]. Without loss of generality, for simplicity of notation we will consider here only the most general model form, with variable proportions, volume, orientation, and shape (referred to as the $[p_k L_k C_k]$ in `Rmixmod`); as such, the model collection is defined solely over a range of numbers of clusters, $(\mathcal{S}_K)_{K\in\mathcal{M}}$.

The parameters of each model in the collection defined in (3) may be estimated using an expectation-maximization (EM)-type algorithm [33]. After solving the density estimation problem, for each model in the collection f is estimated by $\hat{f}_K = f(.|\hat{\theta}_K)$, and the data are clustered using the maximum a posteriori (MAP) rule: for each $i = 1, \ldots, n$ and each $k = 1, \ldots, K$,

$$\hat{z}_{ik} = \begin{cases} 1 & \text{if } t_{ik}\left(\hat{\theta}_K\right) > t_{i\ell}\left(\hat{\theta}_K\right) \; \forall \ell \neq k \\ 0 & \text{otherwise.} \end{cases}$$

Model choice for RNA-seq co-expression

In the mixture model framework, the number of clusters K is typically chosen from the model collection using a penalized selection criterion such as the BIC, [18], ICL [19], or a non-asymptotic penalized criterion whose penalty is calibrated using the slope heuristics (SH) principle [34]:

$$\begin{aligned}
\text{BIC}(K) &= -\mathcal{L}(.|\hat{\theta}_K) + \frac{D_K}{2n}\ln(n) && (5)\\
\text{ICL}(K) &= -\mathcal{L}(.|\hat{\theta}_K) + \frac{D_K}{2n}\ln(n) - \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} t_{ik}(\hat{\theta}_K)\ln(t_{ik}(\hat{\theta}_K)) \\
&= -\mathcal{L}(.|\hat{\theta}_K) + \frac{D_K}{2n}\ln(n) + \text{entropy} && (6)\\
\text{SH}(K) &= -\mathcal{L}(.|\hat{\theta}_K) + \kappa D_K, && (7)
\end{aligned}$$

where $\mathcal{L}(.|\hat{\theta}_K)$ is the loglikelihood evaluated at $\hat{\theta}_K$ (the maximum likelihood estimate of θ_K), D_K represents the number of free parameters for the mixtures in model $\mathcal{S}_K$, n is the number of observations, $t_{ik}(\hat{\theta}_K)$ is the conditional probability that observation i arises from the k^{th} component mixture $f(.|\hat{\theta}_K)$in model $\mathcal{S}_K$, and κ is a multiplicative constant that must be calibrated using the `capushe` R package [21]. The selected model $\hat{K}$ corresponds to the number of clusters K that minimizes the chosen criterion among Equations (5-7).

Although rarely done in practice, penalized criteria like the BIC and ICL may also be used to select among different models or transformations, as was suggested in a different context by Thomas et al. [35] and more recently for RNA-seq data by Gallopin [36]. This is of great interest, as it removes the need for an arbitrary choice of data transformation by using the framework of formal model selection. We illustrate this principle for the choice of number of clusters K and data transformation; in a more general case, a similar procedure could be used to additionally choose among the differerent forms of Gaussian mixture models described in Equation (4) or among different parameteric forms of models. Let $g(\mathbf{x})$ represent an arbitrary monotonic transformation of a dataset $\mathbf{x}$. If the new sample $g(\mathbf{x})$ is assumed to arise from an i.i.d. Gaussian mixture density, $f(.|\theta_K)$, then the initial data $\mathbf{x}$ is an i.i.d. sample from density $f_g(.|\theta_K)$, which is a transformation of $f(.|\theta_K)$ and thus not necessarily a Gaussian mixture density. If J_g denotes the Jacobian of the transformation g and $\hat{\theta}_{(K,g)}$ the maximum likelihood estimate obtained for the model with K clusters and transformation g, we select the pair (K, g) leading to the minimum of the corrected BIC or ICL criteria:

$$\begin{aligned}
\text{BIC}^*(K, g) &= -\mathcal{L}(.|\hat{\theta}_{(K,g)}) + \frac{D_K}{2n}\ln(n) - \ln[\det(J_g)]\\
\text{ICL}^*(K, g) &= -\mathcal{L}(.|\hat{\theta}_{(K,g)}) + \frac{D_K}{2n}\ln(n) + \text{entropy} - \ln[\det(J_g)]. && (8)
\end{aligned}$$

Note that in these expressions, the number of parameters D_K does not depend on the transformation g.

For the purposes of this work, we make use of the corrected ICL criterion defined in Equation (8) to compare between the logit and arcsin transformations in Equations (1) and (2) applied to the expression profiles $\mathbf{p} = (p_{ij})$. In particular, we use the following:

$$\text{ICL}^*_{\text{arcsin}}(K, g_{\text{arcsin}}) = -\mathcal{L}(.|\hat{\theta}_{(K,g_{\text{arcsin}})}) + \frac{D_K}{2n}\ln(n) + \text{entropy} + nq\ln(2) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{q}\ln\left[p_{ij}(1-p_{ij})\right], \quad (9)$$

$$\text{ICL}^*_{\text{logit}}(K, g_{\text{logit}}) = -\mathcal{L}(.|\hat{\theta}_{(K,g_{\text{logit}})}) + \frac{D_K}{2n}\ln(n) + \text{entropy} + nq\ln[\ln(2)] + \sum_{i=1}^{n}\sum_{j=1}^{q}\ln\left[p_{ij}(1-p_{ij})\right]. \quad (10)$$

The values of $\text{ICL}^*_{\text{arcsin}}(K, g_{\text{arcsin}})$ and $\text{ICL}^*_{\text{logit}}(K, g_{\text{logit}})$ can thus be directly compared to choose between the two transformations.

coseq R package

To facilitate co-expression analyses of RNA-seq data using Gaussian mixture models and an appropriate data transformation, we have created the R package coseq (**co**-expression of RNA-**seq** data), freely available on the R-Forge; in this section, we briefly describe some of the options available in this package.

The package is first installed and loaded using the following commands:

```
> install.packages("coseq", repos="http://R-Forge.R-project.org")
> library(coseq)
```

A typical call to coseq to fit a Gaussian mixture model on arcsin- or logit-transformed normalized profiles takes the following form:

```
> run_arcsin <- coseq(counts, K=2:10, model="Normal", transformation="arcsin")
> run_logit <- coseq(counts, K=2:10, model="Normal", transformation="logit")
```

where counts represents a $(n \times q)$ matrix or data frame of read counts for n genes in q samples and K=2:10 provides the desired range of numbers of clusters (here, 2 to 10). We note that this function directly calls the Rmixmod R package to fit Gaussian mixture models [32]. For backwards compatability with our previous method [17], a similar function call may be used to fit a Poisson mixture model on raw counts using the HTSCluster package:

```
> run_pois <- coseq(counts, conds, K=2:10, model="Poisson")
```

where a vector conds is additionally provided to identify the experimental condition associated with each column in counts. In both cases, the output of the coseq function is an S3 object on which standard plot and summary functions can be directly applied; the former uses functionalities from the ggplot2 package [37]. Several examples of the standard plot commands can be seen in the Results section of this work, as well as in the reproducible Rmarkdown document included in the Supplementary Materials. The option of parallelization via the BiocParallel Bioconductor package is also provided.

In addition to the choice of mixture model and transformation to be used, the coseq function provides flexibility to the user to filter normalized read counts according to their mean value if desired, specify library size normalization method (TMM, median ratio, upper quantile, or user-provided normalization factors), and modify `Rmixmod` options (number of iterations, etc). For the specific case of arcsin- and logit-transformed normalized profiles, we provide a convenience function `compareICL` to calculate and plot the corrected ICL model selection criteria defined in Equations (9) and (10). Finally, as RNA-seq expression analyses are often performed on a subset of genes identified as differentially expressed, the `coseq` function can also be directly called on an `DESeqResults` S4 object or integrated with `DGELRT` S4 objects, respectively corresponding to output from the `DESeq2` [11] and `edgeR` [10] Bioconductor packages for RNA-seq differential analyses. For more details and examples, see the full package vignette provided with `coseq`.

Results

In the following, we illustrate co-expression analyses using Gaussian mixture models in conjunction with the proposed transformations on normalized expression profiles for several real RNA-seq datasets. The data were selected to represent several different organisms (pig, mouse, human, fly) in studies for which co-expression is of particular interest (across tissues or across time); additional details on how data were obtained and preprocessed may be found in the Supplementary Materials.

Description of RNA-seq data

Porcine small intestine: Mach et al. [38] used RNA-seq to study site-specific gene expression along the gastrointestinal tract of four healthy 70-day-old male Large White piglets. Samples were collected in three sites along the proximal-distal axis of the small intestine (duodendum, jejunum, and ileum), as well as the ileal Peyer's patch (a lymphoid tissue localized in direct contact with the epithelial intestinal tissue). Complete information regarding sample preparation, sequencing, quality control, and pre-processing are available in the original article [38]. Raw reads are available at NCBI's SRA repository (PRJNA221286 BioProject; accessions SRR1006118 to SRR1006133); in the current work, read counts for genes sharing a common gene symbol or Ensembl gene ID were summed.

Embryonic mouse neocortex: Fietz et al. [39] studied the expansion of the neocortex in five embryonic (day 14.5) mice by analyzing the transcriptome of the ventricular zone (VZ), subventricular zone (SVZ), and cortical plate (CP) using RNA-seq. Laser-capture microdissection, RNA isolation and cDNA library preparation, and RNA sequencing and quantification are described in the Supplementary Materials of Fietz et al. [39]. In our work, raw read counts for this study were downloaded on December 23, 2015 from the Digital Expression Explorer (DEE) [40] using associated SRA accession number SRP013825, and run information was downloaded using the SRA Run Selector. Additional information about the DEE processing pipeline may be found in the Supplementary Materials.

Fetal human neocortex: In the aforementioned study, Fietz et al. [39] also included samples from 6 (13-16 wk postconception) human fetuses taken from four neocortex regions: CP, VZ, and inner and outer subventricular zone (ISZZ and OSVZ, respectively). Raw counts were obtained in the same manner as described above.

Dynamic expression in embryonic flies: As part of the modENCODE project to annotate functional elements of the *Drosophila melanogaster* genome, Graveley et al. [29] characterized the expression dynamics of the fly using RNA-seq over 27 distinct stages of development, from early embryo to ageing male and female adults. As in our previous co-expression work [17], we focus on a subset of these data from 12 embryonic samples that were collected at 2-hour intervals for 24 hours, with one biological replicate for each time point. Phenotype tables and raw read counts were obtained from the ReCount online resource [41].

Results on real RNA-seq data

We used the `coseq` package described in the previous section to fit Gaussian mixture models to the arcsin- and logit-transformed normalized profiles for each of the four datasets described above for $K = 2, \ldots, 40$ clusters (with the exception of the *Drosophila melanogaster* data, for which a maximum value of $K = 60$ was used), using the TMM library size normalization, filtering genes with mean normalized count less than 50, and otherwise using default values for parameters. Concerning the filtering step, screening using either a differential analysis or a threshold on normalized means or coefficients of variation are often applied in practice prior to co-expression analyses to remove features that contribute noise. In all cases, we calculated the corrected ICL values from Equations (9) and (10) to compare between the arcsin and logit transformations; the number of clusters $\hat{K}$ identified for each transformation, as well as the preferred model-transformation pair chosen via the corrected ICL, are shown for each dataset in Table 1. The corrected ICL values across a range of numbers of clusters K are shown in Figure 2 for the Graveley et al. [29] fly and Fietz et al. [39] mouse data; for clarity, we focus our discussion in the main text on these two datasets, but complete and reproducible results (in the form of an Rmarkdown document) for all four RNA-seq datasets may be found in the Supplementary Materials.

For the remainder of the article, the results presented correspond to the model selected via the corrected ICL. It is of interest to investigate the per-cluster covariance structures estimated for the selected models for each of the RNA-seq datasets. As an example, the per-cluster correlation matrices estimated by `coseq` for two selected clusters from the Graveley et al. [29] and Fietz et al. [39] mouse data are shown in Figure 3. It is interesting to note that although the Gaussian mixture model does not explicitly incorporate the experimental condition labels $\mathcal{C}(j)$, the estimated models include large cluster-specific correlations among close time points

Organism	Reference	Conditions (reps)	$\hat{K}_{\text{arcsin}}$	$\hat{K}_{\text{logit}}$
Pig	Mach et al. [38]	4 tissues (4 reps each)	**14**	11
Mouse	Fietz et al. [39]	3 tissues (5 reps each)	**12**	15
Human	Fietz et al. [39]	4 tissues (6 reps each)	**6**	8
Fly	Graveley et al. [29]	12 time pts (1 rep each)	**28**	23

Table 1: Summary of results for Gaussian mixture models fit on transformed normalized RNA-seq profiles. For each dataset, the organism, associated reference, experimental conditions $\mathcal{C}(j)$ and number of biological replicates in each, and number of clusters selected via ICL for the arcsin and logit transformation are provided. Boldface values indicated the final model selected via the corrected ICL.

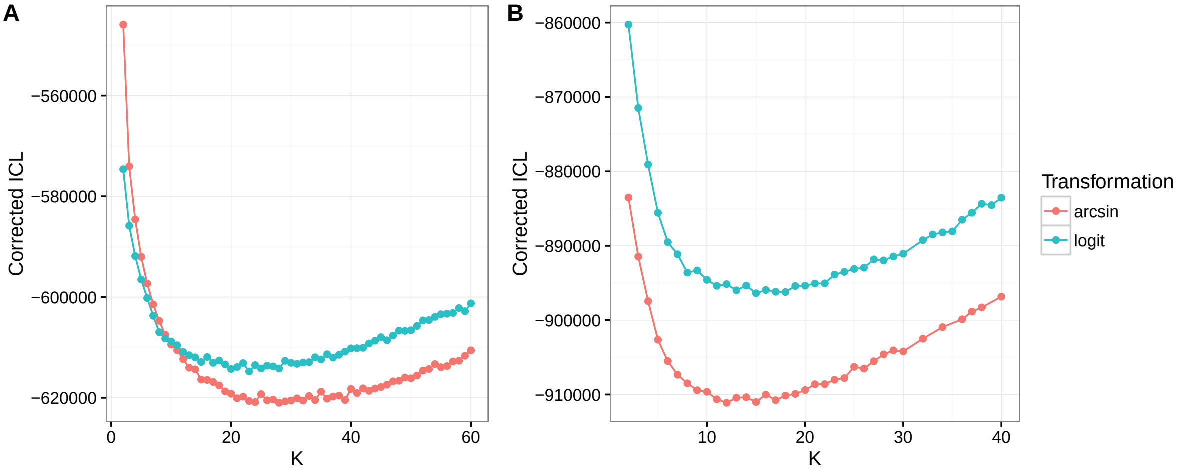

Figure 2: Corrected ICL values for the arcsin (red) and logit (blue) transformed normalized expression profiles over a range of numbers of clusters K for the Graveley et al. [29] fly and Fietz et al. [39] mouse data (A and B, respectively).

(Figures 3A and 3B) or among replicates within each tissue (Figures 3C and 3D). In addition, cluster-specific correlation structures among regions may be clearly seen; for example, in the Fietz et al. [39] mouse data, Cluster 2 is characterized by very large negative correlations between the CP and SVZ/VZ regions, while Cluster 3 instead has a strong negative correlation between the VZ and CP/SVZ regions. This strongly suggests that in these data, the assumption of conditional independence among samples assumed by the Poisson mixture model described in Rau et al. [17] is indeed unrealistic.

There are several ways in which per-cluster expression profiles can be represented graphically, depending on the type of data plotted (normalized counts, normalized expression profiles, or transformed normalized profiles), the type of plot (e.g., line plots or boxplots), and whether replicates within experimental conditions are averaged or plotted independently. Regarding the

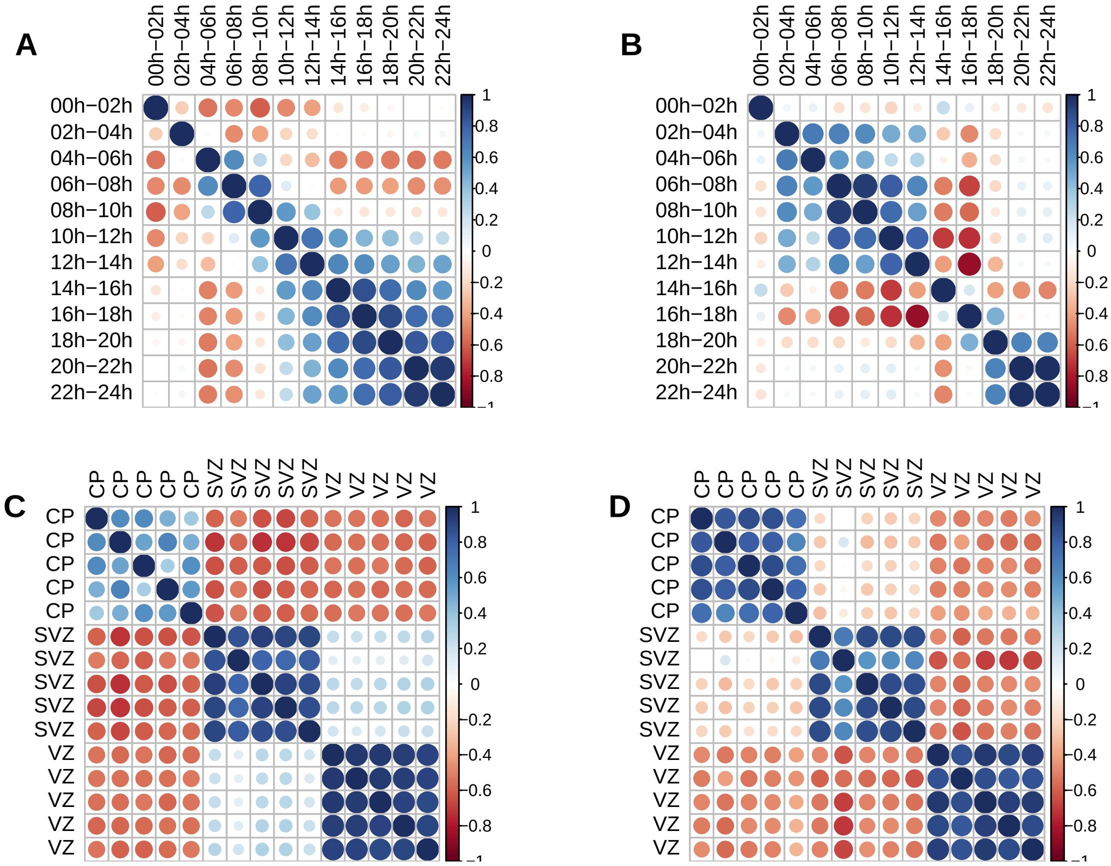

Figure 3: Per-cluster correlation matrices for clusters 25 (A) and 27 (B) from the Graveley et al. [29] fly data and for clusters 2 (C) and 3 (D) from the Fietz et al. [39] mouse data. Dark blue and red represent correlations close to 1 and -1, respectively, and circle areas correspond to the absolute value of correlation coefficients. Correlation matrices are visualized using the `corrplot` R package.

latter point, note that the Gaussian mixture model is fit on the entirety of the data, and replicate averaging is proposed to simplify the visualization of cluster-specific expression. Although the `coseq` package facilitates the implementation of any combination of these three graphical options, our recommendations for visualizating co-expression results are as follows: 1) although “tighter" profiles are observed when plotting the transformed normalized profiles (as these are the data used to fit the model), interpretation of profiles is improved by instead using the untransformed normalized profiles; 2) boxplots are generally preferable when experimental conditions $\mathcal{C}(j)$ represent distinct groups, although line plots can be useful for time-course experiments; 3) averaging replicates prior to plotting often provides clearer distinctions among cluster-specific profiles. Following these recommendations, the cluster-specific profiles iden-

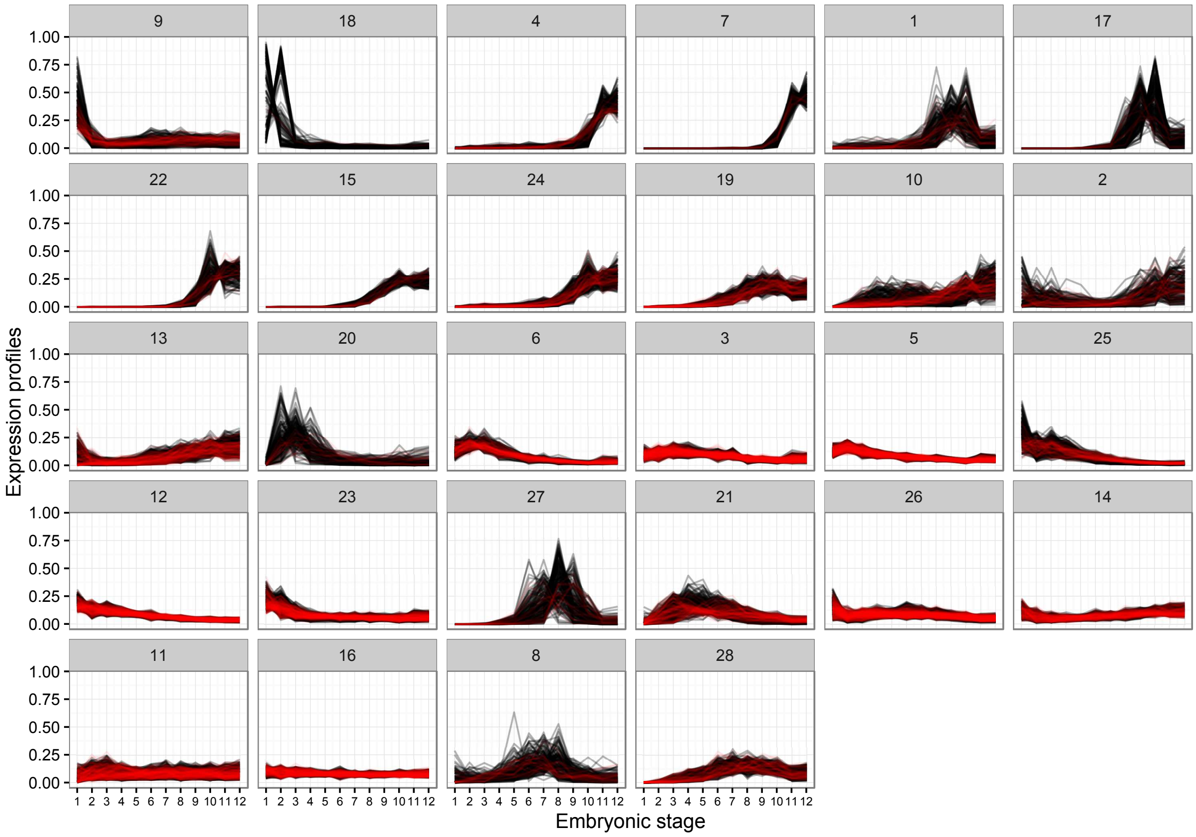

Figure 4: Per-cluster expression profiles for the Graveley et al. [29] data. Clusters have been sorted so that those with similar mean vectors (as measured by the Euclidean distance) are plotted next to one another. Red lines correspond to genes with maximimum conditional probability of cluster membership < 0.8.

tified for the Graveley et al. [29] and Fietz et al. [39] mouse data are shown in Figures 4 and 5.

An additional advantage of model-based clustering approaches is that they facilitate an evaluation of the clustering quality of the selected model by examining the maximum conditional probabilities of cluster membership for each gene:

$$t_{\max}(i) = \max_{1 \le k \le \hat{K}} t_{ik}\left(\hat{\theta}_{\hat{K}}\right), \; i = 1, \ldots, n.$$

Boxplots of the maximum conditional probabilities $t_{\max}(i)$ per cluster for the Graveley et al. [29] and Fietz et al. [39] mouse data are presented in Figure 6. It may be seen that across clusters, the majority of genes in both datasets have a large value (i.e., close to 1) for $t_{\max}(i)$; the number of genes with $t_{\max}(i) > 0.8$ is 7822 (82.1%) and 7382 (82.4%) for the Graveley et al. [29] and Fietz et al. [39] mouse data, respectively. However, the boxplots also illustrate

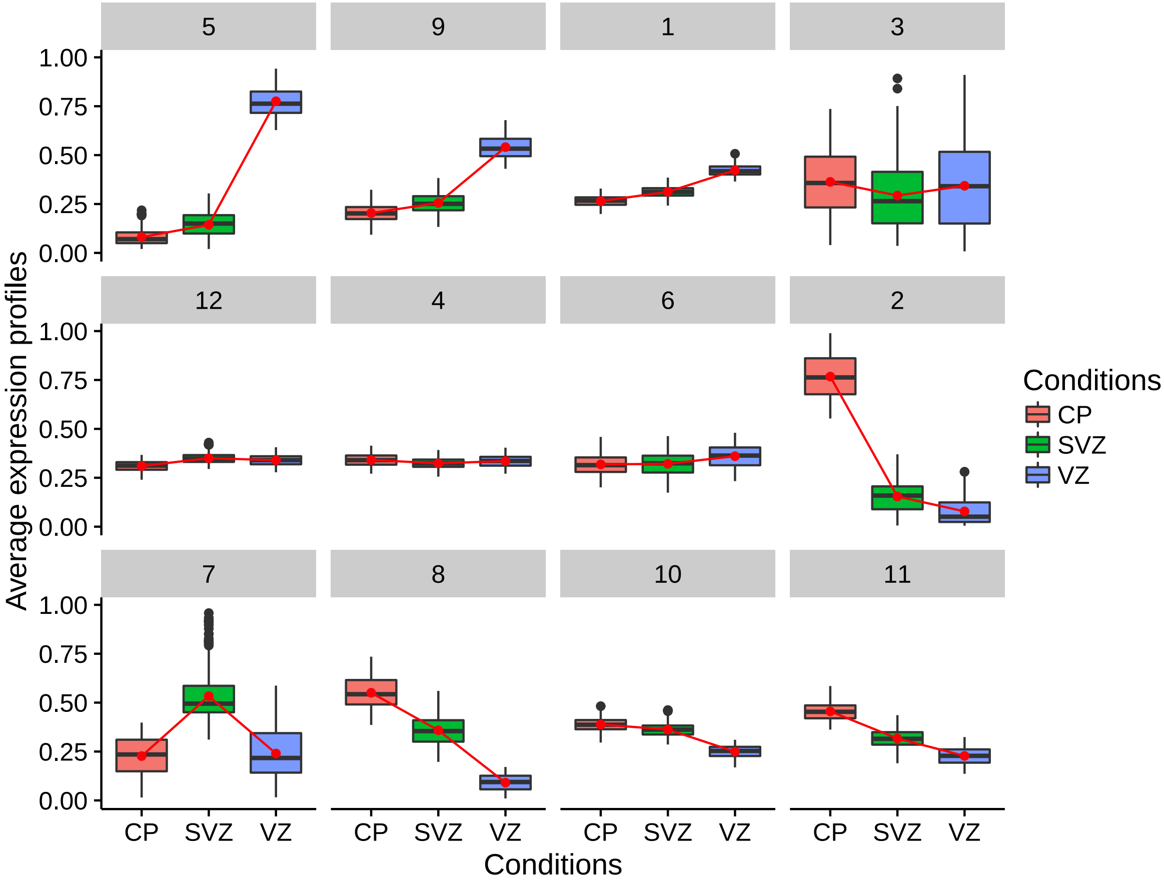

Figure 5: Per-cluster expression profiles for the Fietz et al. [39] data. Clusters have been sorted so that those with similar mean vectors (as measured by the Euclidean distance) are plotted next to one another. Connected red lines correspond to the mean expression profile for each group.

that some genes have a $t_{\max}(i)$ less than this threshold, in some cases as low as 0.4; this indicates that for a small number of genes, the cluster assignment is fairly ambiguous and assignment to a single cluster is questionable (the gene with the smallest $t_{\max}(i)$ in the Fietz et al. [39] mouse data had a conditional probability of 24.8%, 32.2%, 13.0% and 30.0% of belonging to clusters 1, 4, 6, and 12, respectively). In such cases, it may be prudent to focus attention on genes with highly confident cluster assignments (e.g., those with $t_{\max}(i) > 0.8$).

Finally, examining the distribution $t_{\max}(i)$ values within each cluster provides information about the homogeneity and relevance of each cluster. For both datasets, all cases clusters are primarily made up of genes with highly confident $t_{\max}(i)$ values; however some clusters (e.g., Clusters 1 and 4 in the Graveley et al. [29] data) appear to be more homogeneous and well-formed than others (e.g., Clusters 16 and 3 in the same data). These conclusions align with the general observations made about the per-cluster normalized profiles in Figure 4, where it

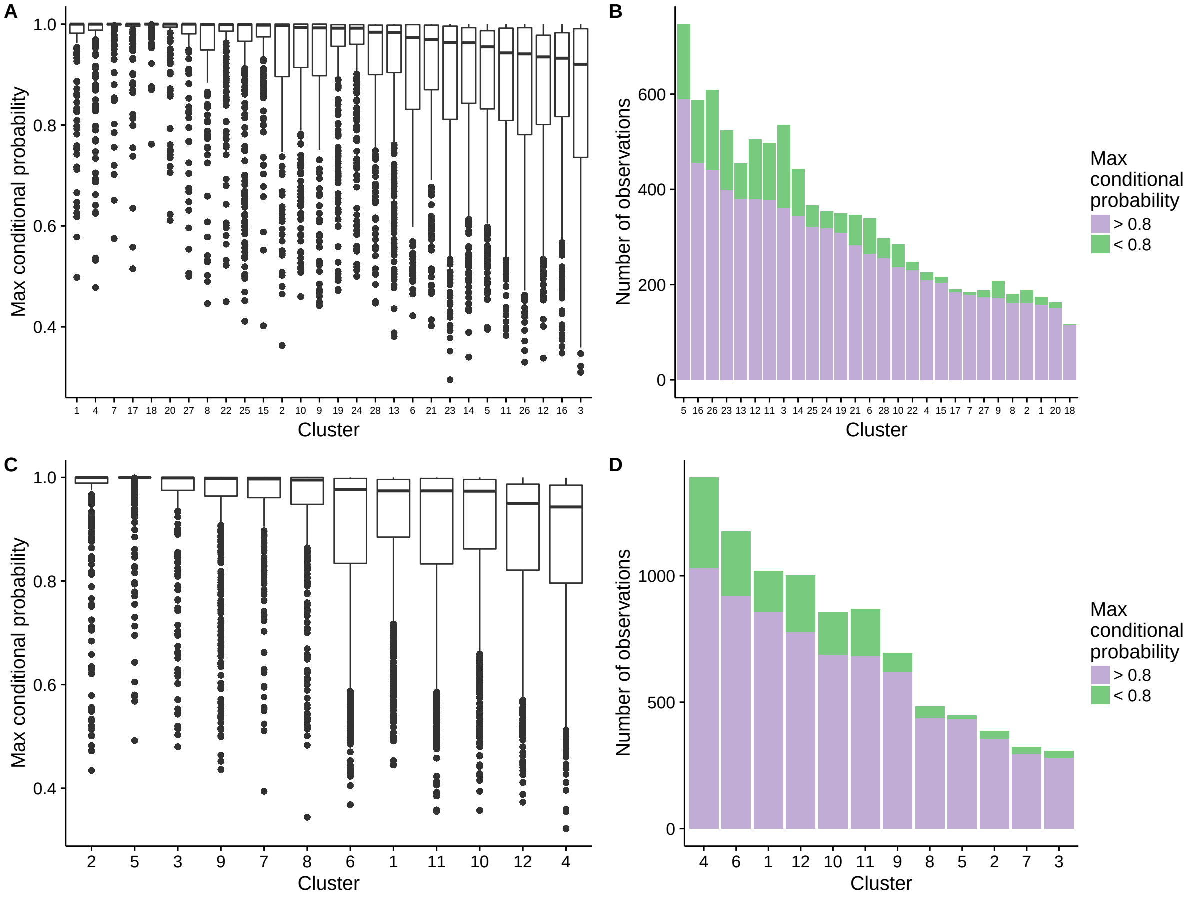

Figure 6: Evaluation of clustering quality for the Graveley et al. [29] data. (left) Maximum conditional probabilities $t_{\max}(i)$ for each cluster, sorted in decreasing order by the cluster median. (right) Barplots of cluster sizes, according to $t_{\max}(i)$ greater than or less than 0.8, sorted according to the number of genes with $t_{\max}(i) > 0.8$.

may be seen that clusters 16 and 3 have quite similar profiles (suggesting that unambiguous assignment to one of these clusters is more difficult).

Discussion and recommendations

In this work, we have primarily addressed the choice of data to be clustered (transformed normalized profiles rather than raw counts) for RNA-seq co-expression analysis under the framework of Gaussian mixture models; we note that many alternative clustering strategies exist based on a different algorithms (e.g., K-means or hierarhical clustering) and distance measures calculated among pairs of genes (e.g., Euclidean distance, correlation, etc). The difficulty of comparing clusterings arising from different approaches is well-known, and it is rarely straightforward to establish the circumstances under which a given strategy may be preferred. However, we have illustrated several advantages in using Gaussian mixture models in conjunction with appropriately defined transformations to identify groups of co-expressed genes from normalized RNA-seq gene expression profiles:

- Mixture models in general have the advantage of providing a rigorous statistical framework for parameter estimation, an objective assessment of the number of clusters present in the data through the use of penalized criteria, and the possibility of performing diagnostic checks on the quality and homogeneity of the resulting clusters. In particular, diagnostic plots on the maximum conditional probabilities of cluster membership provide a global overview of the clustering and an objective explanation of the quality of cluster assignments. Since only a subset of genes are expected to be assigned to biologically interpretable groups, these diagnostic plots help provide a basis for discussion about the choice of genes for follow-up study.

- Gaussian mixtures in particular represent a rich, flexible, and well-characterized class of models that have been successfully implemented in a large variety of theoretical and applied research contexts. For RNA-seq data, this means that the model may directly account for per-cluster correlation structures among samples, which can be quite strong in real RNA-seq data. In this work we considered a single form of Gaussian covariance matrices (the $[p_K L_k C_k]$ form), but any or all of the 28 forms of Gaussian mixture models could be used in practice.

We have also discussed the use of penalized criteria like the ICL and BIC to objectively compare results between different transformations, and potentially among different forms of Gaussian covariance matrices or among different models. For the four datasets considered here, the arcsin transformation of normalized expression profiles was consistently preferred to the logit transformation; as previously mentioned, this is likely due the sensitivity of the latter to very small p_{ij} values. An interesting further direction of research would be to consider approaches able to directly model the compositional nature of normalized profiles p_{ij} without the need to apply an arcsin or logit transformation.

In addition to the choice of clustering method, several practical issues should be considered in co-expression analyses. First, a common question is whether genes should be screened prior to the analysis (e.g., via an upstream differential analysis or filter based on the mean expression or coefficient of variation for each gene). Such a screening step is often used in practice, as genes contributing noise but little biological signal of interest can adversely affect clustering results. A second common question pertains to whether replicates within a given experimental group should be modeled independently or summed or averaged prior to the co-expression analysis. Although technical replicates in RNA-seq data are typically summed prior to analysis, in this work we fit Gaussian mixture models on the full data including all biological replicates; subsequently to visualize clustering results, replicate profiles are averaged for improved clarity of cluster profiles.

Following a co-expression analysis, it is notoriously difficult to validate the results of a clustering algorithm on transcriptomic data, and such results can be evaluated based on either statistical criteria (e.g., between-group and within-cluster inertia measures) or external biological criteria. In practice groups of co-expressed genes are further characterized by analyzing and integrating various resources, such as functional annotation or pathway membership information from databases like the Gene Ontology Consortium. Such functional analyses can be useful for providing interpretation and context for the identified clusters.

References

[1] M. B. Eisen et al. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25):14863–14868, 1998.

[2] D. Jiang, C. Tang, and A. Zhang. Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1370–1386, 2004.

[3] K. Y. Yeung et al. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987, 2001.

[4] S. Anders, P.T. Pyl, and W. Huber. A Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, 2015.

[5] Y. Liao, G. K. Smyth, and W. Shi. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2014.

[6] B. Li and C. N. Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(323), 2011.

[7] N. L. Bray et al. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34:525–527, 2016.

[8] A. Oshlack and M. J. Wakefield. Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, 4(14), 2009.

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

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

[11] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(550), 2014.

[12] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[13] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, number 1, pages 281–297. Berkeley, University of California Press, 1967.

[14] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley-Interscience, 2000.

[15] D. M. Witten. Classification and clustering of sequencing data using a Poisson model. *Annals of Applied Statistics*, 5(4):2493–2518, 2011.

[16] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(R106):1–28, 2010.

[17] A. Rau et al. Co-expression analysis of high-throughput transcriptome sequencing data with poisson mixture models. *Bioinformatics*, 31:1420–1427, 2015.

[18] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[19] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725, 2000.

[20] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc.*, 3:203–268, 2001.

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

[22] Y. Si et al. Model-based clustering for RNA-seq data. *Bioinformatics*, 30(2):197–205, 2014.

[23] C.W. Law et al. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(R29), 2014.

[24] G. K. Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 1(3):1–26, 2004.

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

[26] M.-A. Dillies et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683, 2013.

[27] R. Tibshirani. Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Association*, 83:394–405, 1988.

[28] W. Huber et al. Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical Applications in Genetics and Molecular Biology*, 2(1):Article 3, 2003.

[29] B. R. Graveley et al. The development transcriptome of *Drosophila melanogaster*. *Nature*, 471:473–479, 2011.

[30] J. Aitchison. *The Statistical Analysis of Compositional Data*. Chapman & Hall, 1986.

[31] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781 – 793, 1995.

[32] R. Lebret et al. Rmixmod: The R Package of the model-based unsupervised, supervised, and semi-supervised classification Mixmod library. *Journal of Statistical Software*, 67(6): 1–29, 2015.

[33] A. P. Dempster et al. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1):1–38, 1977.

[34] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138:33–73, 2007.

[35] I. Thomas, P. Frankhauser, and C. Biernacki. The fractal morphology of the built-up landscape. *Landscape of Urban Plan*, 84(2):99–115, 2008.

[36] M. Gallopin. *Classification et inférence de réseaux pour les données RNA-seq*. PhD thesis, Université Paris-Saclay, 2015.

[37] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. URL `http://ggplot2.org`.

[38] N. Mach et al. Extensive expression differences along porcine small intestine evidenced by transcriptome sequencing. *PLoS ONE*, 9(2):1–12, 02 2014.

[39] S. A. Fietz et al. Transcriptomes of germinal zones of human and mouse fetal neocortex suggest a role of extracellular matrix in progenitor self-renewal. *PNAS*, 109(29):11836–11841, 2012.

[40] M. Ziemann et al. Digital Expression Explorer: A user-friendly repository of uniformly processed RNA-seq data. In *ComBio2015*, volume POS-TUE-099, Melbourne, 2015. doi: 10.13140/RG.2.1.1707.5926.

[41] A. C. Frazee, B. Langmead, and J. T. Leek. ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, 12, 2011.

Key Points

- After applying an appropriate transformation, Gaussian mixture models represent a rich, flexible, and well-characterized class of models to identify groups of co-expressed genes from RNA-seq data. In particular, they directly account for per-cluster correlation structures among samples, which are observed to be quite strong in typical RNA-seq data.
- Normalized expression profiles, rather than raw counts, are recommended for co-expression analyses of RNA-seq data. Because these data are compositional in nature, an additional transformation (e.g., arcsin or logit) is required prior to fitting a Gaussian mixture model.
- Penalized model selection criteria like the BIC or ICL can be used to select both the number of clusters present and the appropriate transformation to use; in the latter case, an additional term based on the Jacobian of the transformation is added to the criterion, yielding a corrected BIC or ICL that can be used to directly compare two transformations.

Funding

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grant MixStatSeq (ANR-13-JS01-0001-01).

Acknowledgements

The authors would like to thank Gilles Celeux, Sandrine Laguerre, Béeatrice Laurent, Clément Marteau, and, Marie-Laure Martin-Magniette for helpful discussions.