Clustering transformed compositional data using K-means, with applications in gene expression and bicycle sharing system data

Antoine Godichon-Baggioni, Cathy Maugis-Rabusseau, Andrea Rau

arXiv: arXiv:0000.0000

Clustering transformed compositional data using K-means, with applications in gene expression and bicycle sharing system data

Antoine Godichon-Baggioni*, Cathy Maugis-Rabusseau and Andrea Rau

Institut de Mathématiques de Toulouse
Universit Toulouse III - Paul Sabatier
118 route de Narbonne
31062 Toulouse, France.
e-mail: godichon@insa-toulouse.fr

Institut Mathématiques de Toulouse
INSA de Toulouse,
135 avenue de Rangueil,
31077 Toulouse, France.
e-mail: cathy.maugis@insa-toulouse.fr

Institut National de la Recherche Agronomique
Domaine de Vilvert
78352 Jouy-en-Josas, France.
e-mail: andrea.rau@inra.fr

Abstract: Although there is no shortage of clustering algorithms proposed in the literature, the question of the most relevant strategy for clustering compositional data (i.e., data made up of profiles, whose rows belong to the simplex) remains largely unexplored in cases where the observed value of an observation is equal or close to zero for one or more samples. This work is motivated by the analysis of two sets of compositional data, both focused on the categorization of profiles but arising from considerably different applications: (1) identifying groups of co-expressed genes from high-throughput RNA sequencing data, in which a given gene may be completely silent in one or more experimental conditions; and (2) finding patterns in the usage of stations over the course of one week in the Velib' bicycle sharing system in Paris, France. For both of these applications, we focus on the use of appropriately chosen data transformations, including the Centered Log Ratio and a novel extension we propose called the Log Centered Log Ratio, in conjunction with the K-means algorithm. We use a nonasymptotic penalized criterion, whose penalty is calibrated with the slope heuristics, to select the number of clusters present in the data. Finally, we illustrate the performance of this clustering strategy, which is implemented in the Bioconductor package `coseq`, on both the gene expression and bicycle sharing system data.

MSC 2010 subject classifications: Primary 62H30; secondary 62P10.
Keywords and phrases: Clustering, compositional data, data transformations, K-means.

*Corresponding author

1. Introduction

Compositional data are made up of the relative proportions of a whole and can be represented in the simplex of d parts:

$$\mathcal{S}^d := \left\{ x = (x_1, ..., x_d) \in \mathbb{R}^d | \sum_{i=1}^{d} x_i = 1, x_i > 0, \forall i \right\}.$$

Such data are a common phenomenon in several research domains, including evolutionary ecology (Aebischer et al., 1993; Bingham et al., 2007), geochemistry (Buccianti et al., 2006; Miesch and Chapman, 1977), economics (DeSarbo et al., 1995; Longford and Pittau, 2006), and genomic surveys (Friedman and Alm, 2012); see Pawlowsky-Glahn and Buccianti (2011) for more examples of applications. The statistical analysis of compositional data has been the focus of research for well over a century (Pearson, 1896; Chayes, 1960). One of the initial concerns was to propose appropriate methods able to account for the nature of compositional data, notably the fact that they are subject to a unit sum constraint, which invalidates many standard statistical approaches. For example, Pawlowsky-Glahn and Egozcue (2001) provided a formal definition of metric center and variance for random compositional data, as well as a law of large numbers. Mateu-Figueras et al. (2013) provided a definition for the normal distribution in the simplex, as well as the relevant central limit theorem. The concept of proximity between two compositional vectors has been defined using several proposed distance measures, such as Aitchison's distance (Aitchison, 1982), which is typically based on data transformations such as the Centered Log Ratio (CLR), the Additive Log Ratio (ALR), or the Isometric Log Ratio (ILR) (Egozcue et al., 2003). These compositional data transformations facilitate the use of statistical methods based on Gaussian distributions and Euclidean structures for compositional data.

Many clustering methods exist in the literature, and they can primarily be divided into two classes: model-based methods, such as mixture models (McLachlan and Peel, 2004), and methods based on dissimilarity distances, such as hierarchical clustering (Ward Jr, 1963), K-means (MacQueen et al., 1967), or K-medians (Cardot et al., 2012). However, despite the large number of existing clustering methods, to our knowledge there has been relatively little attention paid to the most appropriate strategy for clustering compositional data (Tauber, 1999; Zhou et al., 1991; Martín-Fernández et al., 1998). In recent work, Rau and Maugis-Rabusseau (2017) proposed the use of data transformations (either the arcsine or logit) and Gaussian mixture models to cluster compositional data arising from high-throughput transcriptome sequencing data. This strategy led to satisfactory results in practice as the proposed transformations removed the linear dependence present among data coordinates, thus enabling estimation of model parameters; however, such an approach implies that the dependencies among coordinates are entirely ignored.

In this work, our aim is twofold: (1) to introduce clustering methods for compositional data based on dissimilarity distances, in particular the K-means

algorithm (MacQueen et al., 1967); and (2) to use or define appropriate transformations for compositional data that account for the dependencies among coordinates. In particular, we investigate three clustering strategies for the task of clustering compositional data via the K-means algorithm with the usual Euclidean distance: (1) using untransformed data, which is arguably the most intuitive but does not directly account for the compositional nature of the data; (2) using data transformed using the CLR (Aitchison, 1982), which is specific to data on the simplex; and (3) transforming data using a novel extension of the CLR transformation, called the Log Centered Log Ratio (logCLR), consisting of a modification designed to provide greater separation for edge case observations, i.e. those with compositional values close to zero in one or more coordinates and those located near the vertices of the simplex. Finally, we select the number of clusters using a nonasymptotic criterion whose penalty is calibrated using the slope heuristics (Baudry et al., 2012; Fischer, 2011). This approach of using a K-means algorithm in conjunction with compositional data transformations is implemented in the Bioconductor package `coseq`, and we apply it to two practical problems: identifying groups of co-expressed genes from high-throughput RNA sequencing data, and finding patterns in the usage of stations in a bicycle sharing system.

The remainder of this paper is organized as follows: in Section 2, we describe the context and data from the transcriptomic and bicycle sharing applications in greater detail. In Section 3, we present the use of the K-means algorithm for the three proposed transformations, as well as the model selection approach used to select the number of clusters. A full analysis using this approach is performed on the transcriptomic and bicycle sharing system data in Sections 4 and 5, respectively. Finally, we provide some conclusions and recommendations in Section 6.

2. Description of motivating data

2.1. *Transcriptomic data*

Gene expression studies now routinely make use of high-throughput sequencing technology, which directly sequences reverse-transcribed RNA molecules in an approach called RNA sequencing (RNA-seq). After aligning sequenced reads to a reference genome and quantifying the number of reads attributed to each gene, RNA-seq data correspond to tables of read counts or pseudo-counts $\left(Y_{i,j}\right)$ representing the number of sequenced reads observed for genes $i = 1, \ldots, n$ in biological samples $j = 1, \ldots, d$, the latter of which may arise from several experimental conditions (e.g., across time, in different tissues). In this work, our focus will be on identifying groups of *co-expressed* genes with the same expression dynamics across all biological samples from an RNA-seq study; these co-expression clusters are often assumed to be involved in similar biological processes or to be candidates for co-regulation.

RNA-seq data tend to be characterized by highly skewed count values covering several orders of magnitude, as well as variable *library sizes* among sam-

ples (i.e., the total number of sequenced reads in a sample). It is often assumed that the read count $Y_{i,j}$ is proportional to the expression of gene i, weighted by the library size of sample j (as gene read counts in samples with larger total numbers of sequenced reads tend to have larger read counts) as well as its length (i.e., the number of nucleotides making up its coding region, as longer genes also tend to have larger read counts). As such, clustering strategies for RNA-seq data must account for both the length of each gene and the library size of each sample. Regarding the former, as is typically done in the context of differential analyses, we will make use of per-sample scaling normalization factors (t_j) calculated using the Trimmed Mean of M-values (TMM) approach (Robinson and Oshlack, 2010). For all $j = 1, \ldots, d$, let

$$\ell_j := t_j \sum_{i=1}^{n} Y_{i,j}, \qquad\qquad s_j := \frac{\ell_j}{\sum_{j'=1}^{d} \ell_{j'}/d},$$

be the normalized library sizes and the associated scaling factors by which raw counts are divided, respectively. As such, read counts normalized for differences in library sizes may be calculated as $Y_{i,j}/s_j$. To account for differences in the length of each gene, similarly to Rau and Maugis-Rabusseau (2017), we calculate the normalized expression *profiles* defined for each gene i by $X_i = (X_{i,1}, \ldots, X_{i,d})$, where

$$X_{i,j} := \frac{Y_{i,j}/s_j + 1}{\sum_{j'=1}^{d} \left(Y_{i,j'}/s_{j'} + 1\right)}, \qquad \forall j = 1, \ldots, d,$$

and a constant of 1 has been added to the normalized expression $Y_{i,j}/s_j$ prior to calculating the profiles due to the presence of 0's in the data. Note that we now have a table of compositional values $(X_{i,j})$ whose rows X_i belong to the simplex $\mathcal{S}^d$. In other words, the normalized expression profile value for a given gene in each sample is now relative to the total number of reads observed across all samples, meaning that this measure is independent of both its absolute expression level and its length; the normalized expression profiles thus facilitate the clustering of expression dynamics, as desired.

In Section 4, we will focus our attention on two sets of RNA-seq data:

- **Embryonic mouse neocortex:** Fietz et al. (2012) 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 place (CP). Details on the sample preparation, sequencing, and quantification are provided in the Supplementary Materials of Fietz et al. (2012). In the current work, raw read counts were downloaded from the Digital Expression Explorer (Ziemann et al., 2015) as described in Rau and Maugis-Rabusseau (2017). The data consist of transcriptome-wide measurements in three tissues with five replicates each. As suggested by Rau and Maugis-Rabusseau (2017), clustering will be performed on the full set of data (after filtering, on $n = 8969$ genes), and visualization of resulting clusters will be done on profiles summed over each tissue.

- **Dynamic expression in embryonic flies:** Graveley et al. (2011) characterized the expression dynamics of the fly using RNA-seq over 27 stages of development, ranging from early embryo to adult males and females. Raw read counts were obtained using the online ReCount resource (Frazee et al., 2011). We focus here on a subset of these data arising from 12 embryonic samples collected at 2-hour intervals over a 24-hour period, with a single replicate for each time point. After filtering, we obtain a subset of $n = 9524$ genes.

For both of the RNA-seq datasets described above, there are a large number of genes whose expression remains unchanged in different tissues or at different time points (i.e., whose profiles are close to the center of the simplex), and a relatively small number of genes with expression highly specific to a single tissue or time point (i.e., whose profiles are close to a vertex of the simplex); see Figure 1 for an example from the mouse neocortex data. In the remainder of the paper, we refer to the latter as observations with *highly-specific* profiles, and we will focus on proposing a clustering method able to highlight these small but important groups of genes.

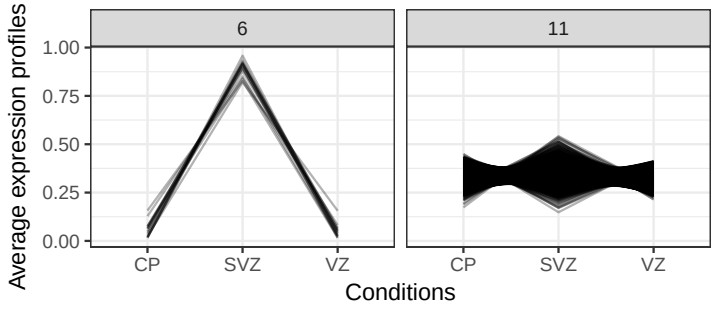

FIG 1. *Example of two groups of genes from the mouse neocortex RNA-seq data displaying normalized expression profiles that are highly-specific to the second tissue (left) and largely unchanged across tissues (right).*

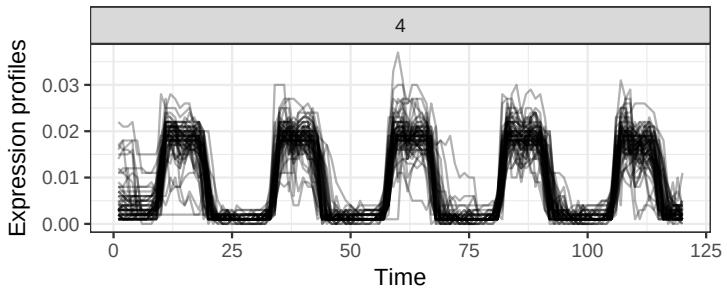

FIG 2. *Example of hourly bicycle station occupancy profiles (collected over a five-week time period) displaying a strong periodic trend in the Velib' bicycle sharing system data.*

2.2. *Bicycle sharing data*

The Velib' bicycle sharing system was introduced in Paris, France in 2007 and is made up of around 14,500 bicycles in 1230 rental stations located around the Paris metropolitan area. Velib' subscribers can check out and return bicycles from any rental stations, and a fleet of vehicules provides a daily overnight redistribution of bicycles among rental stations. Bicycle rental stations located above 60 meters of elevation belong to a special category called V+ stations.

Hourly Velib' station occupancy data, in terms of available bicycles and docks, were downloaded from a period covering 1 week (between 11am on Sunday, August 31st and 11pm on Sunday, September 7th, 2014) from open-data APIs provided by the JCDecaux company, as described in Bouveyron et al. (2015). The data correspond to count tables $\left(Y_{i,j}\right)$, representing the raw occupancy counts (i.e., the number of available bicycles) for stations $i = 1, \ldots, n$ at times $j = 1, \ldots, d$ over the course of this week. Due to the daily habits of Velib' users (e.g., going to work), we note that the station occupancy counts tend to display periodic patterns across time points (see Figure 2). In recent work, to account for differences in the size of rental stations (i.e. the number of docking points), Bouveyron et al. (2015) proposed dividing each row of counts by the total capacity of the station to yield occupancy proportions; they then clustered the rental stations in terms of the number of available docks at any given time. In our work, rather than using the fullness of each station, we instead focus on the dynamics of users' habits across time (e.g., whether bicycles tend to be rented and returned during working hours). In a similar manner to the RNA-seq data described above, we thus calculate the relative occupancy of each Velib' station across the full time period studied; this facilitates clustering stations according to specific user behaviors across time while also accounting for differences in capacity among stations. We thus calculate the occupancy profiles $(X_i)_i$, defined for all i by $X_i = (X_{i,1}, \ldots, X_{i,d})$ where

$$X_{i,j} := \frac{Y_{i,j} + 1}{\sum_{j'=1}^{d} \left(Y_{i,j'} + 1\right)}, \quad \forall j = 1, \ldots, d.$$

Note that as with the RNA-seq data, a constant of 1 is added to occupancy counts due to the presence of 0's. We thus obtain a proportion table $\left(X_{i,j}\right)$, whose rows X_i belong to the simplex $\mathcal{S}^d$.

In Section 5, we focus on the occupancy profiles from the Velib' bicycle sharing data. These data are available in the R package `funFEM` (Bouveyron et al., 2015), and correspond to counts of available bicycles in $n = 1213$ stations at each hour, during one week ($d = 183$). As weekend rental habits tend to be much different from those during weekdays, we focus in particular on the occupancy counts from Monday 12am to Friday 11pm ($d = 120$). Although clustering will be performed on the full set of data, results are visualized after summing across days for each time interval ($d = 24$) since a strong periodicity in station occupancy was observed from day to day.

3. Methods

In what follows, we consider a table of compositional data $X = (X_{i,j})$, with $i = 1, \dots, n$ and $j = 1, \dots, d$. Each row X_i of this table, which we also refer to as *profiles*, belongs to the simplex $\mathcal{S}^d$.

3.1. K-means algorithm

Let $X_1, \dots, X_n$ be a set of d-dimensional points to be clustered into K clusters. We consider the usual Euclidean norm $\|.\|_2$. Let $\mathcal{C}^{(K)} = \{C_k, k = 1, \dots, K\}$ be a partition into K clusters, and let μ_k be the mean of the cluster C_k:

$$\mu_k := \frac{1}{|C_k|} \sum_{i \in C_k} X_i,$$

where $|C_k|$ is the cardinality of cluster k. The aim of K-means is to minimize the sum of squared errors (SSE), defined for each set of clusters $\mathcal{C}^{(K)}$ by

$$\mathrm{SSE}\left(\mathcal{C}^{(K)}\right) := \sum_{k=1}^{K} \sum_{i \in C_k} \|X_i - \mu_k\|_2^2,$$

with $i \in C_k$ if $\|X_i - \mu_k\|_2 = \min_{k'=1,\dots,K} \|X_i - \mu_{k'}\|_2$. Many algorithms have been proposed to implement K-means clustering, and we consider the well-known one introduced by MacQueen et al. (1967). Note that minimizing the SSE is known to be a NP-hard problem, and as such the K-means algorithm can only converge to a local minimum. In what follows, we will consider the K-means based on the Euclidean distance, using either untransformed or transformed data.

Transformations for compositional data: Let $h : \mathcal{S}^d \longrightarrow \mathbb{R}^{d-1}$ be a differentiable and bijective function. The mapping h enables the definition of a Euclidean structure on the simplex (Pawlowsky-Glahn and Egozcue, 2001). As previously noted, the most usual transformations h in the literature for compositional data are the CLR, the ALR and the ILR (Aitchison, 1982). As these different transformations yielded similar results for the two applications considered here, we focus on the CLR : $\mathcal{S}^d \longrightarrow \mathbb{R}^d$, defined for all $x \in \mathcal{S}^d$ by

$$\mathrm{CLR}(x) := \left(\ln\left(\frac{x_1}{g(x)}\right), \dots, \ln\left(\frac{x_d}{g(x)}\right)\right), \tag{3.1}$$

where $g(x)$ is the geometric mean of x. In this case, the transformed values do not belong to $\mathbb{R}^{d-1}$ but to the hyperplane of $\mathbb{R}^d$ with normal vector $(1, \dots, 1)$. Note that there is a strict equivalence between clustering untransformed compositional data with the K-means algorithm using Aitchison's distance (Aitchison, 1982), and clustering CLR-transformed data with the K-means algorithm

using the Euclidean distance, i.e minimizing

$$\mathrm{SSE}_{\mathrm{CLR}}\left(\mathcal{C}^{(K)}\right) := \sum_{k=1}^{K}\sum_{i\in C_k}\left\|\mathrm{CLR}\left(X_i\right) - \mu_{k,\mathrm{CLR}}\right\|_2^2,$$

where $\mu_{k,\mathrm{CLR}}$ is the arithmetic mean of the CLR-transformed data belonging to cluster C_K:

$$\mu_{k,\mathrm{CLR}} := \frac{1}{|C_k|}\sum_{i\in C_k}\mathrm{CLR}\left(X_i\right).$$

Remark that $\mathrm{CLR}^{-1}\left(\mu_{k,\mathrm{CLR}}\right)$ corresponds to the center of the cluster when Aitchison's distance is used (Pawlowsky-Glahn and Egozcue, 2001).

Log Centered Log Ratio transformation: For data of moderate dimension, when a large number of coordinates have very small proportions, the CLR transformation tends to be quite sensitive to small fluctuations close to zero (see Figure 3 for an example on simulated data). This can have a strong undesired effect on clustering results (see Section 4) when a small number of observations have highly-specific profiles. To account for this phenomenon by giving more importance to coordinates with large relative values, we propose a novel extension of the CLR for compositional data called the Log Centered Log Ratio (logCLR). For all $x \in \mathcal{S}^d$, the logCLR is defined by $\mathrm{logCLR}(x) := \left(\mathrm{logCLR}\left(x_1\right), \ldots, \mathrm{logCLR}\left(x_d\right)\right)$, where for all j,

$$\mathrm{logCLR}\left(x_j\right) := \begin{cases} -\left[\ln\left(1 - \ln\left[x_j/g(x)\right]\right)\right]^2 & \text{if } x_j/g(x) \leq 1, \\ \left(\ln\left[x_j/g(x)\right]\right)^2 & \text{otherwise,} \end{cases}$$

and $g(x)$ is the geometric mean of x. The additional log term when $\frac{x_j}{g(x)} \leq 1$ accords less importance in the transformation to samples with relatively weak proportions, while the squared term facilitates the concentration of profiles close to the center of the simplex $\left(\frac{1}{d}, \ldots, \frac{1}{d}\right)$ (see Figure 3). Performing K-means clustering with this transformation amounts to minimizing

$$\mathrm{SSE}_{\mathrm{logCLR}}\left(\mathcal{C}^{(K)}\right) := \sum_{k=1}^{K}\sum_{i\in C_k}\left\|\mathrm{logCLR}\left(X_i\right) - \mu_{k,\mathrm{logCLR}}\right\|_2^2,$$

where $\mu_{k,\mathrm{logCLR}}$ is the arithmetic mean of the transformed data belonging to the cluster C_K:

$$\mu_{k,\mathrm{logCLR}} := \frac{1}{|C_k|}\sum_{i\in C_k}\mathrm{logCLR}\left(X_i\right).$$

Remark 3.1. *Note that for all positive constants p, it is also possible to replace the exponent 2 in the definition of the logCLR transformation by an exponent of p.*

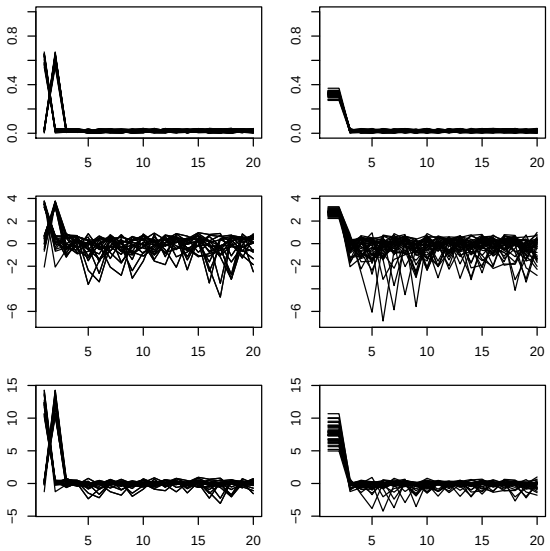

FIG 3. *Simulated highly-specific profiles on the simplex, on untransformed data (top) and after transformation using the CLR (middle) or logCLR (bottom). On the left, two clusters of profiles that are specific to a single sample (either the first or second) are included; on the right, a single cluster of profiles specific to the first pair of samples is included.*

3.2. *Choice of the number of clusters*

Many criteria exist in the literature to select the number of clusters for K-means (Caliński and Harabasz, 1974; Krzanowski and Lai, 1988), and the most common are the Gap statistic (Tibshirani et al., 2001) and the maximization of the averaged Silhouette (Kaufman and Rousseeuw, 2009). However, in our case, these methods tend to lead to the choice of an overly parsimonious model in practice. We instead focus on the selection method introduced by Fischer (2011), which is based on the minimization of a nonasymptotic penalized criterion defined for all positive integers K:

$$\begin{aligned}\text{crit}(K) &:= \sum_{k=1}^{K} \sum_{i \in C_k} \left\| h\left(X_i\right) - \mu_{k,h} \right\|_2^2 + \text{pen}\left(K\right) \\ &= \sum_{i=1}^{n} \min_{k=1,\dots,K} \left\| h\left(X_i\right) - \mu_{k,h} \right\|_2^2 + \text{pen}(K),\end{aligned}$$

where h is the transformation used (identity for untransformed data). The penalty function pen : $\mathbb{N} \longrightarrow \mathbb{R}_+$ is defined by

$$\text{pen}(K) := a_h \sqrt{Knd}.$$

The penalty term is known up to a multiplicative constant a_h, which in practice is calibrated using the slope heuristics method (Birgé and Massart, 2007;

Baudry et al., 2012) implemented in the R package `capushe` (Brault et al., 2011). The number of selected clusters is thus

$$\widehat{K} := \arg\min_{K \leq n} \text{crit}(K).$$

3.3. Conditional probabilities

By associating a clustering partition $\mathcal{C}^{(K)}$ obtained with the K-means algorithm with results obtained with the EM algorithm (Dempster et al., 1977) for spherical Gaussian mixture models, the conditional probability that each observation X_i belongs to a cluster C_k may be calculated as follows:

$$\tau_{i,k,h} := \frac{|C_k| \, \phi\left(h\left(X_i\right); \mu_{k,h}, \sigma^2_{k,h} I_d\right)}{\sum_{k'=1}^{K} |C_{k'}| \, \phi\left(h\left(X_i\right); \mu_{k',h}, \sigma^2_{k',h} I_d\right)},$$

where $\phi\left(.; \mu_{k,h}, \sigma^2_{k,h} I_d\right)$ is the density function of a Gaussian vector of mean $\mu_{k,h}$ and variance $\sigma^2_{k,h} I_d$, and $\sigma^2_{k,h}$ is the within-cluster variance:

$$\sigma^2_{k,h} = \frac{1}{|C_k|} \sum_{i \in C_k} \left\| h\left(X_i\right) - \mu_{k,h} \right\|_2^2.$$

Note that alternative model selection criteria, such as the Bayesian Information Criterion (BIC) or Integrated Completed Likelihood (ICL) criterion (Schwarz, 1978; Biernacki et al., 2000), could also be used to select the number of clusters. Nevertheless, as for the use of the EM algorithm for Gaussian mixture models, for the studied data, the BIC and ICL criteria often lead to the choice of a very large number of clusters for the K-means algorithm.

3.4. Implementation in `coseq`

We have implemented a user-friendly interface for the K-means clustering approach described above for the CLR and logCLR transformations in the R/Bioconductor package `coseq`. Model selection is provided via the slope heuristics as described above, and a variety of customizable graphics of cluster profiles may be easily generated using dedicated plotting functions. Example scripts illustrating clustering analyses for profiles have been included directly in the package vignette.

4. RNA-seq data results

We applied the K-means algorithm using the Euclidean distance described in the previous section to the two RNA-seq datasets described in Section 2 (referred to as the mouse or fly data, respectively), using either untransformed,

CLR-transformed, or logCLR-transformed profiles for $K = 2, \ldots, 40$ clusters. For each dataset and each transformation, the nonasymptotic penalized criterion described in Section 3.2 is used to identify the appropriate number of clusters $\widehat{K}$ (see Table 1); diagnostic plots for the calibration of the slope heuristics penalty as well as profile plots for all identified clusters under each transformation are shown in Appendix. We note that alternative model selection criteria, such as the Gap statistic and averaged Silhouette, led to models with very small numbers of clusters (2 or 3). As such, the following results correspond to the models selected via the nonasymptotic penalized criterion.

	Identity	CLR	logCLR
Mouse	22	20	20
Fly	20	17	21

TABLE 1

Number of clusters $\widehat{K}$ selected for each RNA-seq dataset and each proposed transformation, where model selection was performed by minimizing the penalized criterion defined in Section 3.2 and calibrated with the slope heuristics.

By examining the clusters of profiles identified in each dataset under each transformation, we may note qualitative differences in the results for each approach, in particular for genes with highly-specific profiles. In the case of the mouse RNA-seq data, comparisons among analagous profile clusters for untransformed, CLR-, and logCLR-transformed data indicate that untransformed and CLR-transformed data (Figure 4, top and middle) tend to yield a larger number of small, less-variable clusters situated at the center of the simplex, corresponding to several distinct clusters representing genes that are largely nondifferentially expressed. These two strategies also tend to yield more diffuse clusters for profiles close to the vertices of the simplex (Figure 4, bottom), particularly in genes with high expression in the SVZ tissue relative to the others. On the contrary, using the logCLR transformation with the *K*-means algorithm tends to produce tight clusters on the edges and at the vertices of the simplex (i.e., where expression in one or more of the tissues is close to zero), and a smaller number of diffuse clusters in the center of the simplex. This suggests that untransformed and CLR-transformed RNA-seq profiles tend to facilitate the identification of fine differences among nondifferentially expressed genes but do not tend to separate out genes with distinct profiles; on the other hand, the logCLR transformation leads to a large diffuse grouping of nondifferentially expressed genes and the identification of several smaller, highly specific clusters. We also note that the mouse data contained a single outlier gene which exhibited aberrant expression (i.e., very strong expression in a single replicate in each tissue); only the logCLR transformation grouped this gene alone in its own cluster.

In the case of the fly data, the distinction between the CLR and logCLR transformations is even more marked than for the mouse data. The CLR-transformed profiles with very strong relative expression in either the first or the second time point are incorrectly grouped together (Figure 5 middle, Cluster 2); these

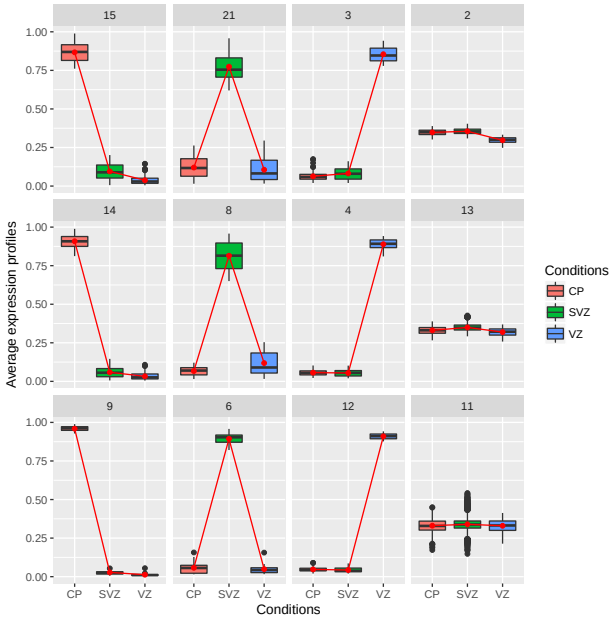

FIG 4. *Examples of clusters of genes from the mouse RNA-seq data: per-cluster normalized expression profiles of selected clusters obtained with the K-means algorithm with no transformation (top), and with the CLR (middle) and logCLR (bottom) transformations are shown. Analogous clusters for each approach have been vertically aligned. Connected red lines correspond to the mean profiles for each condition.*

two sets of highly distinct profiles are correctly separated into two clusters in the case of no transformation (Figure 5 top, Clusters 14 and 17) or the logCLR transformation (Figure 5 bottom, Clusters 4 and 6). This behavior is in fact due to the definition of the CLR transformation (see Equation 3.1). Indeed, as illustrated in Figure 3, the CLR transformation for coordinates close to 0 is highly sensitive to small fluctuations; for the K-means algorithm, this tends to have the effect of grouping together profiles with several zeros in common rather than those with a single non-zero coordinate in common (see for example Cluster 1 in Figure 5, middle). In applications where it is of interest to group together profiles with common strong distinct expression, the use of untransformed or logCLR-transformed data with K-means thus appears to be a more coherent choice.

To focus on a more quantitative measure of the differences among each method, we also calculate the per-cluster squared errors for each dataset using the Euclidean distance. More precisely, for each set of clusters $\mathcal{C}^{(K)}$, we consider the squared errors defined for all $k = 1, \dots, K$ by

$$\mathrm{SE}_h(k) := \sum_{i \in C_k} \|X_i - \mu_k\|_2^2,$$

where $\mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} X_i$. As shown in Figure 6, the main difference between

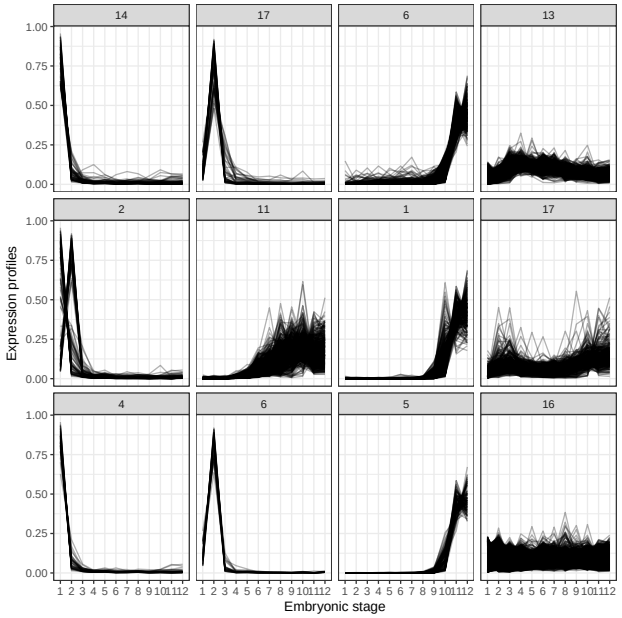

FIG 5. *Examples of clusters of genes from the fly RNA-seq data: per-cluster normalized expression profiles of selected clusters obtained with the K-means algorithm with no transformation (top), and with the CLR (middle) and logCLR (bottom) transformations are shown. Analogous clusters for each approach have been vertically aligned.*

the logCLR transformation and the other approaches is that it tends to produce a larger number of sharp clusters close to the edges or vertices of the simplex, as well as a small number of diffuse clusters situated at the center of the simplex (corresponding to the outlier points in Figure 6). As an example, Clusters $\{8, 11\}$ obtained from the logCLR-transformed mouse neocortex RNA-seq data, which corresponds to a large group of nondifferentially expressed genes, is largely made up of Clusters $\{2, 6, 10, 12, 16, 19, 22\}$ or $\{1, 3, 13, 19\}$ obtained with untransformed or CLR-transformed data, respectively. The conditional probabilities of cluster membership calculated for each approach (see the Appendix for the boxplots of conditional probabilities) indicate that these clusters tend to be made up of genes with fairly low conditional probabilities.

5. Bicycle sharing data results

As in the previous section, we apply the K-means algorithm with Euclidean distance to untransformed, CLR-, and logCLR-transformed profiles from the Velib' bicycle sharing data described in Section 2 for $K = 2, \ldots, 40$ clusters. Using the nonasymptotic penalized criterion described in Section 3.2, we select a clustering model for each of the three transformations ($\widehat{K} = 11, 11, 13$, respectively). Diagnostic plots for the calibration of the selection criterion and full

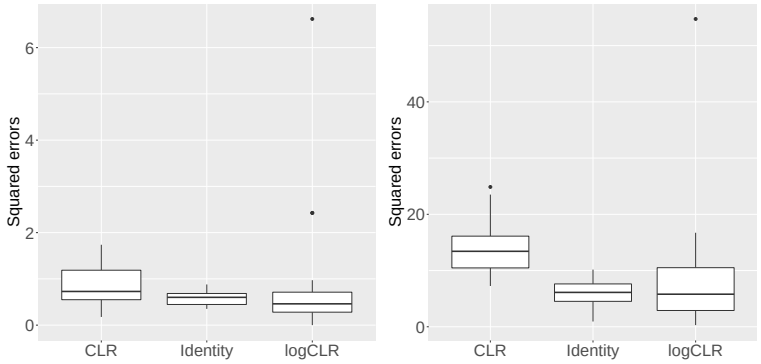

FIG 6. *Distributions of per-cluster squared errors for untransformed, CLR-, and logCLR-transformed profiles for the mouse (left) and fly (right) RNA-seq data.*

profile plots for each transformation are included in the Appendix.

In the case of the bicycle sharing data, the use of the CLR and logCLR transformations leads to the identification of tighter clusters than for untransformed data (Figure 7). As was the case for RNA-seq data, the logCLR transformation in particular leads to a small number of very tight clusters of distinct profiles (i.e., those near the edges or vertices of the simplex). These distinct profiles appear to largely correspond to rental stations with highly periodic rental patterns during the course of the week; for example, Cluster 4 (Figure 7) is made up of rental stations with close to zero available docks overnight on all weeknights.

We note that the clustering partitions obtained with the CLR and logCLR transformations are quite similar; the main differences between the two rely on the fact that the use of the logCLR leads to the choice of a larger number of clusters with slightly tighter profiles. For clarity, in the following we thus focus on the results obtained with the CLR transformation.

As shown in Figure 8, the daily cluster profiles (i.e., after summing occupancy profiles for each time point across the five days) reveal an interesting coherence with the location of the Velib' rental stations; this is all the more striking given that the spatial coordinates of the stations are not included in the model. For example, Cluster 3 (Figure 8, top left) corresponds to stations located in the historical city center of Paris (in black in Figure 9). This area of Paris is a shopping and tourist-oriented district with many restaurants and bars, which explains the fact that users appear to arrive over the course of the morning and to have prolonged departures over the course of the evening. Cluster 7 (Figure 8, top right) corresponds to stations serving users that leave for work in the morning and return home at the end of the workday. These stations (in red in Figure 9) indeed correspond to largely residential areas of Paris. On the other hand, Cluster 2 (Figure 8, bottom left) appears to contain rental stations for professionals arriving at business offices, as users tend to arrive be-

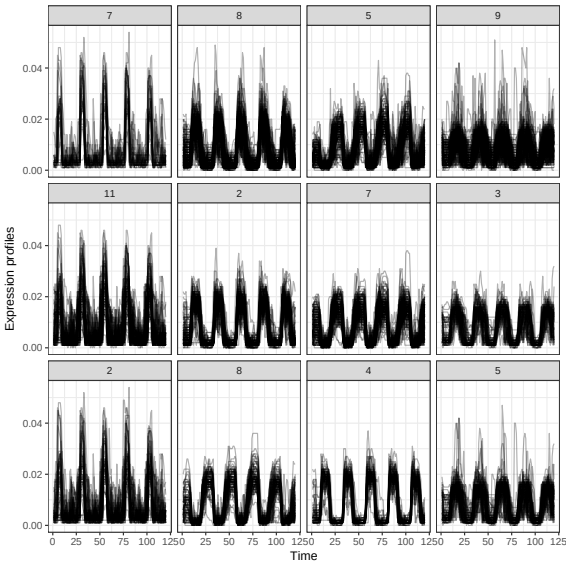

FIG 7. *Examples of weekly clusters of rental stations from the Velib' bicycle station data: available occupancy profiles of selected clusters obtained with the K-means algorithm with no transformation (top), and the CLR (middle) and logCLR (bottom) transformations are shown. Analogous clusters for each approach have been vertically aligned.*

tween 8am and 10am and leave between 6pm to 8pm. This is again confirmed by the map, where these stations (in blue) are located in main business districts such as Bercy, La Défense, and the 8th Arrondissement. Finally, Cluster 11 (Figure 8, bottom right) appears to represent a fairly atypical group of rental stations, as they fill up early in the morning, between 1am to 5am; in addition, their corresponding locations (in orange in Figure 9) are largely residential (especially in the 19th and 20th Arrondissements) and do not necessarily match up to areas of active night life. However, most of these stations belong to the special category of V+ stations located at higher elevations; as such, users tend to rent bicycles in the morning (on the downhill route) but do not return home by bicycle, thus necessitating an overnight redistribution of bicycles.

6. Conclusion and recommendations

In this work, we have described a strategy for clustering compositional data with the *K*-means algorithm and several adaptated transformations. The choice of the appropriate transformation in practice depends strongly on the type of cluster profiles that are of interest for a given context; for example, is it pertinent to separate highly-specific profiles into dedicated clusters (as appears to be the case for the RNA-seq data)? Our recommendations for the choice of compositional data transformation can be summarized into the following

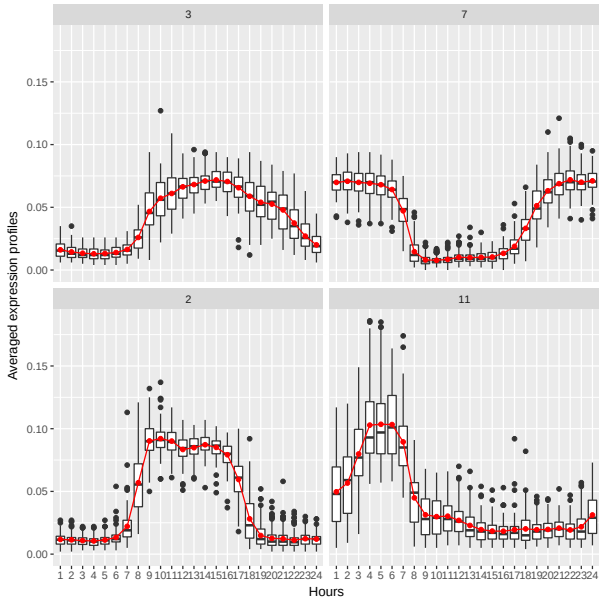

FIG 8. *Examples of daily clusters of rental stations from the Velib' bicycle station data: available occupancy profiles of selected clusters obtained with the K-means algorithm CLR transformation. Connected red lines correspond to the mean profiles for each time point.*

points: 1) if a balanced clustering is desired (i.e., clusters of roughly equal size), the K-means algorithm should be applied on untransformed profiles; 2) in order to highlight groups of highly-specific profiles, the logCLR transformation should instead be used; and 3) if small fluctuations of profile coordinates close to 0 are of primary interest, the CLR transformation may be preferred over the logCLR. Recall that in the two applications studied here, the logCLR transformation was the only option that yielded satisfactory results in all cases.

Since the data studied here originated as count tables that were transformed into compositional tables, it is worth noting that a K-means algorithm with the χ^2 distance could also have been applied on the original counts. This distance is based on a transformation that relies on the multiplication of columns by constants which are, in the case of the applications studied here, very close to one another (in the case of the RNA-seq data, this is due to the normalization of library sizes). As such, results obtained for the K-means algorithm with Euclidean distance on CLR- or logCLR-transformed profiles are very similar to those obtained for the K-means algorithm with χ^2 distance on the original counts.

Finally, several extensions of this work could be considered in future work. For instance, rather than selecting a single number of clusters for a given dataset, it could in some cases be preferable to instead construct a hierarchy of clustering for varying K. To do this, one possibility would be to partition the data

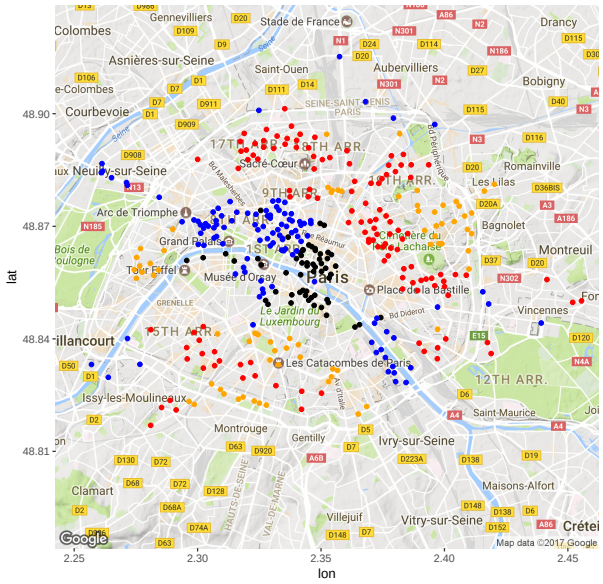

FIG 9. *Geographic location in Paris of selected rental station clusters from the Velib' bicycle station occupancy data obtained with the K-means algorithm and CLR transformation.*

into a moderate number of clusters, for example using the strategy described in this work, prior to aggregating clusters according to an appropriate criterion. Another potential extension is the integration of relevant external data in the clustering approach, such as location (for the Velib' data) or membership in functional pathways (for the RNA-seq data).

References

Aebischer, N. J., Robertson, P. A., and Kenward, R. E. (1993). Compositional analysis of habitat use from animal radio-tracking data. *Ecology*, 74(5):1313–1325.

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 139–177.

Baudry, J.-P., Maugis, C., and Michel, B. (2012). Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470.

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

Bingham, R. L., Brennan, L. A., and Ballard, B. M. (2007). Misclassified resource selection: compositional analysis and unused habitat. *Journal of Wildlife Management*, 71(4):1369–1374.

Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probability theory and related fields*, 138(1-2):33–73.

Bouveyron, C., Côme, E., and Jacques, J. (2015). The discriminative functional mixture model for a comparative analysis of bike sharing systems. *The Annals of Applied Statistics*, 9(4):1726–1760.

Brault, V., Baudry, J.-P., Maugis, C., and Michel, B. (2011). R package "capushe". https://CRAN.R-project.org/package=capushe.

Buccianti, A., Tassi, F., and Vaselli, O. (2006). Compositional changes in a fumarolic field, Vulcano Island, Italy: a statistical case study. *Geological Society, London, Special Publications*, 264(1):67–77.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

Cardot, H., Cénac, P., and Monnez, J.-M. (2012). A fast and recursive algorithm for clustering large datasets with K-medians. *Computational Statistics & Data Analysis*, 56(6):1434–1449.

Chayes, F. (1960). On correlation between variables of constant sum. *Journal of Geophysical research*, 65(12):4185–4193.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B. Methodological*, 39(1):1–38.

DeSarbo, W. S., Ramaswamy, V., and Chatterjee, R. (1995). Analyzing constant-sum multiple criterion data: A segment-level approach. *Journal of Marketing Research*, pages 222–232.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barcelo-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300.

Fietz, S. A., Lachmann, R., Brandl, H., Kircher, M., Samusik, N., Schröder, R., Lakshmanaperumal, N., Henry, I., Vogt, J., Riehn, A., et al. (2012). Transcriptomes of germinal zones of human and mouse fetal neocortex suggest a role of extracellular matrix in progenitor self-renewal. *Proceedings of the National Academy of Sciences*, 109(29):11836–11841.

Fischer, A. (2011). On the number of groups in clustering. *Statistics & Probability Letters*, 81(12):1771–1781.

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

Friedman, J. and Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Computational Biology*, 8(9):e1002687.

Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang, L., Artieri, C. G., van Baren, M. J., Boley, N., Booth, B. W., et al. (2011). The developmental transcriptome of Drosophila melanogaster. *Nature*, 471(7339):473–479.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.

Krzanowski, W. J. and Lai, Y. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, pages

23–34.

Longford, N. T. and Pittau, M. G. (2006). Stability of household income in European countries in the 1990s. *Computational statistics & data analysis*, 51(2):1364–1383.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Martín-Fernández, J., Barceló-Vidal, C., and Pawlowsky-Glahn, V. (1998). A critical approach to non-parametric classification of compositional data. In *Advances in data science and classification*, pages 49–56. Springer.

Mateu-Figueras, G., Pawlowsky-Glahn, V., and Juan José, E. (2013). The normal distribution in some constrained sample spaces. *SORT: statistics and operations research transactions, 2013, vol. 37, núm. 1, p. 29-56*.

McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Miesch, A. and Chapman, R. (1977). Log transformations in geochemistry. *Mathematical Geology*, 9(2):191–198.

Pawlowsky-Glahn, V. and Buccianti, A. (2011). *Compositional data analysis: Theory and applications*. John Wiley & Sons.

Pawlowsky-Glahn, V. and Egozcue, J. J. (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, 15(5):384–398.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 187:253–318.

Rau, A. and Maugis-Rabusseau, C. (2017). Transformation and model choice for RNA-seq co-expression analysis. *Briefings in Bioinformatics*, page bbw128.

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

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

Tauber, F. (1999). Spurious clusters in granulometric data caused by logratio transformation. *Mathematical Geology*, 31(5):491–504.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Zhou, D., Chen, H., and Lou, Y. (1991). The logratio approach to the classification of modern sediments and sedimentary environments in northern South China Sea. *Mathematical Geology*, 23(2):157–165.

Ziemann, M., Kaspi, A., Lazarus, R., et al. (2015). Digital Expression Explorer: A user-friendly repository of uniformly processed RNA-seq data. In *ComBio2015*, volume POS-TUE-099, Melbourne.

Appendix A: Results for the mouse RNA-seq data

In the following, we visualize the clusters of genes identified from the mouse neocortex RNA-seq data after applying the K-means algorithm with the identity, CLR, and logCLR transformations, as well as plots of the conditional probabilities calculated for each. We also present an example of the slope heuristic plot (obtained with the R package `capushe`) for calibrating of the nonasymptotic penalized criterion defined in Section 3.2 of the main manuscript.

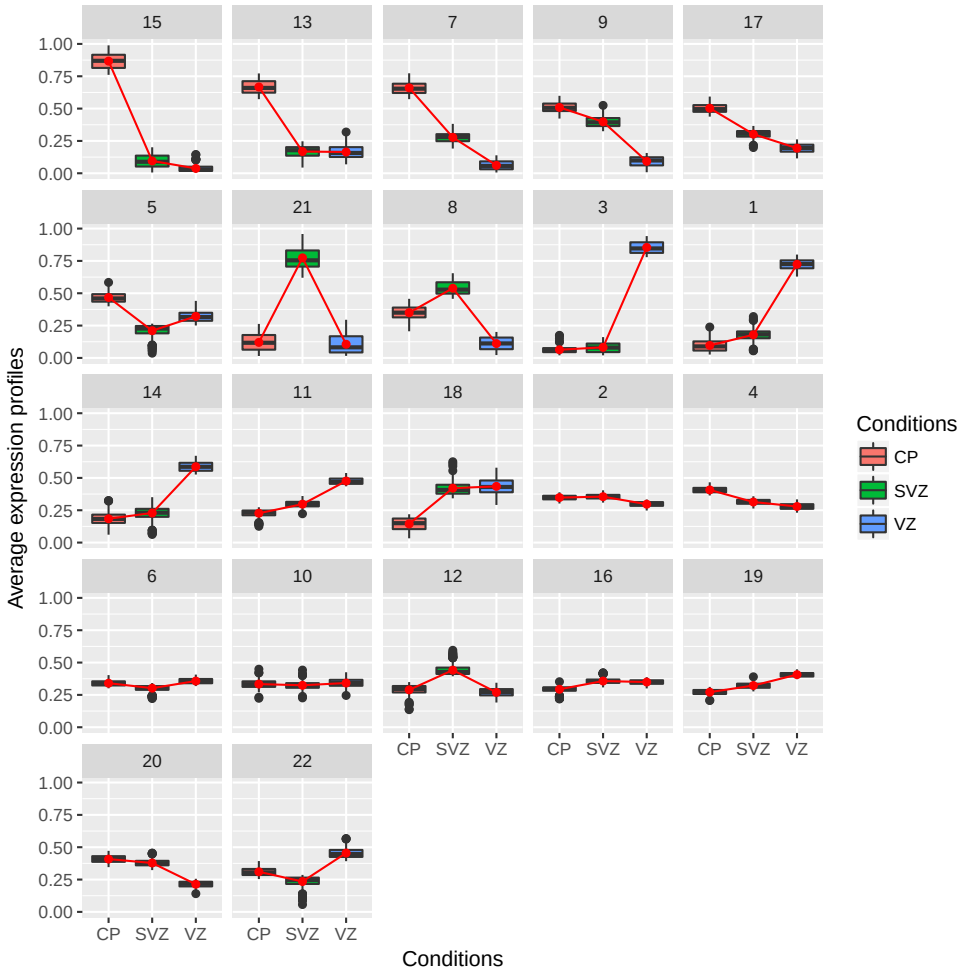

FIG 10. *Clusters of genes from the mouse neocortex RNA-seq data: per-cluster normalized expression profiles of selected clusters obtained with the K-means algorithm and no transformation. Clusters have been arranged so that those with similar average profiles are displayed next to one another. Connected red lines correspond to the mean profiles for each tissue.*

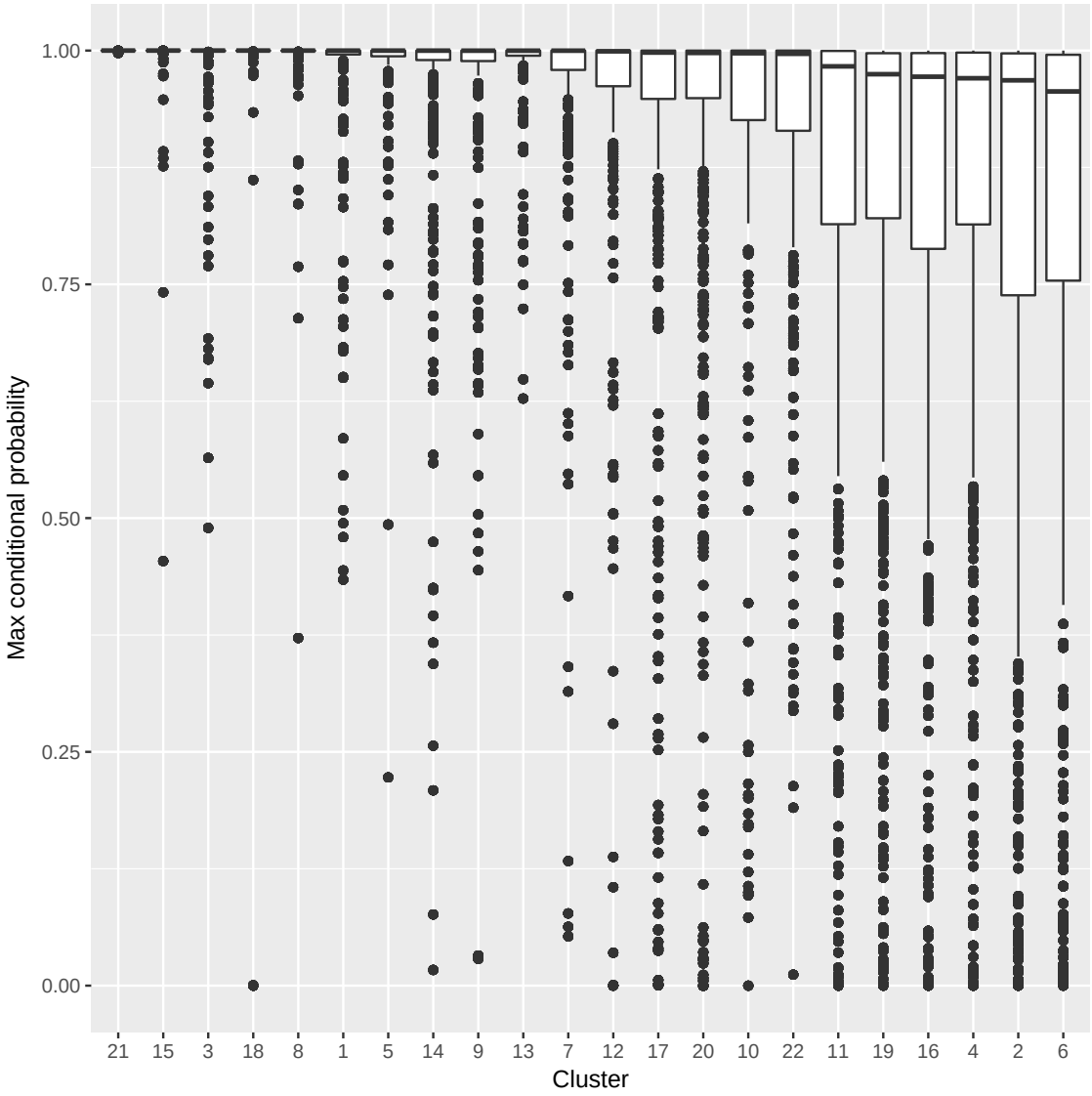

FIG 11. *Per-cluster maximum conditional probabilities from the mouse neocortex RNA-seq data obtained with the K-means algorithm and no transformation. Clusters have been sorted by median per-cluster conditional probabilities.*

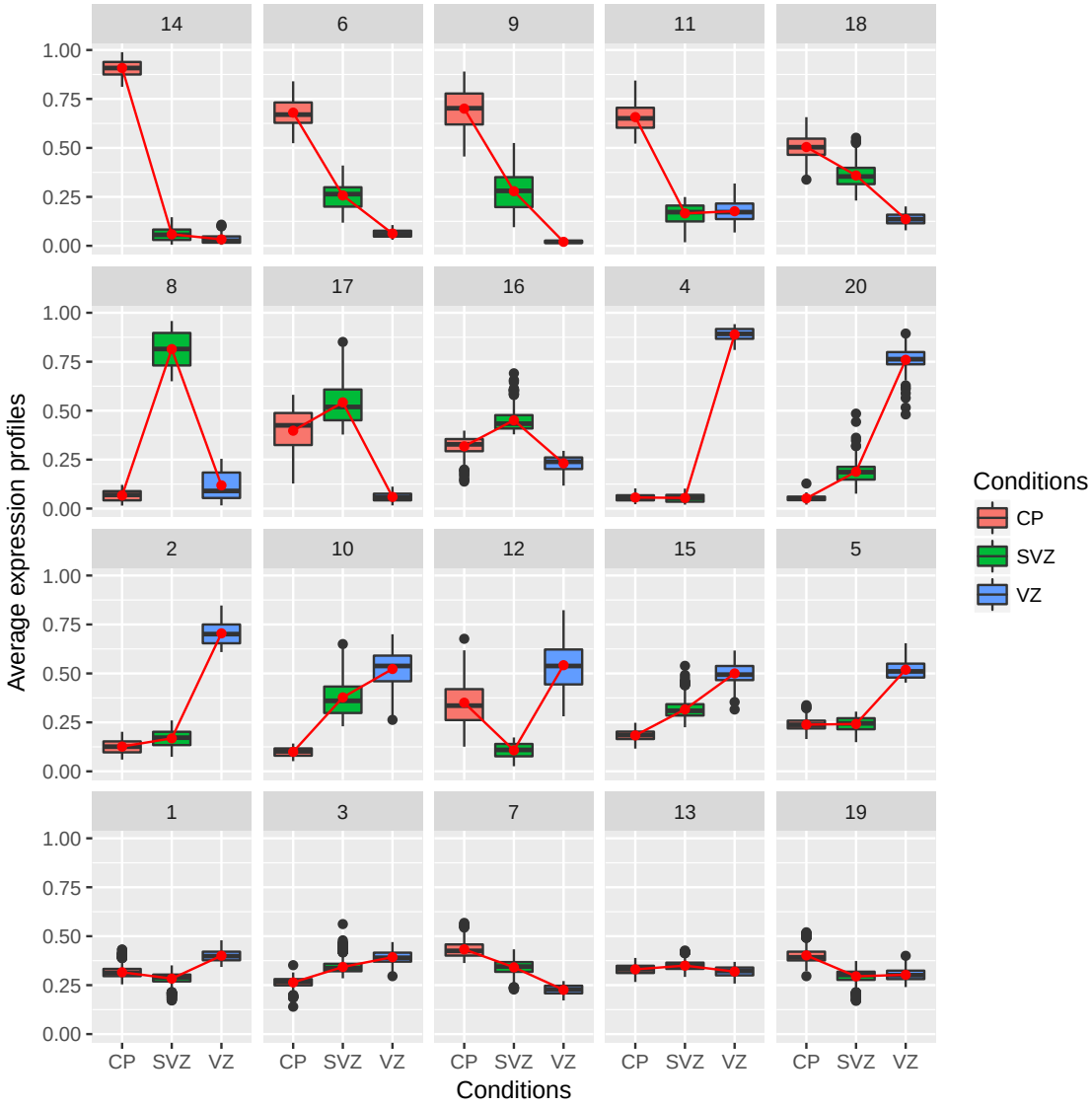

FIG 12. *Clusters of genes from the mouse neocortex RNA-seq data: per-cluster normalized expression profiles of selected clusters obtained with the K-means algorithm and CLR transformation. Clusters have been arranged so that those with similar average profiles are displayed next to one another. Connected red lines correspond to the mean profiles for each tissue.*

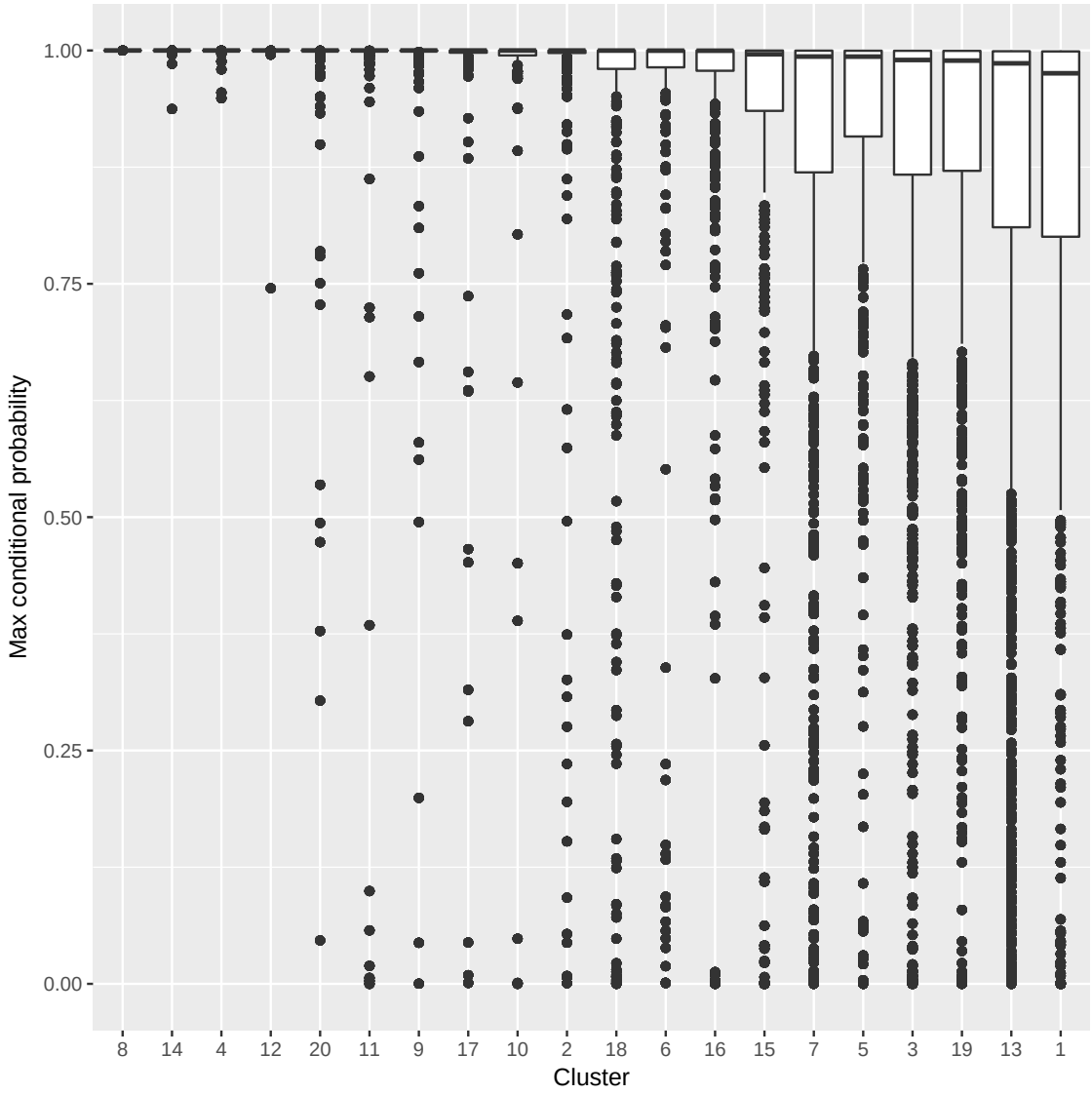

FIG 13. *Per-cluster maximum conditional probabilities from the mouse neocortex RNA-seq data obtained with the K-means algorithm and CLR transformation. Clusters have been sorted by median per-cluster conditional probabilities.*

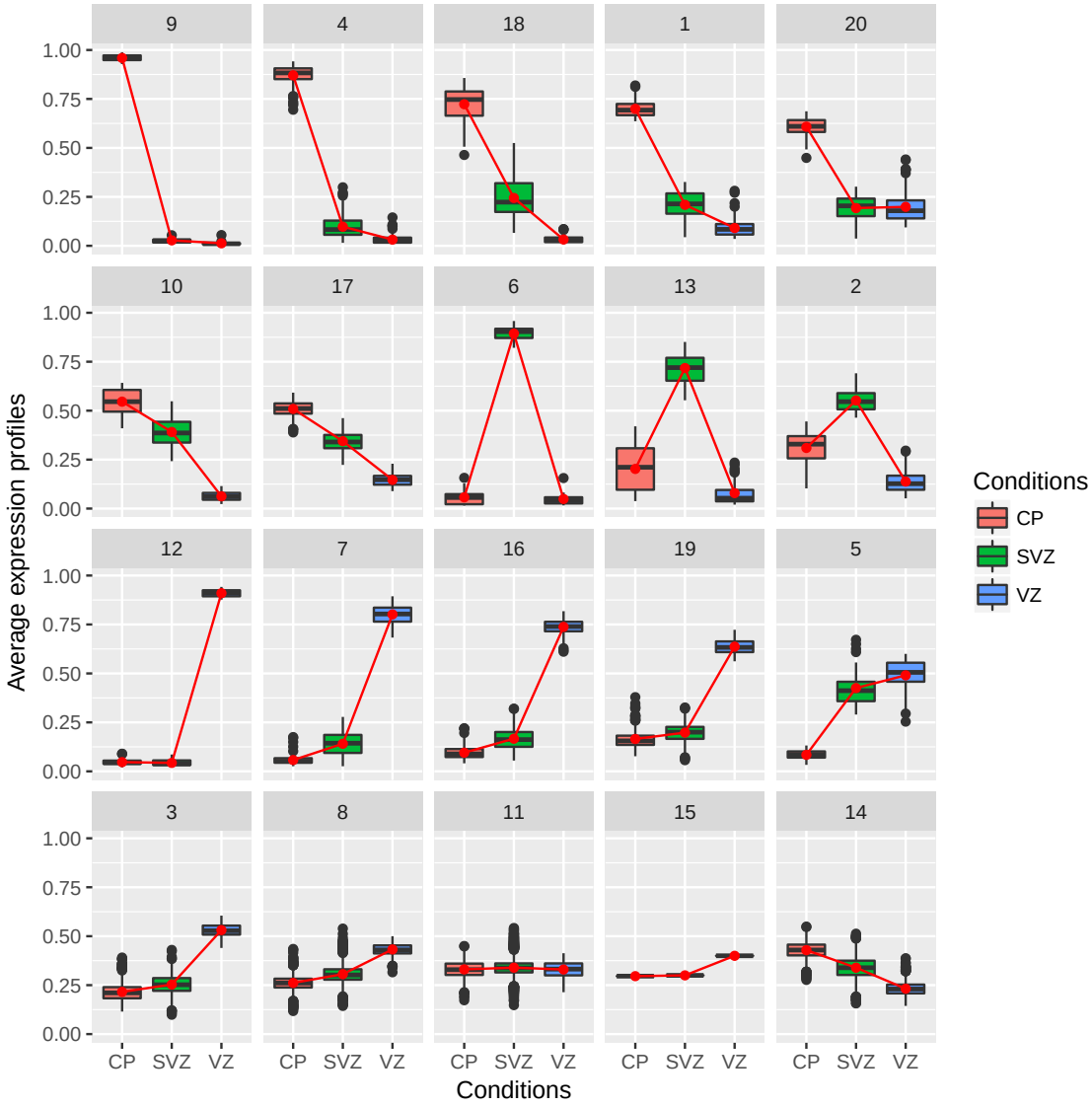

FIG 14. *Clusters of genes from the mouse neocortex RNA-seq data: per-cluster normalized expression profiles of selected clusters obtained with the K-means algorithm and logCLR transformation. Clusters have been arranged so that those with similar average profiles are displayed next to one another. Connected red lines correspond to the mean profiles for each tissue.*

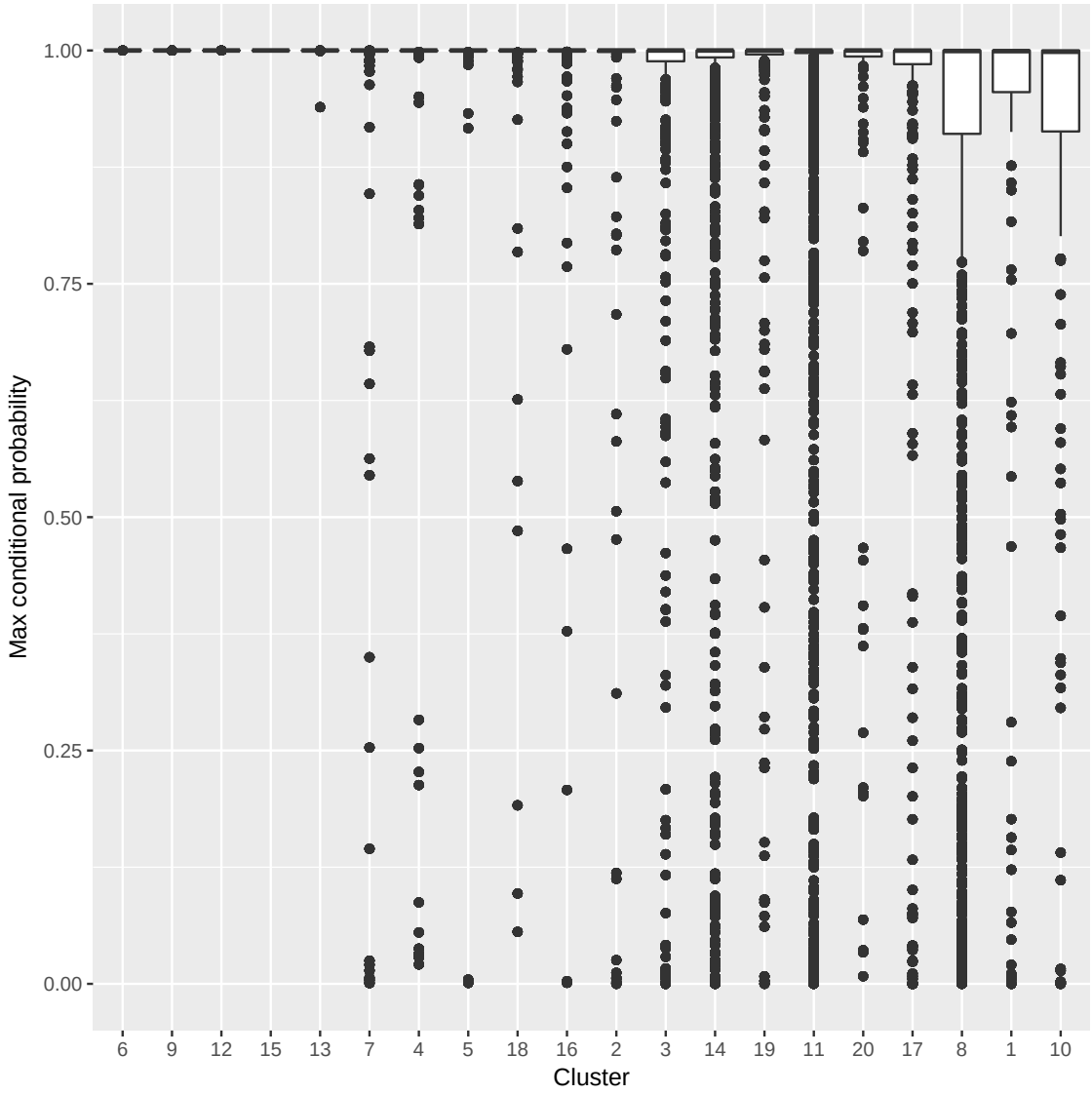

FIG 15. *Per-cluster maximum conditional probabilities from the mouse neocortex RNA-seq data obtained with the K-means algorithm and logCLR transformation. Clusters have been sorted by median per-cluster conditional probabilities.*

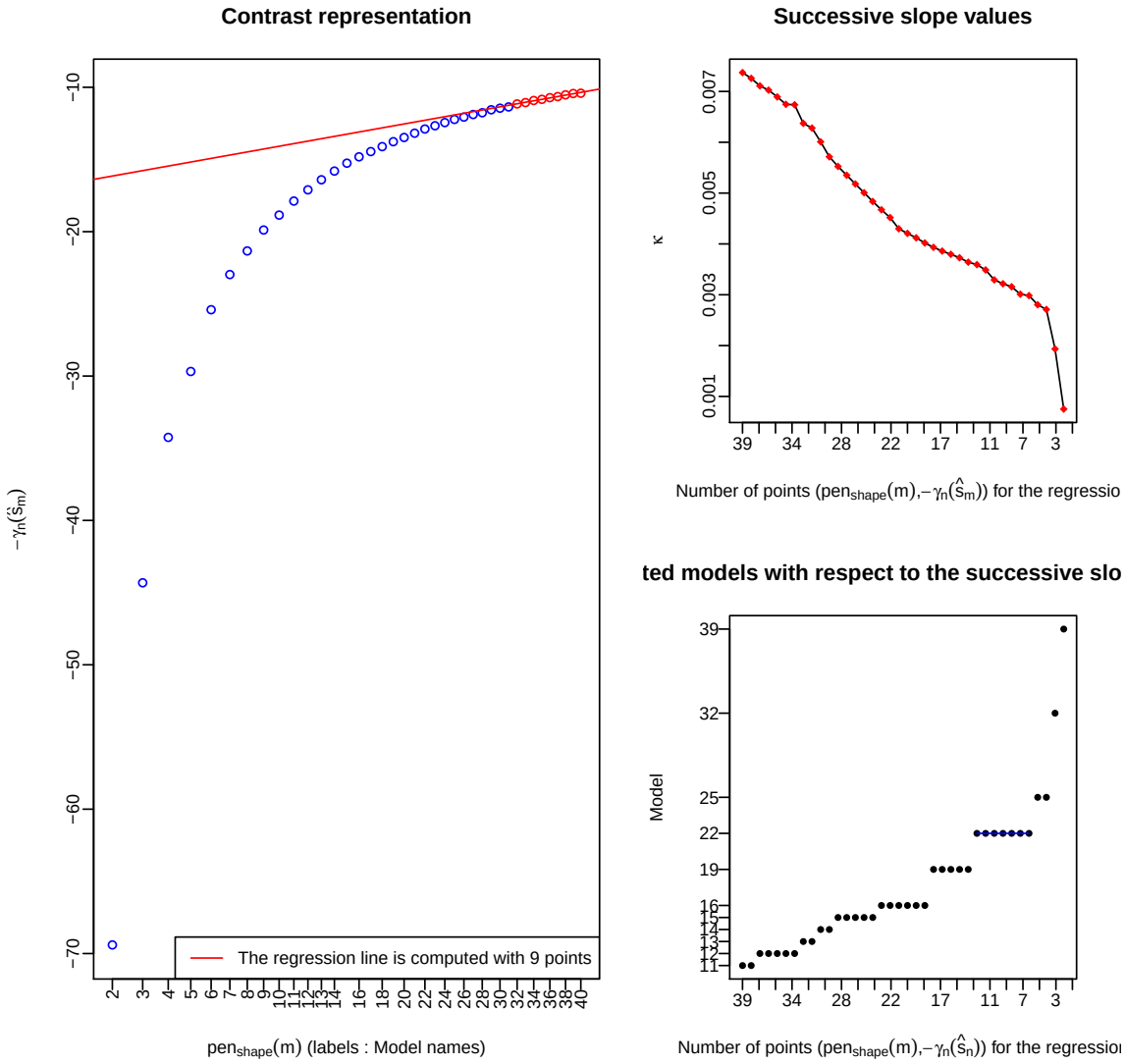

FIG 16. *Example of slope heuristics for clustering the mouse neocortex RNA-seq data with the K-means algorithm and no transformation.*

Appendix B: Results for the fly embryonic RNA-seq data

In the following, we visualize the clusters of genes identified from the fly embryonic RNA-seq data after applying the *K*-means algorithm with the identity, CLR, and logCLR transformations.

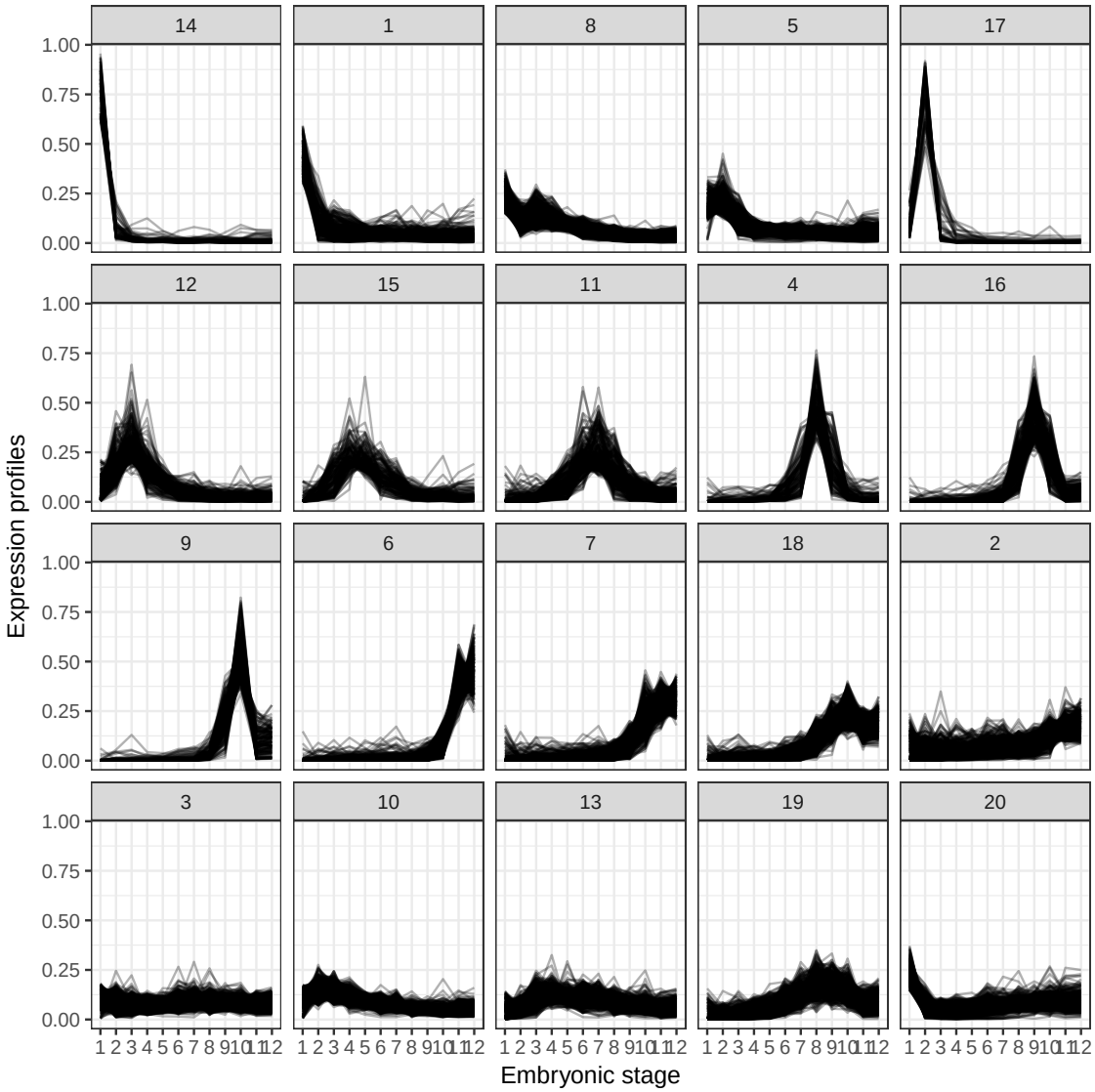

FIG 17. *Clusters of genes from the fly embryonic RNA-seq data: per-cluster normalized expression profiles of selected clusters obtained with the K-means algorithm and no transformation. Clusters have been arranged so that those with similar average profiles are displayed next to one another.*

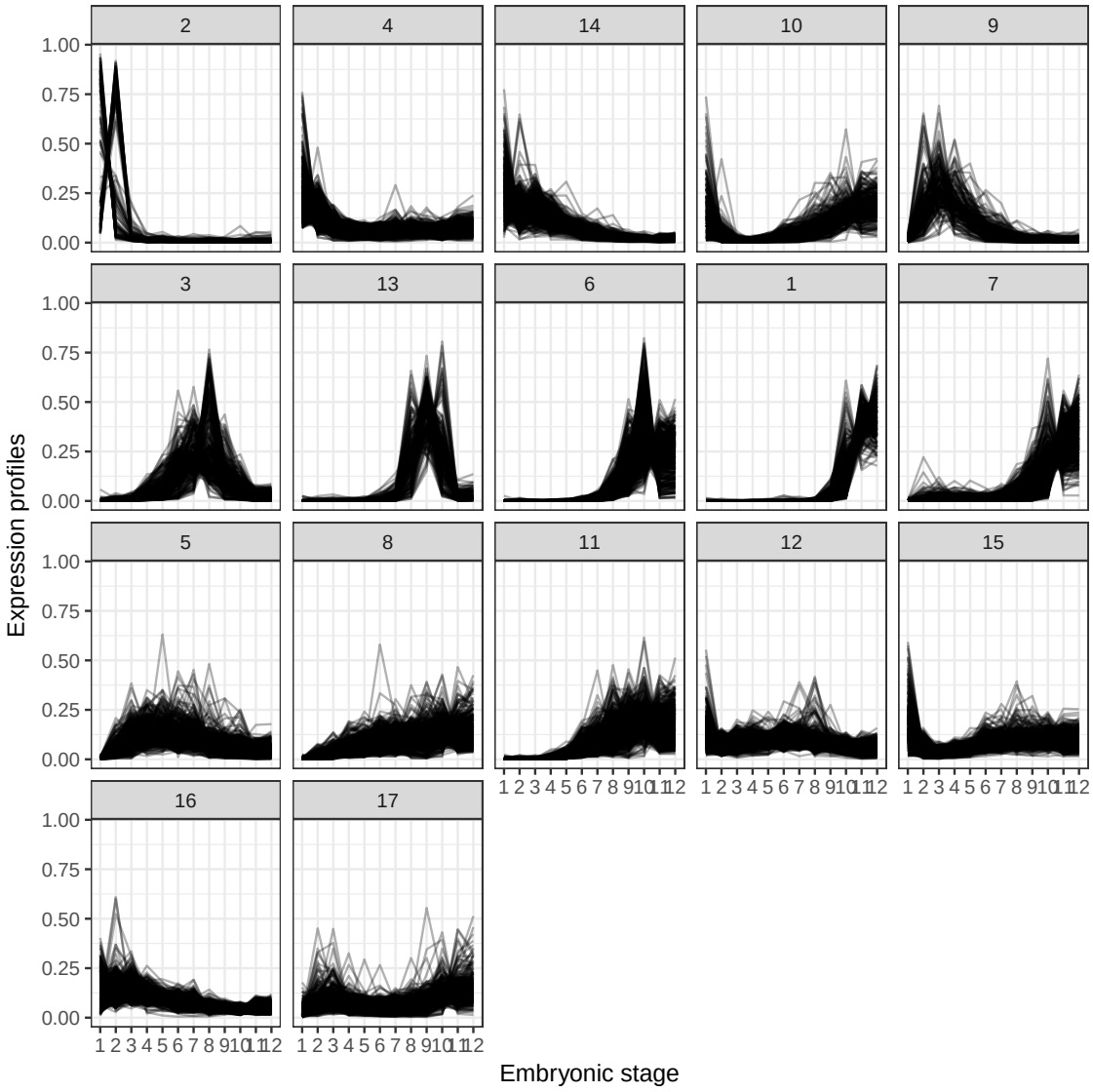

FIG 18. *Clusters of genes from the fly embryonic RNA-seq data: per-cluster normalized expression profiles of selected clusters obtained with the K-means algorithm and CLR transformation. Clusters have been arranged so that those with similar average profiles are displayed next to one another.*

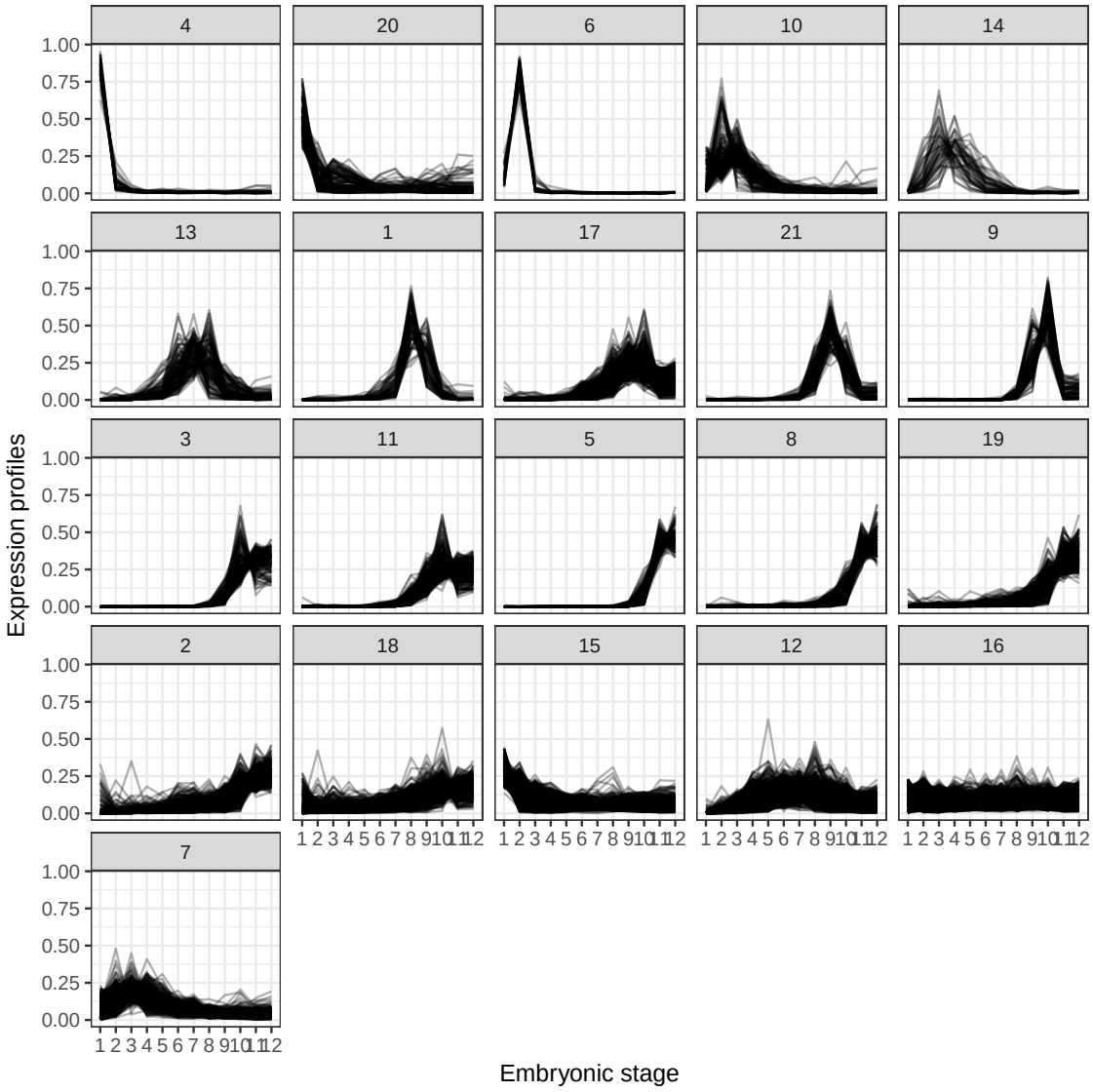

FIG 19. *Clusters of genes from the fly embryonic RNA-seq data: per-cluster normalized expression profiles of selected clusters obtained with the K-means algorithm and logCLR transformation. Clusters have been arranged so that those with similar average profiles are displayed next to one another.*

Appendix C: Results for the Velib' bicycle sharing system data

In the following, we visualize the clusters of genes identified from the Velib' bicycle sharing system data after applying the *K*-means algorithm with the

identity, CLR, and logCLR transformations.

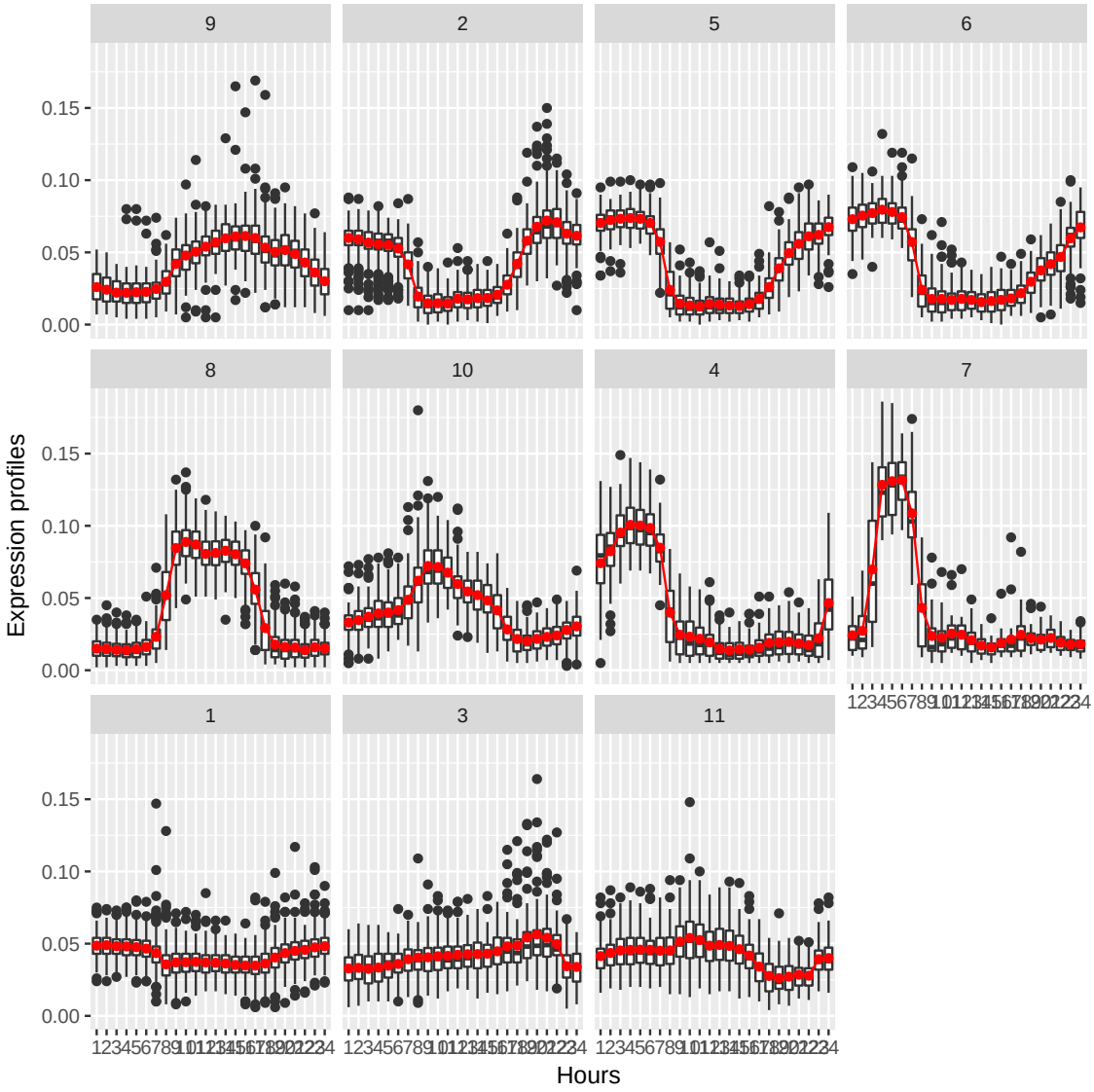

FIG 20. *Clusters of genes from the Velib' bicycle sharing system data: per-cluster occupancy profiles of selected clusters obtained with the K-means algorithm and no transformation. Clusters have been arranged so that those with similar average profiles are displayed next to one another. Connected red lines correspond to the mean profiles for each time point.*

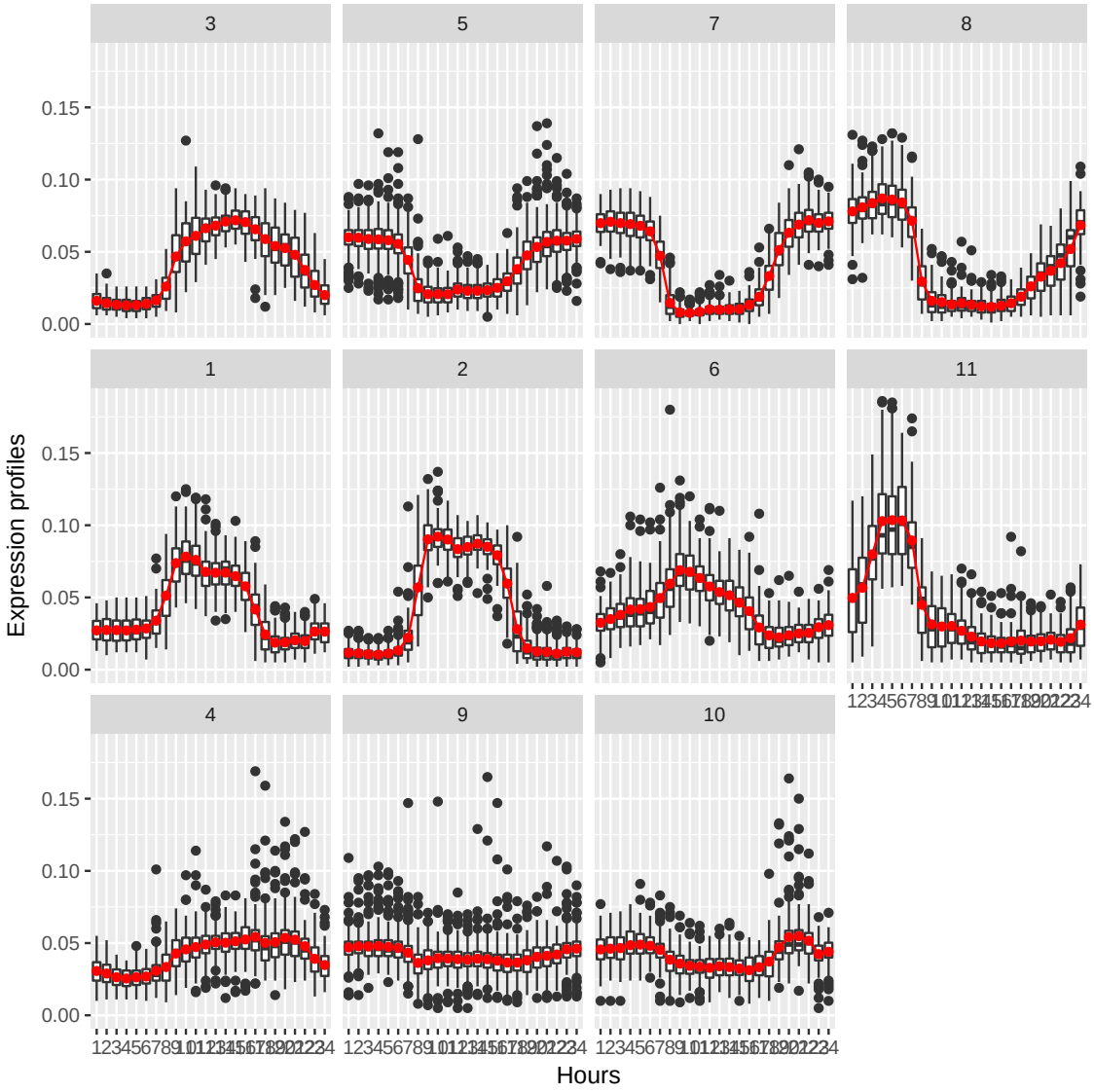

FIG 21. *Clusters of genes from the Velib' bicycle sharing system data: per-cluster occupancy profiles of selected clusters obtained with the K-means algorithm and CLR transformation. Clusters have been arranged so that those with similar average profiles are displayed next to one another. Connected red lines correspond to the mean profiles for each time point.*

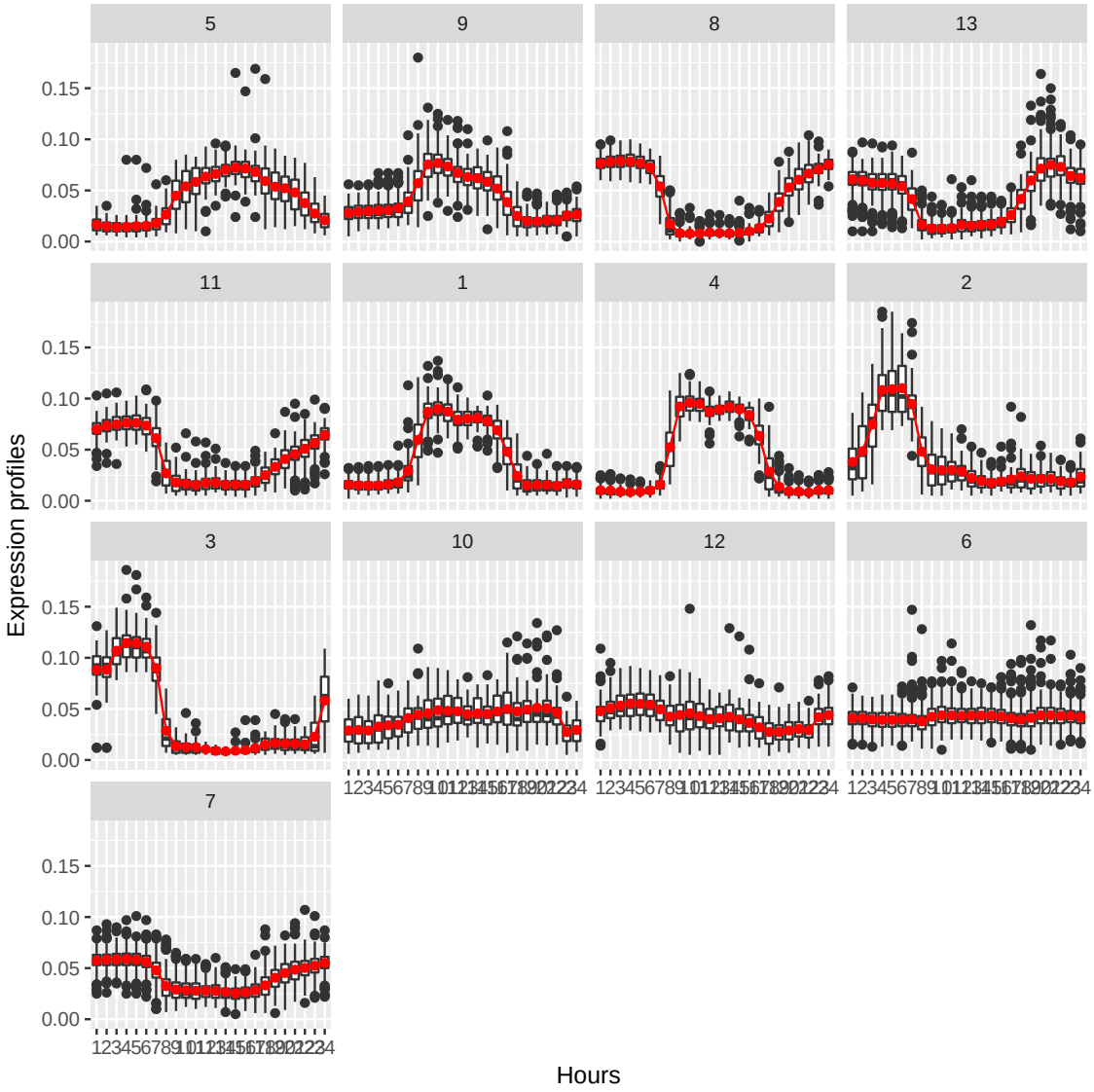

FIG 22. *Clusters of genes from the Velib' bicycle sharing system data: per-cluster occupancy profiles of selected clusters obtained with the K-means algorithm and logCLR transformation. Clusters have been arranged so that those with similar average profiles are displayed next to one another. Connected red lines correspond to the mean profiles for each time point.*