Determination of sets of covariating gene expression using graph analysis on pairwise expression ratios

Emmanuel Curis, Cindie Courtin, Pierre Alexis Geoffroy, Jean-Louis Laplanche, Bruno Saubamea, Cynthia Marie-Claire

Bioinformatics, 35(2), 2019, 258–265
doi: 10.1093/bioinformatics/bty629
Advance Access Publication Date: 13 July 2018
Original Paper

Gene expression

Determination of sets of covariating gene expression using graph analysis on pairwise expression ratios

Emmanuel Curis[1,2,3,*], Cindie Courtin[2], Pierre Alexis Geoffroy[2,4], Jean-Louis Laplanche[2], Bruno Saubaméa[2] and Cynthia Marie-Claire[2]

[1]Laboratoire de Biomathématiques, Faculté de Pharmacie de Paris, Université Paris Descartes, Sorbonne Paris Cité, Paris F-75006, France, [2]UMR S-1144 INSERM—Université Paris Descartes—Université Paris Diderot, Paris F-75006, France, [3]Service de Bioinformatique et Information Médicale, Hôpital Saint-Louis, AP-HP, Paris F-75012, France and [4]Pôle de Psychiatrie et de Médecine, Addictologique GH Saint-Louis – Lariboisière – F. Widal AP-HP, Paris F-75475, France

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on January 12, 2018; revised on June 19, 2018; editorial decision on July 9, 2018; accepted on July 12, 2018

Abstract

Motivation: RNA quantification experiments result in compositional data, however usual methods for compositional data analysis [additive log ratio (alr), centered log ratio (clr), isometric log ratio (ilr)] do not apply easily and give results difficult to interpret. To handle this, a method based on disjoint subgraphs in a graph whose nodes are the quantified RNAs is proposed. Edges in the graph are defined by lack of change in ratios of the corresponding RNAs between conditions.

Results: The methods is suited for qRT-PCR and RNA-Seq data analyses, and leads to easy-to-interpret, graphical results and the identification of set of genes that share a similar behavior when the studied condition changes. For qRT-PCR data, it has better statistical properties than the common $\Delta\Delta C_q$ method.

Availability and implementation: Construction of all pairwise ratio analysis *P*-values matrix, and conversion into a graph was implemented in an R package, named SARP.compo. It is freely available for download on the CRAN repository. Example R script using the package are provided as Supplementary Material; the R package includes the data needed. One of these scripts reproduces the Figure 2 of this paper.

Contact: emmanuel.curis@parisdescartes.fr

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

1 Introduction

Determination of gene-expression level is a key step in understanding cell function under physiological or pathological conditions. Gene-expression level is often defined as the amount of corresponding ribonucleic acid (RNA) transcript that is present in the cell. In differential expression experiments, these amounts are to be compared between several conditions, such as cell type, tissue, pathological versus physiological condition, before and after treatment for example.

Several methods are used to determine these RNA amounts: quantitative reverse transcription polymerase chain reaction (qRT-PCR), microarrays and RNA-seq are amongst the most used, with amounts given as arbitrary units. Over the last two decades qRT-PCR has become the method of choice for quantification of RNA molecules. Its simplicity of use and versatility in terms of starting material has allowed its implementation to gene-expression studies not only in research but also in diagnostic tests, forensic or biotechnology applications. However, even if it became the 'gold standard',

due to technical variations, a large number of published results are not replicated (Bustin and Nolan, 2017). Microarray have been extensively used but it is now outclassed by a more sensitive and reproducible technique: RNA-Seq (Costa *et al.*, 2013; Xu *et al.*, 2016). However, qRT-PCR remains the method of choice to confirm results from RNA-Seq or microarray data.

The number of different transcripts that can be simultaneously quantified varies from a few tens, for qRT-PCR, to several thousands, for RNA-seq and microarrays. These three techniques are based on different physical backgrounds, however in all three methods the result is a number that is assumed to be proportional to the (molar) amount of a specific kind of RNA in a complex mixture of several thousands of different RNAs. However, in all three methods, an experimental constraint is made on the total (mass) quantity of RNA in the sample. Indeed, each step of the isolation of the total RNA to be analyzed involves a loss (e.g. unspecific binding to the plastic materials, columns, pipettes, etc.). Therefore, after biological sample pre-processing a fixed amount of RNA (1 μg of total RNA for example) is taken to perform quantification in all three methods. Despite being necessary for practical reasons, this constraint has strong consequences on the results that can be obtained. After detailing these consequences, this paper presents a method to overcome these limitations and several applications, with a special emphasis on qRT-PCR.

1.1 Experimental model

In a given experimental condition, c, the cell culture contains a mixture of K different RNA molecules, of which $K^* < K$ will be quantified. The value of interest is $q_{i,c}$, the quantity of RNA i in this condition. It corresponds to a mass $m_{i,c} = q_{i,c}\,\mathbb{M}_i$, where $\mathbb{M}_i$ is the molar mass of RNA i. Hence, RNA i represents a molar fraction of $x_{i,c} = \frac{q_{i,c}}{\sum_{k=1}^{K} q_{k,c}}$ of the mixture, and a mass fraction of $y_{i,c} = \frac{m_{i,c}}{\sum_{k=1}^{K} m_{k,c}} = \frac{q_{i,c}\,\mathbb{M}_i}{\sum_{k=1}^{K} q_{k,c}\,\mathbb{M}_k}$.

Before performing the RNA quantification, a given mass of RNA, $\mathfrak{M}_c$, is extracted. This step changes neither the mass fraction, $y_{i,c}$, nor the molar fraction, $x_{i,c}$, (assuming the sample is homogeneous before extraction and all RNA have the same extraction probabilities), but the mass becomes $m^*_{i,c} = \mathfrak{M}_c\, y_{i,c}$ and the amount, $q^*_{i,c} = \frac{m^*_{i,c}}{\mathbb{M}_i} = \frac{\mathfrak{M}_c}{\sum_{k=1}^{K} \mathbb{M}_k\, q_{k,c}}\, q_{i,c}$.

The quantification step quantifies the amount of RNA *i after* extraction, that is $q^*_{i,c}$: the quantification result is $d_{i,c} = f_{i,c}(q^*_{i,c})$, where $f_{i,c}$ is a function depending on the quantification method, expected to be monotonous. Ideally, $f_{i,c}$ is linear: $f_{i,c}(q^*_{i,c}) = \lambda_{i,c}\, q^*_{i,c}$; only this case will be considered hereafter. Hence, experimental expression levels are given by Equation (1)—this expression shows that any change in *any* of the RNA amount, even not amongst the K^* that are quantified, will modify the quantification of *all* RNAs. In other words, *changes in quantified amounts are not directly interpretable*.

$$d_{i,c} = \lambda_{i,c} \frac{\mathfrak{M}_c}{\sum_{k=1}^{K} \mathbb{M}_k\, q_{k,c}}\, q_{i,c} \tag{1}$$

1.2 Compositional data methods

This limitation comes from the extraction step, which turns amounts into molar (or mass) fractions and constrains the sum of all masses to $\mathfrak{M}_c$. This is typical of problems dealing with compositions of a mixture (here, a RNA mixture) and is frequent in various fields of chemistry. Recently, compositional data analysis received focused attention for the analysis of microbial flora composition (Friedman and Alm, 2012; Gloor *et al.*, 2016; Silverman *et al.*, 2017; Tsilimigras and Fodor, 2016). It leads to so-called compositional data; various methods were proposed to overcome this difficulty (Aitchison, 1982, 1984; Filzmoser *et al.*, 2009). It is well established, however, that whatever the method is, only relative changes can be determined (Aitchison, 2003) and that usual statistical methods do not apply directly on these data.

These methods are based on the following ideas: first, compositional data are described by the K-dimensional vector of the molar (or mass) fractions, $\mathbf{x}_{i,c} = (x_{1,c}, \ldots, x_{K,c})$, which is constrained to belong to a portion of an hyper-plane in $[0,\ 1]^K \subset \mathbb{R}^K$. Second, this vector is transformed in a vector in $\mathbb{R}^{K-1}$ without constraint and, hopefully, independent components, on which usual statistical analysis tools can be used.

The most simple method is to select a reference RNA, let's say $i = 1$, divide all values by $x_{1,c}$, and take the logarithm: $\mathbf{z}^{\mathrm{a}}_c = \left(0, \log\frac{x_{2,c}}{x_{1,c}}, \ldots, \log\frac{x_{K,c}}{x_{1,c}}\right)$ of which the first column is useless. This is called the *additive log ratio* (alr) method. However, it does not completely remove the correlations induced by the constraint on $\mathbf{x}_c$, hence is mainly suited for multivariate analysis of the results. In such a context, the whole results do not depend on the choice of the reference RNA (Aitchison, 2003); however, results for invididual RNAs may depend on this choice. In addition, interpretation is not as straightforward as would be expected (Pawlowsky-Glahn and Egozcue, 2016).

An alternative method is to divide each component by the geometric mean of all components, $x^{\mathrm{gm}}_c = \sqrt[K]{\prod_{k=1}^{K} x_{k,c}}$: $\mathbf{z}^{\mathrm{c}}_c = \left(\log\frac{x_{1,c}}{x^{\mathrm{gm}}_c}, \ldots, \log\frac{x_{K,c}}{x^{\mathrm{gm}}_c}\right)$. Known as *centered log ratio*, it gives a K-dimensional vector, hence strong correlations remains between the components. In addition, in typical gene-expression experiment, not all of the components of $\mathbf{x}_c$ are known, which makes this transformation intractable.

The *isometric log ratio* (ilr) method is more complex, leading to a $(K-1)$-dimensional vector $\mathbf{z}^{\mathrm{i}}_c = (z_k)_{1 \le k < K}$ with $z_k = \sqrt{\frac{K-k}{K-k+1}} \log \frac{x_{k,c}}{\sqrt[K-k]{\prod_{j=k+1}^{K} x_{j,c}}}$. Since, in general, K is not known, this method cannot be used in expression experiments. With this method, each individual component of the result vector mixes information from all original components. Consequently, this method is only suited to detect whole changes in the composition, but cannot answer questions about changes in specific components. Hence, even if possible, it would be of little interest for RNA quantification experiments, where questions are on individual RNAs.

2 Materials and methods

As seen above, direct interpretation of observed amounts, $d_{i,c}$, is not possible whatever the experimental method, because results are compositional data. Besides, traditional compositional data analysis method either are difficult to interpret and suited to overall comparisons without trying to identify which component changes, or do not apply, because the total number of components (K) is unknown, several components are not quantified ($K^* < K$) and results are almost always expressed in arbitrary units ('Quantification Point' for qRT-PCR, 'counts' for RNA-seq), so that the total sum that imposes the constraint is also unknown.

2.1 Expression ratios can be interpretated

Let consider the ratio of two RNA amounts in a given condition, $r_{i,i',c} = \frac{q_{i,c}}{q_{i',c}}$. This ratio is unchanged by the extraction process:

$\frac{q^*_{i,c}}{q^*_{i',c}} = \frac{q_{i,c}}{q_{i',c}} = r_{i,i',c}$. Hence, any change in the ratio after sample processing can be tracked down to a change of the ratio in the original sample.

The quantified amount ratio are then, $r^o_{i,i',c} = \frac{d_{i,c}}{d_{i',c}} = \frac{\lambda_{i,c}\, q_{i,c}}{\lambda_{i',c}\, q_{i',c}}$. Let assume that the proportionality constants, $\lambda_{i,c}$ and $\lambda_{i',c}$, do not change between experimental conditions—a condition which is almost always assumed, since any violation to it would prevent any comparison of the results, anyway. Then, changes in quantified amount ratios are due only to changes in the ratio in the original sample. However, the change magnitude may be different unless $\lambda_{i',c} = \lambda_{i,c}$—that is, quantification yield is the same for all RNAs—a much more stringent hypothesis that will not be made in the following paper.

Hence, the whole set of (quantified) ratios, or of their logarithm, is the natural dataset to retrieve information about the effects of the experimental conditions on the gene-expression levels. This set can be seen as a $K^* \times K^*$ matrix (limiting to the effectively quantified ones). Since $r_{i,i',c} = 1/r_{i',i,c}$ and $r_{i,i,c} = 1$, the matrix of the logarithm of the ratios is antisymmetric and it is sufficient to consider its upper-right or lower-left part only.

2.2 Statistical model for expression ratio

Because of the multiplicative shape of the expressions above, a log-linear model is best suited for the analysis. In this model, it is assumed that the column-vector $q_{i,\bullet}$, where $\bullet$ stands for 'all experiments', is the realization of a random vector $\mathbf{Q}_{i,\bullet}$, such that $\log \mathcal{Q}_{i,\bullet} = \mathbf{X}\beta_i + \varepsilon_{i,\bullet}$, where $\mathbf{X}$ is a matrix giving the value of all required covariates describing the changes between experiments ('design matrix'), β_i the vector of effects of these covariates and $\varepsilon_{i,\bullet}$ a random vector of null expectation. The $\varepsilon_{i,\bullet}$ vector is often assumed to be Gaussian and with covariance matrix $\sigma^2\, \mathbf{I}_{K^* \times K^*}$, but these assumptions are not required for the developments below.

Using the notations defined above, and assuming each lowercase value is the realization of a random variable given by the same letter in uppercase, we then have for quantified amounts $\mathbf{D}_{i,\bullet} = \frac{\Lambda_{i,\bullet}\, \mathfrak{M}_\bullet}{\sum_{k=1}^{K} \mathbb{M}_k \mathbf{Q}_{k,\bullet}}\, \mathbf{Q}_{i,\bullet}$. Here, $\Lambda_{i,\bullet}$ is a column-vector that gives the proportionality constant for each experiment and can also be described by a log-linear model, $\log \Lambda_{i,\bullet} = \mathbf{X}^\lambda\, \beta^\lambda_i + \varepsilon^\lambda_{i,\bullet}$: this allows to account for eventual changes in quantification conditions between experiments, either 'intentional' (β^λ_i) or resulting from uncertainties in the experiment set up ($\varepsilon^\lambda_{i,\bullet}$).

Consequently, the column-vector of the logarithm of the ratio of two RNAs is given by

$$\begin{aligned}\log \mathbf{R}_{i,i',\bullet} &= (\log \Lambda_{i,\bullet} + \log \mathbf{Q}_{i,\bullet}) - (\log \Lambda_{i',\bullet} + \log \mathbf{Q}_{i',\bullet}) \\ &= \mathbf{X}^\lambda(\beta^\lambda_i - \beta^\lambda_{i'}) + \mathbf{X}(\beta_i - \beta_{i'}) + \varepsilon_{\mathbf{i},\mathbf{i}',\bullet}\end{aligned}$$

where $\varepsilon_{\mathrm{i},\mathrm{i}',\bullet} = \varepsilon^\lambda_{i,\bullet} - \varepsilon^\lambda_{i',\bullet} + \varepsilon_{i,\bullet} - \varepsilon_{i',\bullet}$.

From this generic model, several observations can be drawn:

- using ratios allows to eliminate the variability coming from the uncertainty in the extraction of the total RNA mass (the $\mathfrak{M}_\bullet$ term disappears),
- using ratios allows to cancel out systematic changes in quantification methods between experiments, like changes in sequencing depth in RNA-Seq or changes in total quantity of fluorescent probe in qRT-PCR, as far as these effects are the same for all RNAs—that is, if $\beta^\lambda_i = \beta^\lambda_{i'}$; however, it somehow doubles the uncertainty introduced by quantification steps (term $\varepsilon^\lambda_{i,\bullet} - \varepsilon^\lambda_{i',\bullet}$ in $\varepsilon_{\mathrm{i},\mathrm{i}',\bullet}$),
- using ratios, no absolute change can be determined, but one can determine if two RNAs experience the same change in amount (that is, $\beta_i = \beta_{i'}$, except eventually for the intercept) or not ($\beta_i \neq \beta_{i'}$).

This model applies to virtually any experimental setting, from the most simple comparison between two experimental conditions to complex designs with longitudinal follow-up of gene expression and several cofactors—as much as the linear model applies to such designs. Examples of such complex designs will be given in the applications part. Inbetween, the simplest case will be considered.

2.3 Simplest case formalization

The simplest case is, in fact, a typical differential expression experiment: two conditions A and B are compared, and several experiments are done in each condition. For sake of simplicity, the case of two experiments for each condition will be detailed here; it immediately generalizes to any number of experiment for each condition. Assuming the two results in condition A are given first, the model gives

$$\log \mathbf{Q}_{i,\bullet} = \begin{pmatrix} \log \mathcal{Q}_{i,1} \\ \log \mathcal{Q}_{i,2} \\ \log \mathcal{Q}_{i,3} \\ \log \mathcal{Q}_{i,4} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \mu_{i,A} \\ \delta_{i,B} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \\ \varepsilon_{i,3} \\ \varepsilon_{i,4} \end{pmatrix}$$

and $\mu_{i,A}$ is the average amount (in log scale) of RNA i in condition A and $\delta_{i,B}$ the change in amount of RNA i induced by condition B. A similar expression can be given for $\mathbf{\Lambda}_{\mathrm{i},\bullet}$. Then, the model for the ratio of RNAs i and i' is given by (using the notation $\Delta\xi = \xi_i - \xi'_i$)

$$\log \mathbf{R}_{i,i',\bullet} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \Delta\mu^\lambda_A + \Delta\mu_A \\ \Delta\delta^\lambda_B + \Delta\delta_B \end{pmatrix} + \varepsilon_{i,i',\bullet}$$

In that model, the effect of condition B on the quantified amount ratio is given by the term $\Delta\delta^\lambda_B + \Delta\delta_B$, in which only $\Delta\delta_B$ contains the relevant, biological information: if $\Delta\delta_B = 0$, then changing A in B does not change the $\frac{q_i}{q'_i}$ ratio—either expressions of genes i and i' are not modified, or they are modified in the same way.

This assumes that $\Delta\delta^\lambda_B$ is null, that is that the change in (the log of) the proportionality constant of the quantification method between conditions A and B is the same for RNAs i and i'—that is, $\delta^\lambda_{i,B} = \delta^\lambda_{i',B}$. Note that this does *not* assume that these constants are the same for the two conditions (that would require $\delta^\lambda_{i,B} = \delta^\lambda_{i',B} = 0$), nor that they are the same for the two RNAs (that would require $\Delta\mu^\lambda_A = 0$ in addition).

2.4 Building groups of equivalent genes

The nature of expression experiments imposes that only changes in ratio can be interpreted. The proposed model allows to test if a given ratio changes between two conditions, using classical statistical methods related with the linear model, either assuming the Gaussian assumption or not. However, interpretation of the results is not straightforward, especially when the number of quantified RNAs, K^*, increases.

To help interpretation, we propose a network-based visualization and analysis method. The idea is as follow:

1. consider the K^* RNAs as the K^* nodes of a fully connected graph.

2. update the connexion matrix of this graph using the results of the tests of the ratio: nodes i and i' are disconnected if the test shows a significant change of the ratio $\frac{d_i}{d_{i'}}$.
3. find groups of strongly connected nodes.

Of note, an alternative approach could be considered to build the graph: starting from a fully disconnected graph and adding connections only between nodes with significant lack of change in the ratio. This would however require the use of equivalence tests and the *a priori* definition of what is a relevant minimal change in the ratio. As this approach is currently seldomly used in biology, only the approach with significant changes will be presented in the applications.

With this approach, if two RNAs are in the same group, their ratio is unchanged so they share the same kind of adaptation to the difference between conditions. If they are in distinct groups, their ratio change between the conditions, so one of them has its amount modified in a different way than the other—either in different directions, or in the same direction but with a different magnitude.

At the group levels, if there is more than one group in the final result, it can be said that at most one of them correspond to a group of RNAs of which the amounts are not influenced by the condition. All other groups correspond to RNAs whose amount is modified by the condition.

Since any method to detect groups could be used on the graph, and any definition of what means a 'significant change' could be used to disconnect nodes, this approach is very general. However, since these choices can have a great impact on the results, they should be decided before analysis. A few guidelines about that are given in the application part. Typically, three classes of methods exist: partitioned graphs, cliques detections and communities detections.

A graph is partitioned if it is possible to find two set of nodes ('subgraphs') such that there is no edge between any node of one set and any node of the other set. In other words, the graph is made by the assembly of two disjoint graphs.

A clique is a subset of nodes that are fully connected, that is any node in the clique is connected to all other nodes of this clique. Maximal cliques correspond to the cliques with the highest number of nodes. Non-overlapping maximal cliques can be good candidates of groups of different behavior. Maximal cliques may however overlap, that is have nodes in common, which makes the interpretation difficult. In addition, finding maximal cliques is a *NP*-complete problem, which means that the computation time exponentially increases with the size of the graph, K^*: such an approach is currently intractable with graphs of more than a few dozens of nodes.

Communities detection algorithms allow to avoid both difficulties. A community is a set of nodes that are more connected between them than with other communities in the graph. Several mathematical formalizations exist, leading to a different algorithm; reviewing them is outside the scope of this paper. Some methods are very time-consuming [like modeling the network as a spin glass and finding its ground state, Reichardt and Bornholdt (2006), or modularity optimization, Brandes *et al.* (2008)], other are less [like Clauset *et al.* method based on the modularity index, Clauset *et al.* (2004)], and different methods often lead to different results for the same graph. Hence, interpretation of the result may be difficult.

A special case of both communities and maximal cliques detection is the detection of disconnected subgraphs. When it occurs, each subgraph correspond to a distinct community, and it contains maximal cliques that are distinct than those of other subgraphs, and the interpretation is straightforward.

2.5 Adjusting *P*-values

Typical usage would be to use a statistical test at a given, prespecified level (for instance, $P_{i,i'} < 0.05$), for each ratio, and disconnect two edges if the test is significant for the corresponding ratio. The question of selecting the test level is then crucial to ensure a conclusion with a correct type I error, P.

This is a difficult question, since the relevant P is for a sentence like 'the graph has two communities' and not for each individual test. Several aspects play in opposite directions: multiple testing problems, correlations between tests, exact formulation of the hypothesis tested... To illustrate this, let assume, that the result is 'one node is isolated' (let us say node 1) in a K-nodes graph, and one would like to be confident at the 0.05 level that it is not a chance result. To isolate this node, *all* the tests of the $K-1$ ratios between this node and the other nodes should be significant; that is, multiplicity correction is not necessary from this point of view—assuming $K-1$ independent tests, all under the null hypothesis, the probability of all of being simultaneously significant would be $P(\cap_{i=2}^{K-1}\{p_{1,i} < 0.05\}) = \prod_{i=2}^{K-1} P(p_{1,i} < 0.05) = 0.05^{K-1}$, well below the desired level, and one could consider to use $0.05^{\frac{1}{K-1}}$ as the level of individual tests. However, the tests are not completely independent because they all share the same denominator, so that would inflate the real type I error. Besides, the test of 'is the node isolated' should be performed for each of the K nodes, hence the type I error for a given node should be, here, corrected for multiplicity. But since the tests are not independent at all, usual corrections like Holm or Bonferroni corrections would overcorrect. Things are even less clear when the results are expressed as communities.

Simulations may help to calibrate the choice of the $P_{i,j}$-value cut-off for individual ratio tests. Assuming that all nodes are in the same class as the null hypothesis (which means that either the experimental conditions induce no change on the RNA levels, or—less likely—that all of them change in exactly the same way), Figure 1 shows the evolution of the type I error with the number of nodes and the choice of the $P_{i,j}$ cut-off, when detecting disjoint sub-networks (see Supplementary Material figure_01.R for the corresponding code). This figure illustrates that the error decreases quickly with the number of nodes (that is, of tested RNAs) and, for instance, for an experiment with 15–20 different RNAs as typical in qRT-PCR, individual test should be conducted with $P_{i,j} < 0.25$ to maintain an overall type I error (of falsely detecting more than one group) around 0.05.

Similar simulations should be done for other definition of RNA group detection.

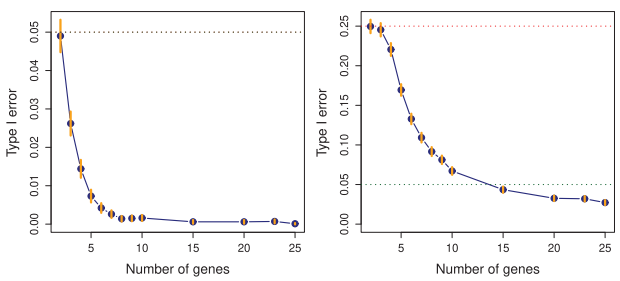

Fig. 1. Evolution of type I error of detecting two completely disjoint groups of RNAs with the number of RNA in the experiment, for individual tests of ratio changes with $P_{i,i'} < 0.05$ (left) or $P_{i,i'} < 0.25$ (right). Upper line: individual tests level α_0; lower dotted line: $P = 0.05$). Vertical segments are ~95% confidence intervals of the type I error estimated from 10 000 simulations

3 Applications to qRT-PCR

In qRT-PCR, the number of quantified RNAs, K^*, is typically small, <100 and typically not more than a few tens. In most cases, reference RNAs are included amongst the K^* RNA quantified. The result is then a 'small' graph, and even time-consuming methods to detect groups can be used. Typically, this means than maximal cliques can be searched for, and that all community detection algorithm can apply. However, since partitioned graphs are the easiest to interpret, this will be the criterion used in the application shown.

Quantification is made by selective amplification of target sequences, assuming more or less complex models (Carr and Moore, 2012; Čikoš and Koppel, 2009; Platts *et al.*, 2008). Roughly, along amplification, fluorescence increases, leading to a S-shaped curve, and RNA is quantified as the abscissa of a special point on this curve: the cycle at which the curvature of the amplification curve is maximal (quantification point, C_q), the first cycle that gives a detectable fluorescence (cycle threshold, C_t), the position of first derivative maximum (crossing point, C_p)... This quantification can be interpreted as $\log d_{i,c}$. We will not consider here the various models that may be used to convert fluorescence signal into arbitrary unit, quantified RNA level—as all other analysis methods, our proposed methods expects reliable quantifications, whatever their origin. See for instance (Ruijter *et al.*, 2013; Spiess *et al.*, 2015, 2016; Tellinghuisen and Spiess, 2014) for discussions and comparisons of quantification methods.

The traditional analysis method is the so-called '$\Delta\Delta C_t$' method (Livak and Schmittgen, 2001), where one Δ correspond to the division of $d_{i,c}$ by the quantified amount of a reference gene $d_{i_r,c}$ (normalization step), and the second to the difference of the logarithms between two conditions.

This method only applies when two conditions are compared. In that case, it is equivalent to the 'additive log ratio' method in compositional data, and shares the same limitations. Compared to the 'all ratios' approach, it gives only one line (or one column) of the ratios matrix. Hence, it only allows to detect RNAs that behave differently from the reference gene. Since the reference gene RNA level is expected not to be affected by the condition tested, it may detect RNAs with changed expression, but does not allow to classify them in different groups. Besides, the fact that the reference gene RNA is not modified is a strong assumption that is rarely verified in all samples and conditions tested, and selecting appropriate reference genes is a complicated matter (Ling and Salvaterra, 2011; Radonić *et al.*, 2004; Xu *et al.*, 2015).

To limit the impact of unexpected variations in reference gene amounts, the use of the geometric mean of the expression levels of several reference genes is advised (Vandesompele *et al.*, 2002). However, interpretation of such ratios is not straightforward, unless all the reference genes share the same behavior (and in that case, using several genes would add little value). This neither overcomes the other limitations.

Our method can be seen as a generalization of the $\Delta\Delta C_t$ method that bypasses all the limitations given above: any experimental design can be handled and interpretation of the resulting graph gives all the available information in the data. In addition, using several reference gene allows to test the hypothesis that they indeed have the same variation: they should belong to the same group. Because this method has less assumptions and is less sensitive to unexpected changes between experiments, it is expected to increase reproducibility of the conclusions obtained, expressed as groups of genes, in agreement with MIQE Guidelines, point 8.1 'Normalization' (Bustin *et al.*, 2009).

3.1 Application to a complex two factors design

This example considers the data published in Geoffroy *et al.* (2017), that also presents the rational for the experiment. Briefly it aimed to compare expression levels in cells of two kind of patients suffering from bipolar disorders and treated by lithium, a reference treatment for this pathology: patients who resolve their symptoms using lithium ('excellent responders', ER, $n = 20$) and patients who do not ('non-responders', NR, $n = 16$)—see the publication for the classification criteria. Lymphoblastoid cell lines from these patients were cultivated with and without lithium for 4 days (original experiment also includes 8 days and 2 days, that are not considered here), then qRT-PCR is performed on a set of 17 candidate RNA (transcripts of circadian genes) and 2 reference genes (*HPRT* and *SDHA*). In this design, RNA levels may change between patients (factor *P*), between ER and NR in average (factor *G*) or because of lithium (factor *L*). It is expected that either ER and NR patients should have different RNA levels at baseline, or that lithium effect on RNA level should differ between ER and NR patients, hence the G: Li interaction term is of interest. *P* is a nuisance, but important, factor.

To analyze this complex design, a linear mixed effects model is suitable, where the patient is a random effect on the intercept. This writes, using a reference-level (ER patients, without lithium) contrast coding, and after log-transformation, as

$$\begin{aligned} Q_{i,p,l,g} &= \mu_{i,0} + U_{i,p} + \\ &\quad \delta_{i,\mathrm{NR}}\, 1_{g=\mathrm{NR}} + \delta_{i,\mathrm{Li}^+}\, 1_{l=\mathrm{Li}^+} + \\ &\quad \delta_{i,\mathrm{I}}\, 1_{g=\mathrm{NR}}\, 1_{l=\mathrm{Li}^+} + \\ &\quad \varepsilon_{i,p,l,g} \end{aligned}$$

where p is the patient, $U_{i,p}$ its random effect for RNA i, $\mu_{i,0}$ the average level of RNA i (in log-space) for ER patients in absence of lithium, $\delta_{i,\mathrm{NR}}$ the difference between average level of RNA i between NR and ER patients without lithium, δ_{i,Li^+} the effect of lithium on RNA i level in ER patients and $\delta_{i,\mathrm{I}}$ the *additional* effect of lithium on RNA i level in NR patients, compared to ER patients. $1_{y=Y}$ is an indicator function, that $=1$ if $y = Y$ is true for the given data, $=0$ otherwise, and $\varepsilon_{i,p,l,g}$ is the residual error term.

With this model, terms of interest are $\delta_{i,\mathrm{NR}}$ (which is non-null if and only if ER and NR patients have different basal levels of RNA i) and $\delta_{i,\mathrm{I}}$ (which is non-null if and only if lithium has different effects on RNA i level between ER and NR patients).

When applied to ratio of RNA i and j levels, the model becomes (with $\Delta\xi = \xi_i - \xi_j$)

$$\begin{aligned} R_{i,j,p,l,g} &= \Delta\mu_0 + \Delta U_p + \\ &\quad \Delta\delta_{\mathrm{NR}}\, 1_{g=\mathrm{NR}} + \Delta\delta_{\mathrm{Li}^+}\, 1_{l=\mathrm{Li}^+} + \\ &\quad \Delta\delta_{\mathrm{I}}\, 1_{g=\mathrm{NR}}\, 1_{l=\mathrm{Li}^+} + \\ &\quad \Delta\varepsilon_{i,p,l,g} \end{aligned} .$$

Hence, testing the model coefficients will say if two RNAs have the same *change* in basal level between ER and NR ($\Delta\delta_{\mathrm{NR}} = 0$) or the same differential lithium effect between ER and NR patients ($\Delta\delta_{\mathrm{I}} = 0$). Consequently, RNAs can be grouped using either the first or the second criteria, leading to two different graphs. Since reference genes were used at three different dilutions, the graph contains $17 + 2 \times 3 = 23$ nodes, hence a $P < 0.25$ cut-off should be used to have $\alpha < 0.05$, according to simulation results presented in Figure 1. However, since *P*-values were obtained through Wald tests that assume an infinite sample size and are known to be liberal because of this, a conservative $P < 0.20$ cut-off value was used, leading to the graphs given in Figure 2. All analyses were done using R (R Core Team, 2016) (Supplementary Material figure_02.R). The model was fitted using restricted maximum-likelihood with the lmer function (package lme4) (Bates *et al.*, 2015). The graph was built

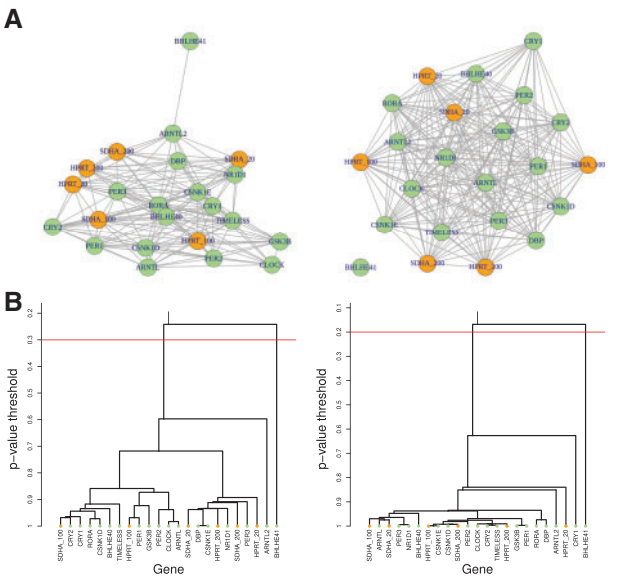

Fig. 2. (**A**) Graphs of candidate circadian genes in qRT-PCR, for basal differences (left) and differential lithium effects (right) between ER and NR patients. Circles with green background correspond to candidate circadian genes, circles with orange background to reference genes (HPRT_x and SDHA_x). (**B**) Corresponding dendrogramms. Red line is the threshold used to obtain the graphs above

and analyzed using the igraph package (Csardi and Nepusz, 2006). Sets of genes with similar behavior were defined using disjoint subnetworks (partitioned graph). Supplementary Material exemple_shiny.R provides the code to build an interactive version of Figure 2, using shiny. Results of the analysis using normalization by the geometric average of the expression levels of the two reference RNAs are given in Table 1.

At baseline, the network displays only one graph, with all nodes being connected to at least another one. There is no evidence for any change in basal circadian gene expression between excellent responders and non-responders patients. The dendrogram shows that a minimal threshold of 0.24 would be needed to cut the link between BHLHE41 and ARNTL2, making BHLHE41 a potential candidate. The next scission needs a much higher threshold of 0.60, isolating ARNTL2, suggesting no other clear candidate. The classical analysis neither detects any change, even before any multiplicity correction.

In contrary, differential lithium effects graph shows that the BHLHE41 gene is isolated, connected with no other node. This suggests that lithium effect on this gene expression is different between excellent responders and non-responders, at the 0.05 type I error level used to select the threshold. Of note, classical analysis detects two genes (BHLHE41 and CRY1) without multiplicity correction, but none of them is kept after multiplicity correction by Holm's method, so it is not possible with the classical analysis to confirm any of them. The analysis of the dendrogram shows that the next cut occurs for a threshold of about 0.63, far above from the selected one, and this cut isolates CRY1. This suggests that CRY1 detection using uncorrected classical analysis may be a false-positive, whereas significance of BHLHE41 may be related to a true signal. To confirm this, a simulation was made to compare the two methods (see Supplementary Material tableau_02.R for details). Its results, given in Table 2, confirm that the uncorrected classical analysis has an unacceptably high false-positive rate and, despite a high power, a very low probability of finding the correct genes. At comparable type-I error rate, our method has a comparable power than the multiplicity corrected classical analysis and, when detecting something, the probability that it is correct is acceptable ($\approx \frac{2.1}{6.7} = 31\,\%$), without relying on the reference gene invariance hypothesis.

Table 1. Results of qRT-PCR analysis using the changes in ratio of RNAs of interest with the geometric average of reference RNAs ('classical analysis')

Gene	Basal		Diff. Li effect	
	Raw	Holm	Raw	Holm
ARNTL	0.505	1	0.685	1
ARNTL2	0.345	1	0.250	1
BHLHE40	0.734	1	0.924	1
BHLHE41	0.092	1	0.037	0.589
CLOCK	0.258	1	0.766	1
CRY1	0.390	1	0.006	0.098
CRY2	0.588	1	0.720	1
CSNK1D	0.934	1	0.567	1
CSNK1E	0.744	1	0.765	1
DBP	0.725	1	0.935	1
GSK3B	0.424	1	0.938	1
NR1D1	0.636	1	0.621	1
PER1	0.707	1	0.951	1
PER2	0.246	1	0.495	1
PER3	0.988	1	0.657	1
RORA	0.990	1	0.464	1
TIMELESS	0.037	0.631	0.915	1

Note: *P*-values are expressed without ('raw') and after multiplicity correction by Holm's method ('Holm').

Table 2. Comparison of analysis methods to detect that 1 gene changes between two conditions amongst 23 genes, including 6 reference genes

Method	α	$1-\beta$	γ
$\Delta\Delta Ct$	56.3 % (55.3; 57.2)	62.5 % (61.6; 63.5)	7.9 % (7.4; 8.4)
H-$\Delta\Delta Ct$	5.2 % (4.8; 5.6)	7.0 % (6.5; 7.5)	2.2 % (1.9; 2.5)
Graphs, $P = 0.30$	5.2 % (4.7; 5.6)	6.7 % (6.2; 7.2)	2.1 % (1.8; 2.4)

Notes: Results of 10 000 simulations; $\Delta\Delta Ct$: 'classical analysis', without multiplicity correction; H-$\Delta\Delta Ct$: 'classical analysis', with multiplicity correction by Holm's method; graphs: our method, with the given threshold. α: type-I error rate; $1-\beta$: power (probability of rejecting the null hypothesis that no gene-expression changes); γ: probability of correctly detecting the gene that changes, and only that gene. Values are given with their 95% confidence intervals.

4 Applications to RNA-Seq

Basically, in a RNA-Seq experiment, RNAs extracted from the culture are retro-transcribed into DNA. This DNA is splitted in small size fragments and these fragments are sequenced. Each obtained sequence is then mapped on a reference library to identify the original transcript. Amounts of a given RNA are expressed in number of times a given fragment was associated to this RNA sequence—'counts'.

The conversion of this values in quantities raises several issues, and there is numerous discussions on the methods to have meaningful and comparable values, known as normalization steps (Li *et al.*, 2015; Risso *et al.*, 2014). Discussion of these methods and issues is outside the scope of this paper; as for qRT-PCR, we will assume that a method that gives reliable $d_{i,c}$ was used. Let's just note that, since the changes in ratios do not depend on having the same quantification yield for all RNAs and all conditions, but only that these yields do not change differently between RNA between the different conditions, our method should be less sensitive to the choice of the normalization method, as far as this method still leads to an approximate linear relationship between the true amount and the quantified, normalized amount.

However, in RNA-Seq, an additional difficulty arises: quite often, several transcripts are separately quantified for a given gene, but interest is on the gene level. If gene g has n_g different transcripts, $n_g \ll K$, then its expression level could be defined as $q_c^g = \sum_{j=1}^{n_g} q_{j,c}$ (assuming its transcripts are numbered first), and the corresponding quantified amount is $d_c^g = \frac{\mathfrak{M}_c}{\sum_{k=1}^{K} \mathbb{M}_k q_{k,c}} \sum_{j=1}^{n_g} \lambda_{j,c} q_{j,c}$. Hence, the ratio of quantified amounts between genes g and g' is $\frac{d_c^g}{d_c^{g'}} = \frac{\sum_{j=1}^{n_g} \lambda_{j,c} q_{j,c}}{\sum_{j=n_g}^{n_g+n_{g'}} \lambda_{j,c} q_{j,c}}$.

This is in general different from the original ratio, unless the proportionality constants are the same for all transcripts of each genes, hence factorizes. This is however a reasonable hypothesis, since the different transcripts should be chemically close. Hence, we will not consider this difficulty in the applications below.

Since the number of different RNAs detected, K^*, is huge, the network is too large for several methods of group detection, especially the maximal cliques method and several algorithms of community detection. However, this huge number of nodes also guarantees than even high individual P-value cutoffs can be used to detect disconnected sub-networks.

Another difficulty is the discrete nature of the quantification data, at least as given by most normalization methods. This is usually handled using count process models, like Poisson law or negative binomial law (Love *et al.*, 2014) and a generalized linear model approach. However, there is no such standard model for ratio of counts data. Hence, we instead suggest the use of a log-normal law, which is often a good continuous approximation of the negative binomial law, allowing for skewness and independent position and dispersion parameters. This should work well unless very low counts are obtained, with the special case of 0 counts that is difficult to handle since it leads to infinite ratios. However, since the way to decide if two nodes are connected or not is free, this limitation can be overcome by replacing the test on counts by, for instance, a test on the proportion of values below the quantification limit.

5 Discussion

Expression data are by nature compositional, due to the way these experiments are done: a pre-defined mass of total RNA is analyzed for quantification. This is a basic limitation of these studies that remains whatever the quantification method used. Hence, it should be taken into account in the analysis and the interpretation of the results. To our knowledge, this was not yet done for microarray and RNA-Seq experiments.

In qRT-PCR experiments, this limitation is handled through the use of a set of reference, 'housekeeping', genes. This approach has several drawbacks, including the assumption that reference gene RNA levels are not modified across the various conditions, the difficulty to choose the number and the nature of these reference genes. Furthermore, it only allows a partial interpretation of the results, since it only detects candidate genes that behave differently than reference genes, but cannot easily detect that two candidate genes behave in different ways. Last, because each candidate gene is tested separately, corrections for multiple testing are needed (like Bonferroni or Holm methods), which make the detection of a signal all the more difficult, than there are different genes tested in the experiment.

The method proposed here allows to interpret the data respecting their compositional status, whereas overcoming most of these limitations. First, since groups of genes of homogeneous behavior are built, interpretation of the results is easy and can be done graphically, with the drawing of the obtained network. Second, reference genes if present are handled similarly to candidate genes. If all reference genes belong to the same final set of nodes, then assumption that they all behave similarly (in particular that they are not affected) is reasonable, and using this set as a reference 'no change' set, interpretation of the other sets can be done. However, if the reference genes are splitted between different sets of nodes, then the underlying assumption for the selection does not hold. But this does not prevent the interpretation of each set of genes as behaving differently. Last, if set of nodes are defined as disjoint subgraphs, as demonstrated by the simulations done and the application, this method keeps a good power to detect signal when the number of candidate genes increases. This is because it does not need multiplicity correction. Instead, to keep a '$P < 0.05$' threshold for the global test 'disjoint graphs do exist', individual tests should be performed with a P-value cut-off that increases with the number of nodes, ensuring a reasonable power.

Compared to the usual $\Delta\Delta Ct$ method, often used in qRT-PCR, it also presents the advantage to be usable in any experimental context, and not only in paired samples, two-conditions experiments. Since the $\Delta\Delta Ct$ is a special case of the 'reference gene' method, remarks of the previous paragraph also apply.

Concerning RNA-Seq and microarrays experiments, the method is time and place-consuming, because the number of tests to be done on the ratios increases as $\frac{K^*(K^*-1)}{2} \sim K^{*2}$, where K^* is the number of quantified RNAs—for a typical RNA-Seq experiment that detects around 16 000 different RNAs, that means around 128 million comparisons and a 16 000 nodes graph to analyze. Because of this, simulations are tedious and the optimal cut-off selection difficult. Besides, one may expect that the risk of having genes that present intermediate variations between other pairs of genes increases, meaning that disjoint subgraph are harder to find and that the power of the method is diminished. Consequently, despite the principle and necessity of analyzing ratios holds, further investigations are needed to select more powerful graph analysis methods.

6 Conclusion

There is still work to be done to fully characterize the method, especially concerning its power in different situations and its properties when other definition of nodes sets are used than disjoint subgraphs. However, in its current state, the method seems ready to be used in several contexts. It is easy to implement in any statistical software, for the individual ratios comparisons part, and the graph can be built with any devoted software.

Acknowledgements

We thank Bruno Blanchet for a discussion that gave the idea of exploring maximal cliques in graph. We thank one of the referees for suggestions on visualization that led to the hierarchical tree display of the results.

Funding

This work was supported in part by INSERM, AP-HP (Assistance Publique-Hôpitaux de Paris), the Bio-Psy Labex (ANR-11-IDEX-0004-02) and the Fondation FondaMental research prize (Prix Marcel Dassault 2013).

Conflict of Interest: none declared.

References

Aitchison,J. (1982) The statistical analysis of compositional data. *J. R. Stat. Soc. Series B Methodol.*, **44**, 139–177.

Aitchison,J. (1984) The statistical analysis of geochemical compositions. *Math. Geol.*, **16**, 531–564.

Aitchison,J. (2003). A concise guide to compositional data analysis. In: *2nd Compositional Data Analysis Workshop*. http://leg.ufpr.br/lib/exe/fetch.php/pessoais:abtmartins:a_concise_guide_to_compositional_data_analysis.pdf.

Bates,D. *et al.* (2015) Fitting linear mixed-effects models using lme4. *J. Stat. Softw.*, **67**, 1–48.

Brandes,U. *et al.* (2008) On modularity clustering. *IEEE Trans. Knowl. Data Eng.*, **20**, 172–188.

Bustin,S. and Nolan,T. (2017) Talking the talk, but not walking the walk: rt-qpcr as a paradigm for the lack of reproducibility in molecular research. *Eur. J. Clin. Invest.*, **47**, 756–774.

Bustin,S.A. *et al.* (2009) The miqe guidelines: minimum information for publication of quantitative real-time pcr experiments. *Clin. Chem.*, **55**, 611–622.

Carr,A.C. and Moore,S.D. (2012) Robust quantification of polymerase chain reactions using global fitting. *PLoS One*, **7**, e37640.

Tsilimigras,M.A. and Fodor,A. (2016) Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann. Epidemiol.*, **26**, 330–335.

Clauset,A. *et al.* (2004) Finding community structure in very large networks. *Phys. Rev. E*, **70**, 066111.

Costa,C. *et al.* (2013) Comprehensive molecular screening: from the rt-pcr to the rna-seq. *Trans. Lung Cancer Res.*, **2**, 87–91.

Csardi,G. and Nepusz,T. (2006) The igraph software package for complex network research. *InterJ. Complex Syst.*, **1695**, 1–9.

Filzmoser,P. *et al.* (2009) Univariate statistical analysis of environmental (compositional) data: problems and possibilities. *Sci. Total Environ.*, **407**, 6100–6108.

Friedman,J. and Alm,E.J. (2012) Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.*, **8**, e1002687–e1002681.

Geoffroy,P.A. *et al.* (2017) Lithium response in bipolar disorders and core clock genes expression. *World J. Biol. Psychiatry*, 1.

Gloor,G.B. *et al.* (2016) It's all relative: analyzing microbiome data as compositions. *Ann. Epidemiol.*, **26**, 322–329.

Li,P. *et al.* (2015) Comparing the normalization methods for the differential analysis of illumina high-throughput rna-seq data. *BMC Bioinformatics*, **16**, 347.

Ling,D. and Salvaterra,P.M. (2011) Robust rt-qpcr data normalization: validation and selection of internal reference genes during post-experimental data analysis. *PLoS One*, **6**, e17762.

Livak,K.J. and Schmittgen,T.D. (2001) Analysis of relative gene expression data using real- time quantitative pcr and the $2^{-\Delta\Delta C_T}$ method. *Methods*, **25**, 402–408.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol.*, **15**, 550.

Ruijter,M. *et al.* (2013) Evaluation of qpcr curve analysis methods for reliable biomarker discovery: bias, resolution, precision, and implications. *Methods*, **59**, 32–46.

Pawlowsky-Glahn,V. and Egozcue,J.J. (2016) Spatial analysis of compositional data: a historical review. *J. Geochem. Explor.*, **164**, 28–32.

Platts,A.E. *et al.* (2008) Real-time pcr quantification using a variable reaction efficiency model. *Anal. Biochem.*, **380**, 315–322.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radonić,A. *et al.* (2004) Guideline to reference gene selection for quantitative real-time pcr. *Biochem. Biophys. Res. Commun.*, **313**, 856–862.

Reichardt,J. and Bornholdt,S. (2006) Statistical mechanics of community detection. *Phys. Rev. E*, **74**, 016110.

Risso,D. *et al.* (2014) Normalization of rna-seq data using factor analysis of control genes or samples. *Nature Biotechnol.*, **32**, 896.

Silverman,J.D. *et al.* (2017) A phylogenetic transform enhances analysis of compositional microbiota data. *Elife*, **6**, e21887.

Spiess,A.-N. *et al.* (2015) Impact of smoothing on parameter estimation in quantitative dna amplification experiments. *Clin. Chem.*, **61**, 379.

Spiess,A.-N. *et al.* (2016) System-specific periodicity in quantitative real-time polymerase chain reaction data questions threshold-based quantitation. *Nature Sci. Rep.*, **6**, 38951.

Tellinghuisen,J. and Spiess,A.-N. (2014) Comparing real-time quantitative polymerase chain reaction analysis methods for precision, linearity, and accuracy of estimating amplification efficiency. *Anal. Biochem.*, **449**, 76–82.

Vandesompele,J. *et al.* (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.*, **3**, research0034.

Čikoš,v. and Koppel,J. (2009) Transformation of real-time pcr fluorescence data to target gene quantity. *Anal. Biochem.*, **384**, 1–10.

Xu,H. *et al.* (2015) The dilution effect and the importance of selecting the right internal control genes for rt-qpcr: a paradigmatic approach in fetal sheep. *BMC Res. Notes*, **8**, 58.

Xu,J. *et al.* (2016) The fda's experience with emerging genomics technologies-past, present, and future. *AAPS J.*, **18**, 814–818.