PLOS ONE

RESEARCH ARTICLE

Independent Component Analysis (ICA) based-clustering of temporal RNA-seq data

Moysés Nascimento[1]*, Fabyano Fonseca e Silva[2], Thelma Sáfadi[3], Ana Carolina Campana Nascimento[1], Talles Eduardo Maciel Ferreira[2], Laís Mayara Azevedo Barroso[1], Camila Ferreira Azevedo[1], Simone Eliza Faccione Guimarães[2], Nick Vergara Lopes Serão[4]

1 Department of Statistics, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil, **2** Department of Animal Science, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil, **3** Department of Exact Sciences, Federal University of Lavras, Lavras, Minas Gerais, Brazil, **4** Department of Animal Science, Iowa State University, Ames, Iowa, United States of America

* moysesnascim@gmail.com

Citation: Nascimento M, Silva FFe, Sáfadi T, Nascimento ACC, Ferreira TEM, Barroso LMA, et al. (2017) Independent Component Analysis (ICA) based-clustering of temporal RNA-seq data. PLoS ONE 12(7): e0181195. https://doi.org/10.1371/journal.pone.0181195

Editor: Quan Zou, Tianjin University, CHINA

Received: January 31, 2017

Accepted: June 27, 2017

Published: July 17, 2017

Data Availability Statement: All dataset related to simulation, real data, as well the R software codes are available from the zenodo database (DOI: 10.5281/zenodo.571134).

Funding: The authors thank CAPES, FAPEMIG and FUNARBE for the financial support. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Gene expression time series (GETS) analysis aims to characterize sets of genes according to their longitudinal patterns of expression. Due to the large number of genes evaluated in GETS analysis, an useful strategy to summarize biological functional processes and regulatory mechanisms is through clustering of genes that present similar expression pattern over time. Traditional cluster methods usually ignore the challenges in GETS, such as the lack of data normality and small number of temporal observations. Independent Component Analysis (ICA) is a statistical procedure that uses a transformation to convert raw time series data into sets of values of independent variables, which can be used for cluster analysis to identify sets of genes with similar temporal expression patterns. ICA allows clustering small series of distribution-free data while accounting for the dependence between subsequent time-points. Using temporal simulated and real (four libraries of two pig breeds at 21, 40, 70 and 90 days of gestation) RNA-seq data set we present a methodology (ICAclust) that jointly considers independent components analysis (ICA) and a hierarchical method for clustering GETS. We compare ICAclust results with those obtained for K-means clustering. ICAclust presented, on average, an absolute gain of 5.15% over the best K-means scenario. Considering the worst scenario for K-means, the gain was of 84.85%, when compared with the best ICAclust result. For the real data set, genes were grouped into six distinct clusters with 89, 51, 153, 67, 40, and 58 genes each, respectively. In general, it can be observed that the 6 clusters presented very distinct expression patterns. Overall, the proposed two-step clustering method (ICAclust) performed well compared to K-means, a traditional method used for cluster analysis of temporal gene expression data. In ICAclust, genes with similar expression pattern over time were clustered together.

Introduction

Gene expression time series (GETS) analysis aims to characterize sets of genes according to their longitudinal patterns of expression, improving the understanding of the biological processes and regulatory mechanisms of genes that share similar expression profiles over time [1]. Specifically, in GETS studies, given the large number of genes evaluated, such as those using RNA-seq data, summarization of expression profiles into a small number of clusters that include genes with similar expression over time is a typical and useful strategy to deal with the high dimensionality of GETS data sets.

In general, the methods used for gene clustering can be split into two groups. One composed by traditional methods, such as hierarchical clustering [2] and k-means [3] methodologies, which consider observations at each time as independent variables for the clustering process. K-means optimizes the variance of the clusters, whereas hierarchical methods minimize the radius of the clusters. In general, k-means outperforms hierarchical clustering, since is likely to be a poor choice for further computational analysis of the resulting clusters. [4, 5]. Although these methods are of easy application and interpretation, they have as disadvantage the fact that the temporal dependence between time-points is not taken into account in the clustering process. In the other group we have the so called model-based cluster methods [6, 7], which require normality of the data, and cluster membership is decided based on maximizing the likelihood of data points given the cluster models [1]. However, RNA-seq data, which has discrete distribution (counts of reads), is not suitable to be used in these methods that assume normality of the data [8]. In general, GETS analysis present small number of the temporal expression measures.

The application of model-based methodologies in RNA-seq data (discrete variable) presents some challenges, such as the lack of normality (assumed in several models) and the small number of temporal observations (small series), which leads to poor estimation of the effects used in the clustering process. In this context, a methodology that can be used to cluster small series of distribution-free data while accounting for the dependence between subsequent time-points should be used for temporal analysis RNA-seq data. The Independent Component Analysis (ICA) [9] is a statistical procedure that uses a transformation to convert raw time series data into sets of values of independent variables, which can be used for cluster analysis to identify sets of genes with similar temporal expression patterns. ICA is a powerful methodology, especially when traditional methods, such as principal component and factor analyses, are ineffective, since these can still find intrinsic factors that support the observational data [9].

In summary, in this paper we propose a methodology named ICAclust that jointly considers ICA and a hierarchical method for clustering temporal RNA-Seq data. The proposed methodology was applied to GETS using temporal simulated and real RNA-seq data.

Material and methods

Independent Component Analysis

Independent Component Analysis (ICA) uses the existence of independent factors (latent variables) in multivariate data and decomposes an input data set into statistically independent components [9].

Assume $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m)^T$ as the random vector. ICA approach assumes that $\mathbf{Y}$ can be modelled as linear combination of n independent components $\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n)^T$, with some matrix of unknown coefficients $\mathbf{A} = [a_{ij}]$, named mixing matrix, $\mathbf{Y}_{m\times N} = \mathbf{A}_{m\times n} \cdot \mathbf{S}_{n\times N}$.

Considering the observed data set, y_{ij} corresponds to the mean expression value (considering multiple replicates) at time j for the i^{th} series (gene). Therefore, each serie y_i is

decomposed into a linear combination given by $y_i = a_{i1}s_1 + a_{i2}s_2 + \cdots + a_{ik}s_n$, for every i = 1, 2, . . ., m, so that each series is represented by the coefficients of each independent component of the mixture.

ICA has been used for dimension reduction [10]. This possibility is specially interesting for situations approaching high dimensional problems. Aiming reduction of dimensionality, a number $k \leq n$ of independent components (IC) can be selected by using principal component analysis (PCA) as pre-processing for ICA, so that, $\mathbf{Y_{m\times N}} \propto \mathbf{A_{m\times k}} \cdot \mathbf{S_{k\times N}}$, where **A** can be approximated by the product **KR**, where **K** is an orthogonalization matrix and **R** the matrix that maximizes the statistical independence of the columns of the matrix **S**. However, because of the low number of temporal observations in GETS analysis, we should use k = m.

To verify the significance of independence hypothesis between the independent components the non-parametric Hoeffding test [11] was performed. Hoeffding test computes D statistics, which represents the distance between F(x,y) and G(x) H(y), where F(x,y) is the joint cumulative distribution function (CDF) of X and Y, and G and H are marginal CDFs.

Two-step algorithm (ICAclust) for clustering genes with similar gene expression patterns

Gene clustering was performed using a two-step approach, called ICAclust, in which ICA is initially applied to convert raw time series data, $\mathbf{Y}_{m\times N} = (y_{ij})$ into statistically independent components. Thus, the new data set, composed by elements of matrix of independent component **S**, can be used as input variables in hierarchical cluster analysis using Ward's method [12], with the number of cluster being defined by Mojena's criterion [13]. In the Ward's method, the goal at each stage of clustering is minimize the increment of the within-group error sum of squares by combining two individuals. Considering two groups, A and B, the increment is defined by $I_{AB} = \frac{n_A n_B}{n_A + n_B}(\bar{y}_A - \bar{y}_B)^T(\bar{y}_A - \bar{y}_B)$, where n_A represents the number of individuals of A, n_A represents the number of individuals of B, $\bar{y}_A$ and $\bar{y}_B$ are vectors giving the means of the variables of groups A and B, respectively. Mojena's criterion, suggests that one should select the number of groups corresponding to the first stage in the dendrogram satisfying the condition: $\alpha_{j+1} > \bar{\alpha} + cS_\alpha$, where $\alpha_0, \alpha_1, \ldots, \alpha_{n-1}$ are the fusion levels corresponding to stages with n, n-1, . . ., 1 clusters. The terms $\bar{\alpha}$ and S_α are the mean and standard error of α's, respectively; and c is a constant equal to 3.50 [13]. Fig 1 shows a scheme of the proposed method ICAclust. The resulting clusters from this analysis contain genes with similar expression patterns over time.

Real data

GETS analyses were performed with four libraries of two pig breeds (Piau and commercial breed) at 21, 40, 70 and 90 days of gestation. Animals were raised at the Pig Breeding Farm from Federal University of Viçosa, Brazil. Pregnant gilts were euthanized at each day of gestation following the procedures described at [14]. For every breed, three sows were used for each time point and embryos/fetuses were collected (four library per breed). *Longissimus dorsi* muscle samples were collected from the embryos/fetuses, except for those at 21 days post gestation, in which the whole embryo was used. The collected material was placed in tubes with RNAlater solution (Ambion, Carlsbad, CA, USA) and stored at 4˚C overnight and at -80˚C prior to RNA isolation. The procedures for obtaining the embryos and fetuses were approved by the Ethics Committee for Animal Use at UFV (protocol no. CEUA-UFV 85/2013), in accordance with current Brazilian federal legislation.

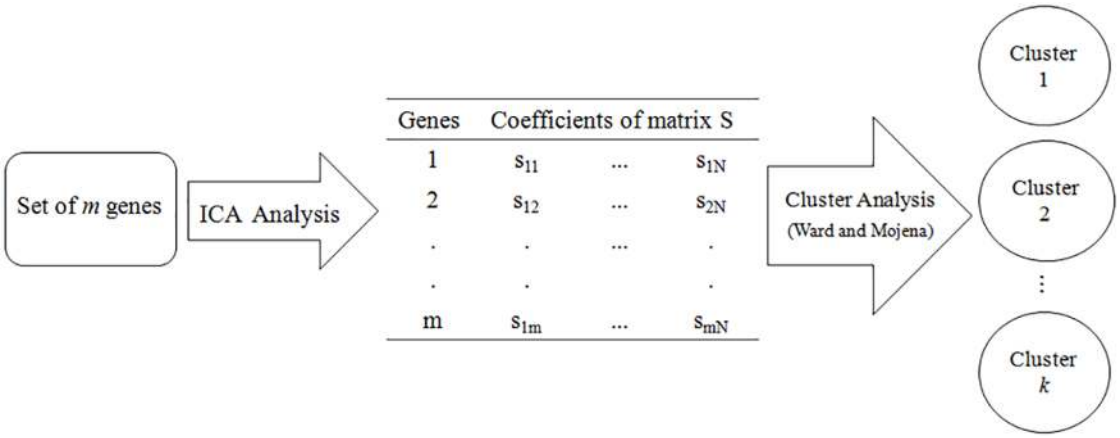

Fig 1. Flowchart summarizing the ICAclust approach.

https://doi.org/10.1371/journal.pone.0181195.g001

Total RNA was isolated with RNeasy Mini Kit (Qiagen, Valencia, CA, USA). The total concentration of RNA was estimated in a spectrophotometer NanoVue TMPlus (GE Healthcare, Freiburg, Germany) and quality checked at the Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA). rRNA were depleted using RiboMinus Eukaryote kit (Invitrogen, Carlsbad, CA). Then, RNA was fragmented by enzyme RNAse III, followed by purification and cDNA synthesis. Resulting samples were used for whole transcriptome library preparation for sequencing in SOLiD™ v.4 platform (Life Technologies Corporation, Carlsbad, CA, USA). RNA sequencing and all RNA processing procedures were performed using protocols and kits (SOLiD™ Total RNA-Seq) as recommended by Applied Biosystems. RNA sequencing was performed at the Research Center René Rachou (BH/MG), Minas Gerais, Brazil.

The data were visualized with fastQC and treated with Prinseq-Lite (v. 0.20.4; [15]. Reads were mapped by Bowtie software using *Sus scrofa* build 10.2 (Sscrofa10.2) as reference. After that, transcripts that had at least ten mapped reads across all the libraries were selected for subsequent analyses. On average, 36.75 million reads were obtained per library.

The proposed method ICAclust was applied to 458 genes that presented differential expression between breeds (FDR < 0.05) by empirical Bayesian approach based on posterior probabilities using the R package baySeq [16].

Application to simulated data

We evaluated the proposed clustering method performance using simulated datasets, each one with 458 genes divided into 6 clusters. RNA-seq quantification is based on read counts (discrete variable), and thus, count distributions such as Poisson or negative binomial are the usual choices to account for the biological phenomenon under study [17]. Additionally, since our problem approaches temporally dependent measures, a multivariate count data distribution seems appropriate for this situation. Therefore, gene expression levels were generated over 4 time points using a multivariate Poisson model with a heterogeneous first-order autoregressive covariance structure. The gene expression time series for each gene in each cluster was sampled from $Y_{ik} \sim P(\boldsymbol{\lambda}_k, \Sigma_k)$, where Y_{ik} is the time series (a vector with dimension 1 x 4) of gene i (i = 1,2,. . .,g_i) in cluster k (k = 1,2,. . .,6), $\boldsymbol{\lambda_k} = [\lambda_{1k} \ldots \lambda_{4k}]^T$ is the rate of occurrence vector, which were sampled from $\lambda_{tk} \sim N(\lambda_t, \sigma^2_{\lambda t})$ and the correlation

matrix Σ_k is given by:

$$\sum = \begin{bmatrix} \sigma_{t1} & \phi_k & \phi_k^2 & \phi_k^3 \\ & \sigma_{t2} & \phi_k & \phi_k^2 \\ & & \sigma_{t31} & \phi_k \\ & & & \sigma_{t14} \end{bmatrix},$$

where ϕ_{1k} is the autoregressive parameter and σ_{tk}^2 is the variance in each time.

The number of genes (458) and longitudinal points (i.e. 4 time points) were chosen based of the real dataset presented in the previous section. The number of clusters (i.e. 6), and number of genes in each cluster, as well as the values of $\boldsymbol{\lambda}_k$, and σ_{tk} were determined according to the results using the real data and will be presented later. For ϕ_{1k}, values used in the simulation were obtained by averaging the estimates obtained from each resulting cluster k, i.e., $\phi_{1k} = \sum_{i=1}^{I_k} \widehat{\phi}_{1k}/I_k$, where $\widehat{\phi}_{1k}$ is the estimate of ϕ_{1k} for each gene belonging to cluster k, and I_k is the number of genes in cluster k.

In order to compare the clustering method (ICAclust) presented here with a traditional clustering method, the simulated datasets were also analyzed using the, k-means algorithm [18]. The choice of k-means is due to its general use in temporal gene expression clustering [7, 8, 19] and for its overall better performance over hierarchical clustering [4, 5].

The comparison between ICA clustering methodology and k-means was evaluated by mean correct classification rate (CCR), which was computed as the ratio between the numbers of genes clustered into the true cluster (from simulation) and the total number of genes over 10 replicated datasets. It is important to emphasize that, unlikely the proposed method, the number of clusters need to be defined prior to analysis in k-means. Since, in general, the number of clusters is unknown, we simulated the data using different number of clusters (k = 2, 3, 4, 5, 6 and 7). For ICAclust, we used different values of the Mojena's constant (c = 2.25, 2.50, 2.75, 3.00, 3.25, 3.50 and 3.75) to evaluate their effect on the clustering process. These values were chosen aiming to expand the optimal clustering range (2.77–3.50) suggested by [13], and thus, being more conservative in our comparisons.

Computational features

The simulation process was carried out with the function *gen.PoisBinOrd* of the PoisBinOrd R package (http://R-cran.org). The proposed method, denoted by ICAclust, was implemented in R, through the combination of fastICA [20], hclust R functions and the Mojena's criterion. The R scripts for implementation of the proposed clustering method, and the real and simulated data sets are freely accessible at https://zenodo.org/record/571134#.WQsuLca1vIU.

Results

Real data

The six scatterplots used to visualize relationships between the four independent components, D statistics of Hoeffding and their associate p-values are shown in Fig 2. As expected, the data showed a random pattern indicating absence of any association. Overall, p-values were greater than 0.01.

Using the four independents components as variables, genes were grouped into six distinct clusters (clusters A to F) with 89, 51, 153, 67, 40, and 58 genes each, respectively. In general, it can be observed that the 6 clusters presented very distinct expression patterns (Fig 3). Fold-change is presented as the $\log_2$ of the ration between the reads in the Piau and Commercial

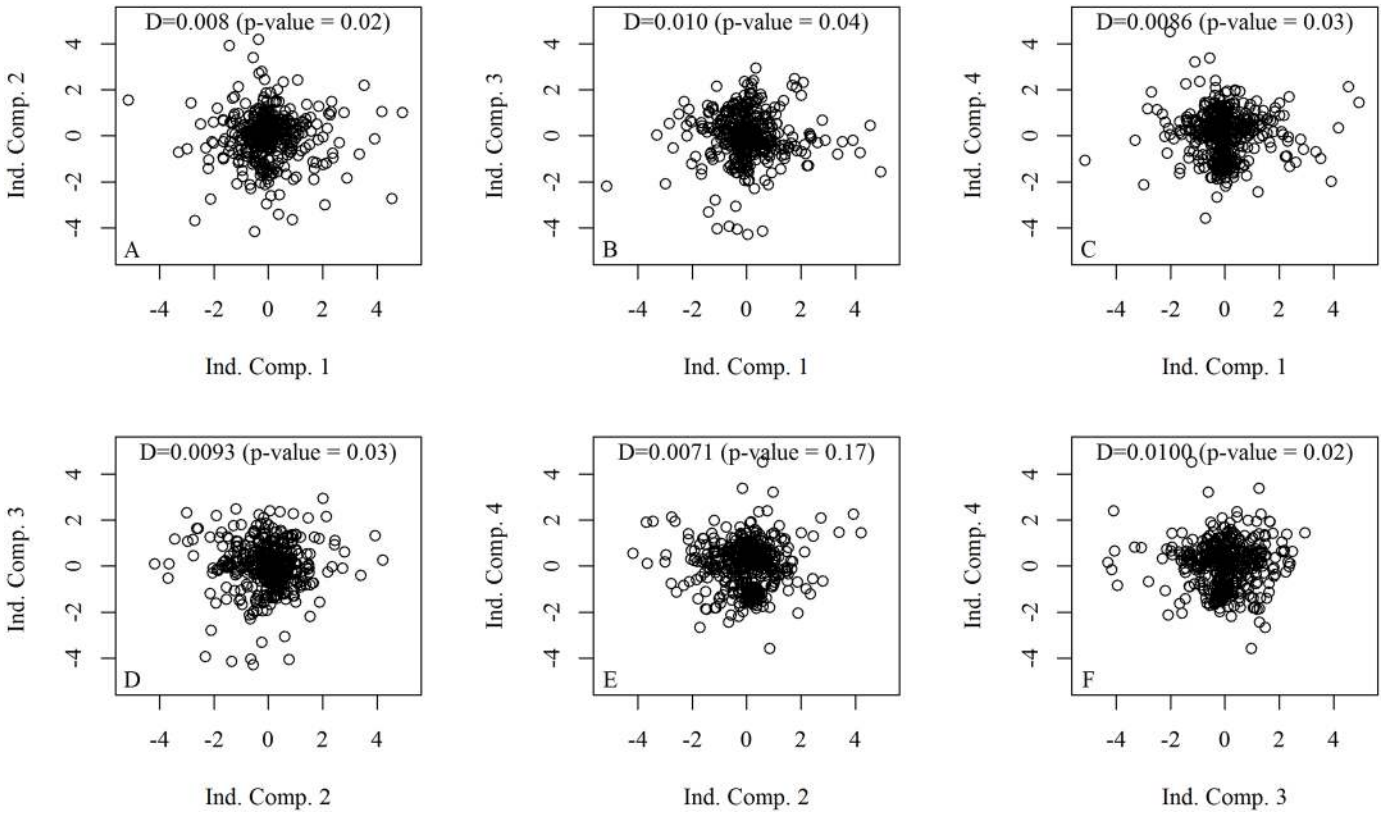

Fig 2. Scatterplots based on independent components, D statistics of Hoeffding and their associated p-values. The six scatterplots used to visualize relationships between the four independent components are represented in figures A to F.

https://doi.org/10.1371/journal.pone.0181195.g002

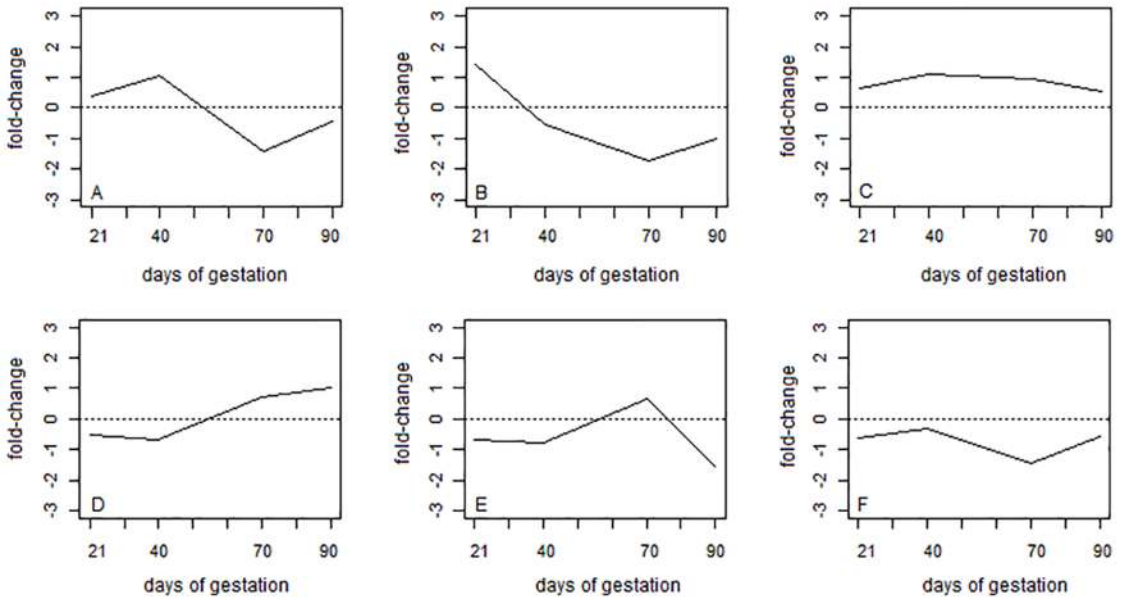

Fig 3. Time series average expression of six gene clusters found by the ICAclust. Fold-change = $\log_2$(Piau/Commercial). The six clusters are represented in figures A to F.

https://doi.org/10.1371/journal.pone.0181195.g003

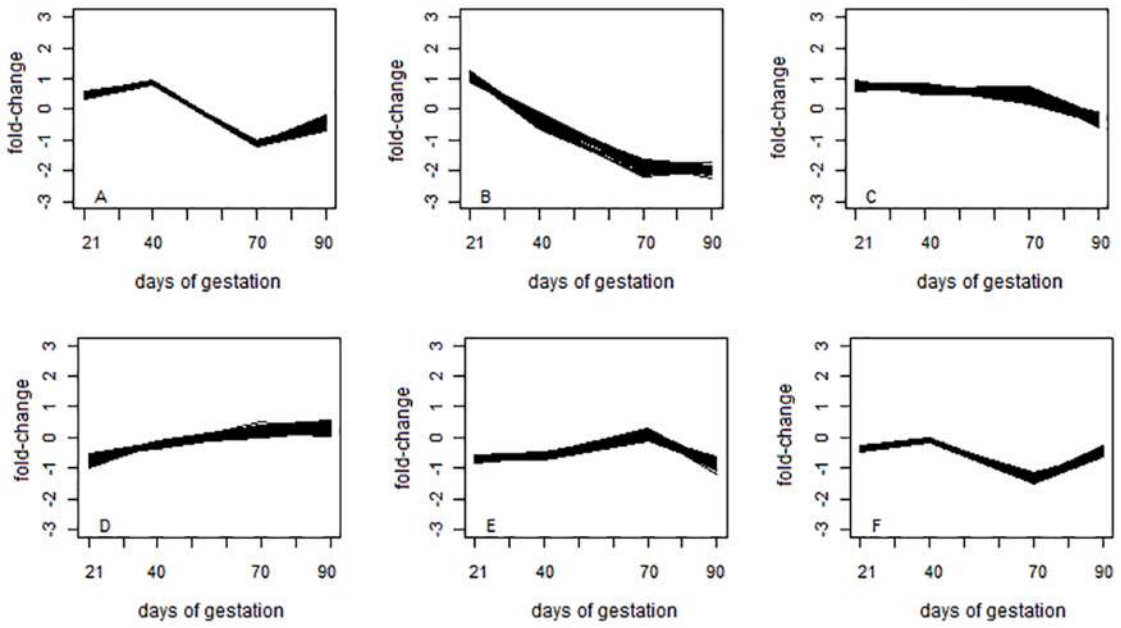

Fig 4. Average expression profile of over ten simulated data set considering the number of clusters determined according to the results using the real data. Fold-change = $\log_2$(Piau/Commercial). The six clusters are represented in figures A to F.

https://doi.org/10.1371/journal.pone.0181195.g004

breeds. Among the various differences, it can be observed that the genes that make up groups A and D had opposite average expression pattern across time. While Piau had greater overall expression than Commercial animals at 21 and 40 days of gestation in cluster A (Fig 3A), the opposite trend was found at days 70 and 90 of gestation in cluster D, where Commercial animals had greater expression than Piau. Genes belonging to the second cluster (Fig 3B) presented greater expression values in the Piau breed only at the beginning of gestation (i.e. 21 days). However, genes belonging to cluster C (Fig 3C) had greater expression in the Piau breed at all time points. In contrast, genes in cluster F had greater expression in the Commercial breed throughout the whole gestation period (Fig 3F). Furthermore, genes belonging to cluster E presented higher expression in Piau compared to Commercial only at 70 days of gestation (Fig 3E).

Simulated data

Ten replicates of gene expression profiles were simulated to compare the ICAclust and k-means methodologies. The average of expression profiles across all replicates is presented in Fig 4. The simulated data set presented similar temporal expression pattern to those clusters obtained in the real data analysis, showing that the simulation process was able to capture the same temporal relationship presented in the real data.

Performance of the clustering methods based on the simulated data is presented in Fig 5. Correct classification rate (CCR), which considers the ratio between the numbers of genes clustered into the true cluster (from simulation) and the total number of genes, was used for evaluation over the 10 replicated datasets. As depicted in Fig 5A, k-means clustering had a lower CCR compared to ICAclust in all cases. The k-means method presented the highest average CCR (86%) when the number of clusters specified for analysis was the same as the number of simulated clusters (i.e. 6), thus, representing the best case scenario for this method.

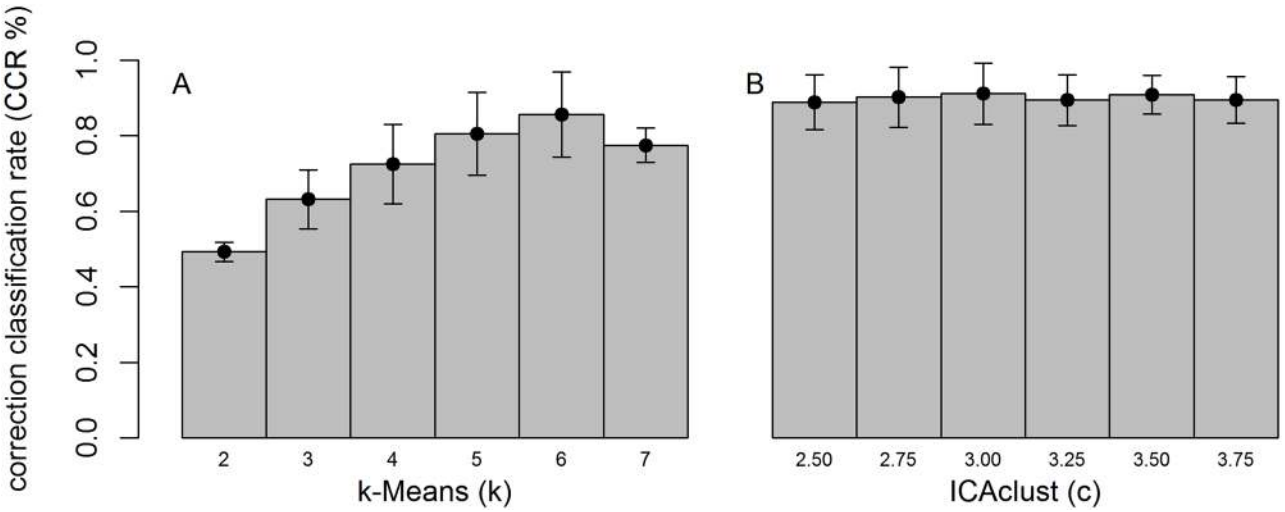

Fig 5. Average correct classification rate (CCR, %) for each clustering method across 10 replicates. The performance of k-means and ICAclust clustering methods are represented, respectively in figures A and B. Error bars represent the CCR standard deviation of 10 replicates.

https://doi.org/10.1371/journal.pone.0181195.g005

Furthermore, CCR values decreased as the number of clusters specified for the analysis moved away from 6.

On average, all ICAclust results had great CCR than k-means, and ranged from 89% to 92%, for c between 2.50 and 3.75, respectively (Fig 4B). Moreover, ICAclust presented, on average, an absolute gain of 5.15% over the best k-means scenario (k = 6). Considering the worst scenario for k -mean (k = 2), the gain was of 84.85%, when compared with the best ICA-clust result (c = 3.00). Differently than for k-means, which requires to have the number of clusters (k) defined to perform cluster analysis, ICAclust uses Mojena's criterion to determine the number of clusters automatically at end of the clustering process, and thus, increases the CCR.

The modal number of clusters identified by ICAclust was 6 over all replicates and Mojena's constant (c) values. Time series average expression of clusters, considering the modal number of clusters (k = 6), found by ICAclust is presented in Fig 6. The patterns presented in this figure are similar to those obtained from the simulated data (Fig 6), indicating that this methodology was able to create the same results as those using the real data (Fig 4).

Time series average expression of clusters found by k-means considering k = 2, 3, 4, 5, 6 and 7 are presented in Fig 7.

The correct simulated pattern was only observed when the correct number of groups (k = 6) was informed for k-means method. However, in real life, the correct number of clusters is unknown. When a lower number of clusters is considered ($k < 6$) in k-means, unique gene expression patterns can hidden. On the other hand, when the number of clusters used for analysis is higher than true value, some gene expression patterns can be split into two new groups.

Although the simulated data set was generated according to the results from the real data analysis, the comparison between the proposed ICAclust and the traditional k-means methods is reasonable. While ICAclust performed well using a traditional hierarchical method without losing information about the relationship between observations, the traditional k-means method did not account for this dependence, leading to worse results compared to those obtained by ICAclust.

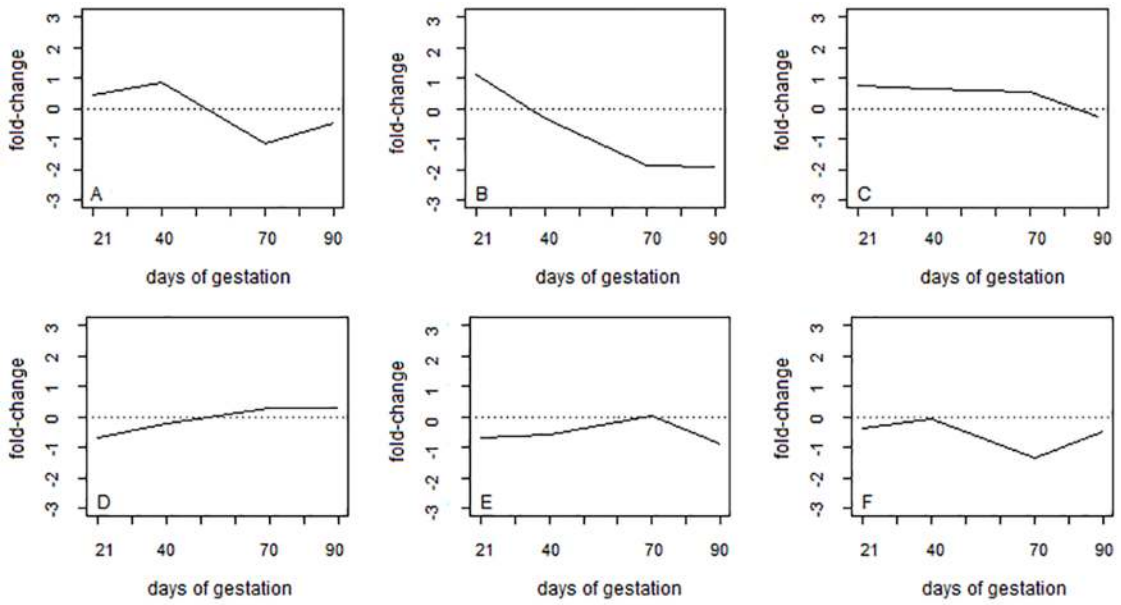

Fig 6. Time series average expression of six gene clusters found by the ICAclust considering values for c between 2.50 and 3.75. Fold-change = $\log_2$(Piau/Commercial). The six clusters are represented in figures A to F.

https://doi.org/10.1371/journal.pone.0181195.g006

Discussion

In this paper we have presented the ICAclust methodology, which can be used to decompose RNA-seq data into statistically independent components and to group genes into mutually exclusives clusters.

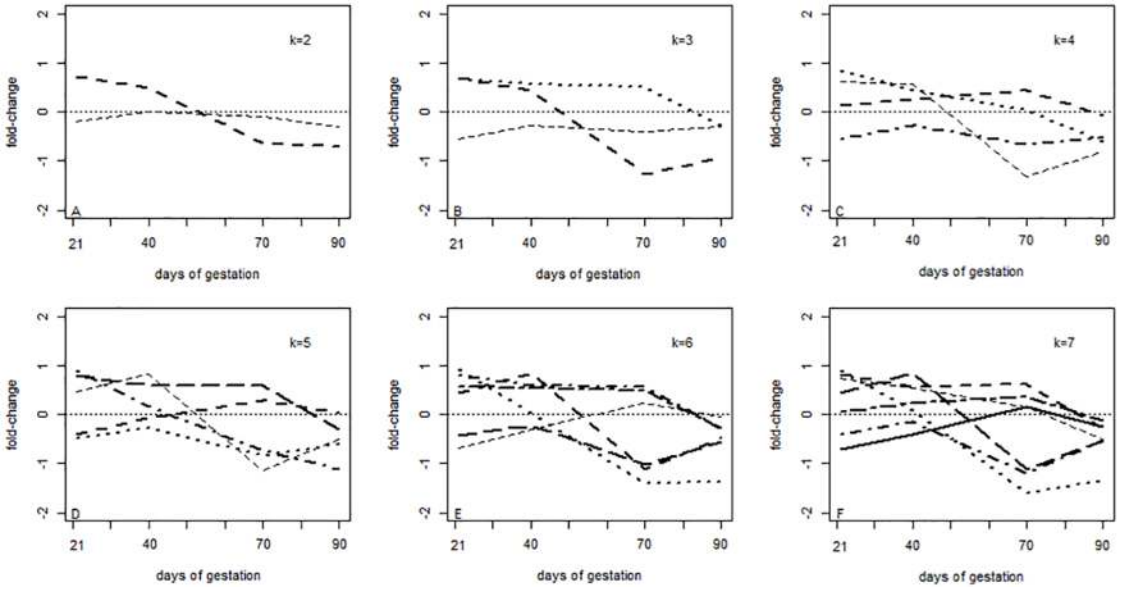

Fig 7. K-means clusters. Fold-change = $\log_2$(Piau/Commercial). Time series average expression of clusters found by k-means considering k = 2, 3, 4, 5, 6 and 7 are represented, respectively, in figures A to F.

https://doi.org/10.1371/journal.pone.0181195.g007

Independent Component Analysis (ICA) decomposes an input data set into statistically independent components. In the last decade, ICA approach was proposed as PCA, to reduce the dimensionality of the data [10]. However, differently than for PCA, where components are independent only in the presence of multivariate normality of the input variables, in ICA we can obtain statistically independent components even in the absence of multivariate normality. Therefore, ICA seems ideal for data with mixed distributions, such as those found in high-dimensional and non-normal RNA-seq data sets. Since ICA naturally takes the temporal dependency into account through its underlying model when decomposing variables [21], ICA gives the opportunity to quickly generate independent components and then group them based (e.g. genes) based on the temporal dependence among them. In our methodology, the number of latent variables (i.e. independent components) is equal to the number of samples in gene expression data, i.e., 100% of original information is used in the clustering process. The small number of RNA-seq samples does not present any problems, as ICA has been successfully used in many studies with microarray data and cluster analyses [22]. In addition, differently than the most clustering methodologies, our method outputs the number of the clusters automatically at the end of clustering process. ICAclust methodology is simpler and quicker than based-model methods [6, 7], specifically those using a Bayesian approach, since these require evaluating convergence of the chains.

Although several advantages of ICAclust have been reported here, one possible disadvantage is that this method is not model-based. Some model-based approaches have been specially indicated for time course RNA-seq studies [23], such as the autoregressive time-lagged regression and hidden Markov models. However, when working with short gene expression time series, as in the present study (only four temporal measures), the model-based methods can lead to poor clustering performance [24].

One interesting point to be exploited under a gene clustering approach is to examine how the genes are assigned to clusters as the number of clusters increases. Schonlau [25] proposed a relevant method denominated "clustergram", which enables to visualize the clustering formation and give insight on the optimal number of clusters. Since we used the Mojena criterion to identify this number, one future implication might be to update the ICAclust to provide information requested for "clustergram" implementation. Finally, with the rapid increase in the size of high-throughput genomic data, other efficient algorithms, such as MapReduce [26] and trie trees [27], could be considered in the future to improve the computational performance for read alignment.

Conclusions

The proposed two-step clustering method (ICAclust) performed well compared to k-means, a traditional method used for cluster analysis of temporal gene expression data. In ICAclust, genes with similar expression pattern over time were clustered together. Compared to k-means method, ICAclust methodology present some advantages: (i) the dependence between observations are take account in the clustering process through of independent components that are linear combinations of original variables; (ii) it is not necessary to define the number of the clusters prior to analysis, as these are obtained automatically using Mojena's criterion; (iii) ICAclust does not make any assumptions about the data distribution, i.e., it can be used for discrete data such as RNA-seq data; and (iv) it performed well to small number of temporal observations. However, more studies using different RNA-seq data sets are needed to further validate results found in this study.

Author Contributions

Conceptualization: Moysés Nascimento, Fabyano Fonseca e Silva, Thelma Sáfadi, Nick Vergara Lopes Serão.

Formal analysis: Moysés Nascimento, Ana Carolina Campana Nascimento, Talles Eduardo Maciel Ferreira, Laís Mayara Azevedo Barroso, Camila Ferreira Azevedo.

Investigation: Simone Eliza Faccione Guimarães.

Methodology: Moysés Nascimento, Fabyano Fonseca e Silva, Thelma Sáfadi, Ana Carolina Campana Nascimento, Nick Vergara Lopes Serão.

Software: Moysés Nascimento, Ana Carolina Campana Nascimento.

Writing – original draft: Moysés Nascimento, Fabyano Fonseca e Silva, Ana Carolina Campana Nascimento, Nick Vergara Lopes Serão.

Writing – review & editing: Moysés Nascimento, Nick Vergara Lopes Serão.

References

1. Schiliep A, Schonhuth A, Steinhoff C. Using hidden Markov models to analyze gene expression time course data. Bioinformatics. 2003; 19: 264–272.
2. Reeb PD, Bramardi SJ, Steibel JP. Assessing Dissimilarity Measures for Sample Based Hierarchical Clustering of RNA Sequencing Data Using Plasmode Datasets. PLoS ONE. 2015; 7: e0132310.
3. Li P, Ponnala L, Gandotra N, Wang L, Si Y, Tausta SL, et al. The developmental dynamics of the maize leaf transcriptome. Nat. Genet. 2010; 42: 1060–1067. https://doi.org/10.1038/ng.703 PMID: 21037569
4. D'haeseleer P. How does gene expression clustering work? Nature Biotechnology. 2005; 23: 1499–1501. https://doi.org/10.1038/nbt1205-1499 PMID: 16333293
5. Datta S, Datta S. Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics. 2003; 19: 459–466. PMID: 12611800
6. Ramoni MF, Sabastiani P, Kohane IS. Cluster analysis of gene expression dynamics. Proceedings of the National Academy of Sciences of America. 2002; 99: 9121–9126.
7. Nascimento M, Safadi S, Silva FF, Nascimento ACC. Bayesian model-based clustering of temporal gene expression using autoregressive panel data approach. Bioinformatics. 2012; 4: 1–5.
8. Oh S, Song S, Dasgupta N, Grabowski GA. The analytical landscape of static and temporal dynamics in transcriptome data. Frontiers in Genetics. 2014; 5: 35. https://doi.org/10.3389/fgene.2014.00035 PMID: 24600473
9. Hyvärinen A, Karhunen J, Oja Erkki. Independent Component Analysis. J. Wiley, New York; 2001.
10. Wang J, Chang C. Independent Component Analysis-Based Dimensionality Reduction with Applications in Hyperspectral Image Analysis. IEEE Transactions on Geoscience and Remote Sensing. 2006; 44: 1586–1600.
11. Hoeffding W. A non-parametric test of independence. Annals of Mathematical Statistics. 1948; 19: 293–325.
12. Ward JH. Hierarquical grouping to optimize an objective function. Journal of the American Statistical Association. 1963; 58: 236–244.
13. Mojena R. Hierarchical grouping method and stopping rules: an evaluation. Computer Journal. 1977; 20: 359–363.
14. Sollero BP, Guimarães SE, Rilington VD, Tempelman RJ, Raney NE, Steibel JP, et al. Transcriptional profiling during foetal skeletal muscle development of Piau and Yorkshire-Landrace cross-bred pigs. Animal Genetics. 2011; 42: 600–12. https://doi.org/10.1111/j.1365-2052.2011.02186.x PMID: 22035001
15. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011; 27: 863–864. https://doi.org/10.1093/bioinformatics/btr026 PMID: 21278185
16. Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. Bioinformatics. 2010; 11: 422. https://doi.org/10.1186/1471-2105-11-422 PMID: 20698981

17. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics. 2007; 23: 2881–2887. https://doi.org/10.1093/bioinformatics/btm453 PMID: 17881408

18. McQueen JB. Some methods for classification and analysis of multivariate observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1967; 1: 281–297.

19. Buettner F, Natarajan KN, Casale PF, Proserpio V, Scialdonw A, Theis FJ, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol. 2015; 33: 155–60. https://doi.org/10.1038/nbt.3102 PMID: 25599176

20. Marchine JL et al. fastICA: FastICA Algorithms to perform ICA an Projection Pursuit. 2010; 1: 1–13. Available from: http://CRAN.R-project.org/package=fastICA, R package version.

21. Calhoun VD, Adalı T. Multi-subject Independent Component Analysis of fMRI: A Decade of Intrinsic Networks, Default Mode, and Neurodiagnostic Discovery.IEEE reviews in biomedical engineering. 2012; 5: 60–73. https://doi.org/10.1109/RBME.2012.2211076 PMID: 23231989

22. Kong W, Vanderburg CR, Gunshin H, Rogers JT, Huang X. A review of independent component analysis application to microarray gene expression data. BioTechniques. 2008; 45: 501–520. https://doi.org/10.2144/000112950 PMID: 19007336

23. Oh S, Song S, Grabowski G, Zhao H, Noonan JP. Time series expression analyses using RNA-seq: a statistical approach. Biomed Res. Int. 2013; 203681. https://doi.org/10.1155/2013/203681 PMID: 23586021

24. Bar-Joseph Z. Analyzing time series gene expression data. Bioinformatics. 2004; 20: 2493–2503. https://doi.org/10.1093/bioinformatics/bth283 PMID: 15130923

25. Schonlau M. Visualizing non-hierarchical and hierarchical cluster analyses with clustergrams. Comput. Stat. 2004; 19: 95–111.

26. Zou Q, Hu Q, Guo M, Wang G. HAlign: Fast multiple similar DNA/RNA sequence alignment based on the centre star strategy. Bioinformatics. 2015; 15:2475–81.

27. Zou Q, Li XB, Jiang WR, Lin ZY, Li GL, Chen K. Survey of MapReduce frame operation in bioinformatics. Comput. Stat. 2004; 19: 95–111. Brief Bioinform. 2014; 15:637–47.