Journal of Biotechnology 326 (2021) 1–10

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Multi-assignment clustering: Machine learning from a biological perspective

Benjamin Ulfenborg [a,*], Alexander Karlsson [b], Maria Riveiro [b,c], Christian X. Andersson [d], Peter Sartipy [a], Jane Synnergren [a]

[a] School of Bioscience, University of Skövde, Skövde, Sweden
[b] School of Informatics, University of Skövde, Skövde, Sweden
[c] Department of Computer Science and Informatics, School of Engineering, Jönköping University, Jönköping, Sweden
[d] Takara Bio Europe AB, Gothenburg, Sweden

ARTICLE INFO

Keywords:
Clustering
Multiple cluster assignment
K-means
Transcriptomics
Annotation enrichment
Pathways

ABSTRACT

A common approach for analyzing large-scale molecular data is to cluster objects sharing similar characteristics. This assumes that genes with highly similar expression profiles are likely participating in a common molecular process. Biological systems are extremely complex and challenging to understand, with proteins having multiple functions that sometimes need to be activated or expressed in a time-dependent manner. Thus, the strategies applied for clustering of these molecules into groups are of key importance for translation of data to biologically interpretable findings. Here we implemented a multi-assignment clustering (MAsC) approach that allows molecules to be assigned to multiple clusters, rather than single ones as in commonly used clustering techniques. When applied to high-throughput transcriptomics data, MAsC increased power of the downstream pathway analysis and allowed identification of pathways with high biological relevance to the experimental setting and the biological systems studied. Multi-assignment clustering also reduced noise in the clustering partition by excluding genes with a low correlation to all of the resulting clusters. Together, these findings suggest that our methodology facilitates translation of large-scale molecular data into biological knowledge. The method is made available as an R package on GitLab (https://gitlab.com/wolftower/masc).

1. Introduction

Over the last two decades, the amount of available molecular data has increased dramatically. State-of-the-art techniques such as sequencing, mass spectrometry and other high-throughput 'omics' technologies enable efficient generation of extensive experimental datasets that provide unique resources for advanced data mining. These datasets provide unprecedented insights into molecular pathways in cells and their role in various diseases (Meng et al., 2016). However, many challenges remain on how to best perform the analysis of this large-scale molecular data (Li and Chen, 2014; Sulakhe et al., 2014).

One of the most common strategies for analyzing high-dimensional data is to find clusters of objects that share similar characteristics. Many clustering algorithms are based on machine learning techniques designed to discover hidden structures in the data (Xu and Wunsch, 2010). Unlabeled data objects are grouped into clusters so that objects in the same cluster are more similar to each other than to objects assigned to other clusters. A cluster is inherently a subjective structure that does not have a precise and formal definition. In theory, data points that are in the same group should have similar properties and/or features, while data points in separate groups should have different properties and/or features (Gan et al., 2007).

Along with the large number of clustering algorithms available, a rich literature on cluster analysis has developed over the years and several reviews account for the variety of algorithms, application fields, types, weaknesses, and strengths (Aggarwal, 2014; Jain, 2010; Wenskovitch et al., 2017). Examples of clustering algorithms commonly used in biomedical research are presented in the review by Xu and Wunsch

Abbreviations: ECM, Extracellular matrix; HTA, Human Transcriptome Array; KEGG, Kyoto Encyclopedia of Genes and Genomes; MAPK, Mitogen-activated protein kinase; MAsC, Multi-assignment clustering; NAFLD, Non-alcoholic fatty liver disease; PPAR, Peroxisome proliferator-activated receptor; RMA, Robust Multichip Average; SPIA, Signaling Pathway Impact Analysis; TGF, Transforming growth factor.

* Corresponding author.
E-mail address: benjamin.ulfenborg@his.se (B. Ulfenborg).

https://doi.org/10.1016/j.jbiotec.2020.12.002
Received 17 June 2020; Accepted 3 December 2020
Available online 4 December 2020

(Xu and Wunsch, 2010). Conventional clustering algorithms have been used or adapted to gene expression data, and algorithms have also been developed specifically for gene expression data analysis (Si et al., 2014). K-means (Hartigan and Wong, 1979) is one of the most well-known clustering algorithms and frequently used in biomedical data analysis (Jain, 2010). The main advantages of K-means are that it is transparent, relatively fast, and has low memory requirement since it only computes the distances between the data point and the centroid (group center) of the clusters until convergence.

In most clustering analyses of large-scale molecular data that are performed today, the objects are restricted to be assigned to a single cluster, potentially disregarding high similarity to several other clusters. The cluster with the *highest* similarity based on a selected similarity measure is the one that the object will be assigned to. From a modeling perspective, this treats the clusters as being mutually exclusive, which is a poor description of the underlying biological context where genes or proteins may be involved in several processes and signaling pathways simultaneously (Fallah et al., 2019; Nikolaou et al., 2013). These pathways control much of the functions that take place in cells and tissues where gene transcripts, enzymes, proteins, and other molecules serve as key components (Devkota and Wuchty, 2020).

One of the basic assumptions when using clustering techniques in the analysis of molecular data is that genes with highly similar expression profiles may be involved in the same molecular processes. Thus, when interpreting clustering results, it is relevant to explore over-representation of genes in known pathways, or enrichment for specific functional annotations. However, since the majority of clustering analyses today do not consider multiple assignment of objects, the downstream analysis may become less informative, since genes of a specific pathway can be split into several clusters. Many clusters lack objects that participate in similar biological processes, as the other members in these processes may have been assigned and locked to other clusters. This results in lower statistical power in subsequent enrichment analyses and affects the correct mapping of clusters to their true biological functions.

The problem that genes need to be part of multiple clusters may be handled through different types of fuzzy clustering techniques, e.g. Ferraro & Giordani (Ferraro and Giordani, 2015), which allow objects to belong to more than one cluster with a certain degree (Bandyopadhyay et al., 2007; Fu and Medico, 2007; Wu et al., 2011; Zhang et al., 2011). There are also other approaches for imprecise clustering based on the assignment of meta clusters (Liu et al., 2015; Wu et al., 2011), and model-based approaches using mixture models (Mitchell, 1999), where one can obtain probabilistic membership information to the different clusters. These approaches provide alternative perspectives on the clustering, but do not assign objects into multiple clusters simultaneously.

In contrast to previous research and given that a substantial portion of gene expression profiling analyses are performed under the assumption of mutually exclusive assignment to clusters, our aim was to investigate the effects of multi-assignment clustering on the downstream analysis. We hypothesized that annotation enrichment analysis of clusters could be rendered more sensitive and biologically meaningful by performing it on non-mutually exclusive clustering, i.e. when genes are allowed to belong to more than one cluster. In this work, we implemented and compared the results for both single- and multi-assignment clustering using K-means with Pearson correlation distance measure on two independent datasets of different sizes. A threshold on the correlation was used to assign the genes to none or several clusters.

2. Material and methods

2.1. Description of omics datasets

In order to test the hypothesis that assignment of genes to multiple clusters supports the biological interpretation of clusters, two transcriptomics datasets from previous gene expression studies performed in our group have been re-analyzed and the efficiency of the proposed multi-assignment clustering approach evaluated. The datasets are publicly available in ArrayExpress with accession numbers E-MTAB-5219 and E-MTAB-5367. E-MTAB-5219 (referred to as the Mesoderm dataset) represents a time series transcriptomics dataset from human embryonic stem cells during differentiation towards an early cardiac phenotype. The dataset consists of eleven time points, representing cells sampled daily from day 0 to day 10 during the differentiation process, as described in detail previously (Ulfenborg et al., 2017). E-MTAB-5367 (referred to as the Hepatocyte dataset) represents time series data from six different human pluripotent stem cell lines during differentiation towards hepatocytes. This dataset has five time points with two biological replicates for each time point and cell line. The experimental setup and technical details of this dataset have been described previously (Ghosheh et al., 2017).

2.2. Data analysis

2.2.1. Preprocessing of data

Raw gene expression signals were background-corrected and normalized with the Robust Multichip Average (RMA) function in the oligo package (Carvalho and Irizarry, 2010) with version 1.48.0 (R version 3.6.0). For both datasets, gene expression was calculated by taking the mean of the biological replicates at each time point. Probes were mapped to Entrez gene IDs with NetAffx annotation from Affymetrix (ThermoFisher Scientific, https://www.thermofisher.com). For the Mesoderm dataset, the HuGene 1.0 ST V1 NA36 annotation against the hg19 reference genome was used. For the Hepatocyte dataset, the HTA 2.0 R3 NA36 annotation against the hg19 reference genome was used. When multiple probes mapped to the same Entrez gene ID, only the probe with the highest expression value was retained. The datasets were filtered to remove probes without Entrez gene IDs, probes with a $\log^2$ expression below 5 in all time points and probes with a coefficient of variation below 10 %. Following this preprocessing, the Mesoderm dataset contained 1224 genes and 11 time points, and the Hepatocyte dataset contained 3219 genes and 5 time points. Fold changes used for downstream pathway analysis were calculated by taking the gene expression of the last time point divided by the expression in the first time point.

2.2.2. Multi-assignment clustering scheme

We utilized a simple similarity-clustering schema based on thresholding. The basic idea is to let a data point x belong to more than one cluster if that point is sufficiently similar, given by the threshold δ, to several clusters. More formally, we propose a clustering schema based on the centroids $c_1, \ldots, c_K$ as the result from K-means clustering by:

$$C(x) = \{c_i : \Gamma_x(c_i) \geq \delta,\ i \in \{1, \ldots, K\}\}$$

where $\delta \in [0, 1]$ is a threshold and $\Gamma_x(c_i)$ is defined as:

$$\Gamma_x(c_i) = 0,\ \textit{if}\ \alpha(x, c_i) \leq 0$$

$$\Gamma_x(c_i) = \alpha(x, c_i),\ \textit{otherwise}$$

where $\alpha(x, c_i)$ denotes the Pearson correlation coefficient. Hence, instead of assigning each point to a single cluster, we utilize a threshold in order to determine several cluster assignments as illustrated in Fig. 1.

2.2.3. Clustering analysis

K-means clustering was carried out in R with the amap package (Caussinus et al., 2003), version 0.8.17. We chose K-means as a basis for our multi-assignment clustering technique, primarily due to its ease of interpretation and that it is one of the most commonly used clustering algorithms within bioinformatics research (Rodriguez et al., 2019; Xu and Wunsch, 2010). The number of clusters was set to maximize the intracluster correlations while maintaining low intercluster correlations

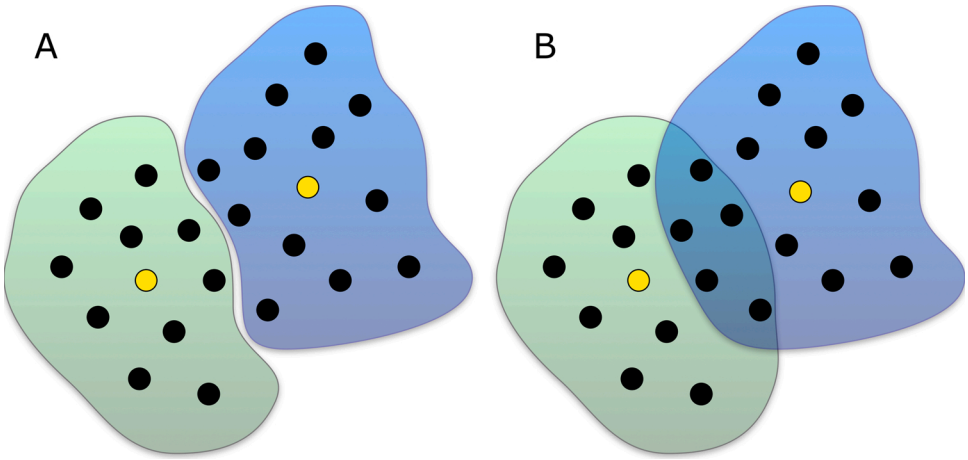

Fig. 1. Schematic illustration of the multi-assignment clustering approach.
The idea behind multi-assignment clustering is to allow each gene in the dataset to be assigned to more than one cluster simultaneously. In the standard clustering analysis (Panel A), clusters are treated as mutually exclusive and genes (black) can never appear in more than one cluster. With multi-assignment clustering (Panel B), genes that are sufficiently similar to several cluster centroids (yellow) will be assigned to each of them. This makes the clusters partially overlapping and drops the assumption of exclusive cluster assignment.

as described previously (Ulfenborg et al., 2017). This resulted in 10 cluster centroids for each dataset, here referred to as the standard clustering results. The distance measure for the clustering was set to Pearson. The centroids from the standard clustering results were used in the next step to calculate correlations $\alpha(x, c_i)$ between gene expression profiles and the clusters.

2.2.4. Defining threshold for multi-cluster assignment

To enable downstream enrichment analysis, we wanted to achieve sufficiently large clusters after applying the δ threshold. A mean cluster size of 200 genes was considered as a minimum as this size is suggested to be required for follow-up enrichment analysis (Huang et al., 2009). Higher thresholds result in fewer genes assigned to the clusters, since the gene correlation to the cluster centroid must be higher than δ for the gene to belong to that cluster. To find a suitable threshold, a step-wise parameter search with δ from 0.5 to 1.0 was performed for the datasets investigated (Fig. 2). For both datasets we found that $\delta \geq 0.9$ resulted in sufficient cluster sizes for downstream pathway enrichment analysis. Applying this threshold for correlation, the mean cluster size was 220 for the Mesoderm dataset and 446 for the Hepatocyte dataset. The approach that allows genes to be assigned to multiple clusters is referred to as Multi-Assignment Clustering (MAsC).

2.2.5. Pathway analysis

For each of the centroids extracted from the standard K-means clustering results, the genes were assigned to zero, one or multiple clusters based on the similarity to each of the ten centroids, i.e. when Pearson correlation $\geq$ δ. Genes in the resulting clusters were then assessed with respect to their enrichment of KEGG molecular pathway annotations. Significantly enriched pathways among the genes in each cluster were identified with the SPIA package in R (Tarca et al., 2009). The number of significantly enriched pathways detected in the clusters from MAsC was calculated and compared to the results from the standard K-means clustering. The pathways reported as significant were also investigated and compared between the clustering approaches.

3. Results

3.1. Descriptive comparison of cluster properties between clustering approaches

The MAsC approach generated on average larger cluster sizes compared to the standard K-means algorithm as shown in Figs. 3 and 4 for the Mesoderm and the Hepatocyte datasets, respectively. As expected, the cluster sizes increased when MAsC was used, and a larger

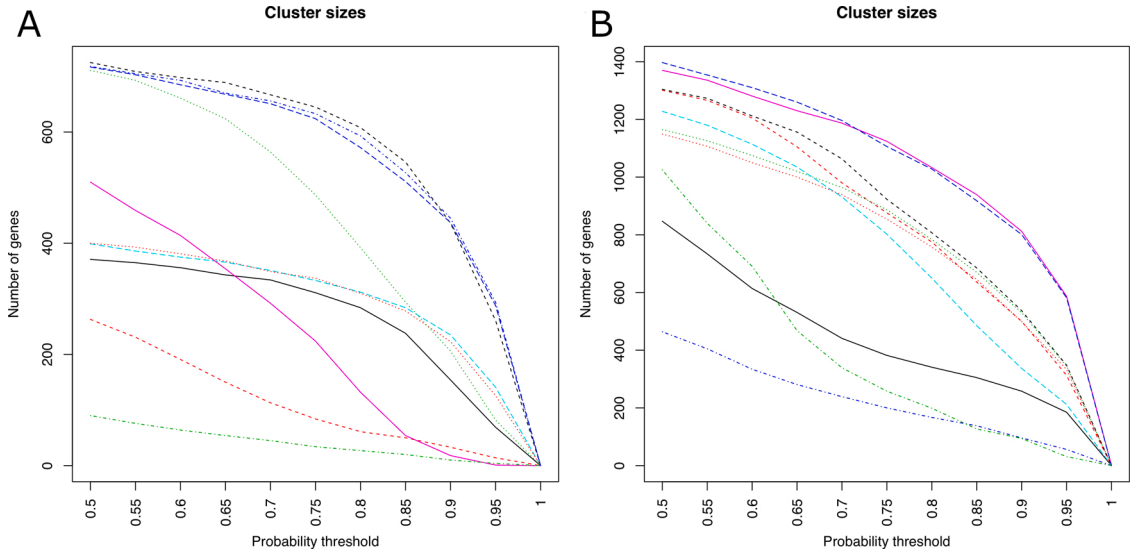

Fig. 2. Exploring different thresholds for multi-assignment clustering.
By varying the threshold δ from 0.5 to 1 different cluster sizes are obtained. Genes are assigned to all clusters where the correlation between the gene and cluster centroid is $\geq$ δ. Apart from allowing genes to be assigned to multiple clusters, this also results in removal of certain genes from all clusters, when correlation to all centroids is below δ. Panel A represents the Mesoderm dataset and panel B represents the Hepatocyte dataset.

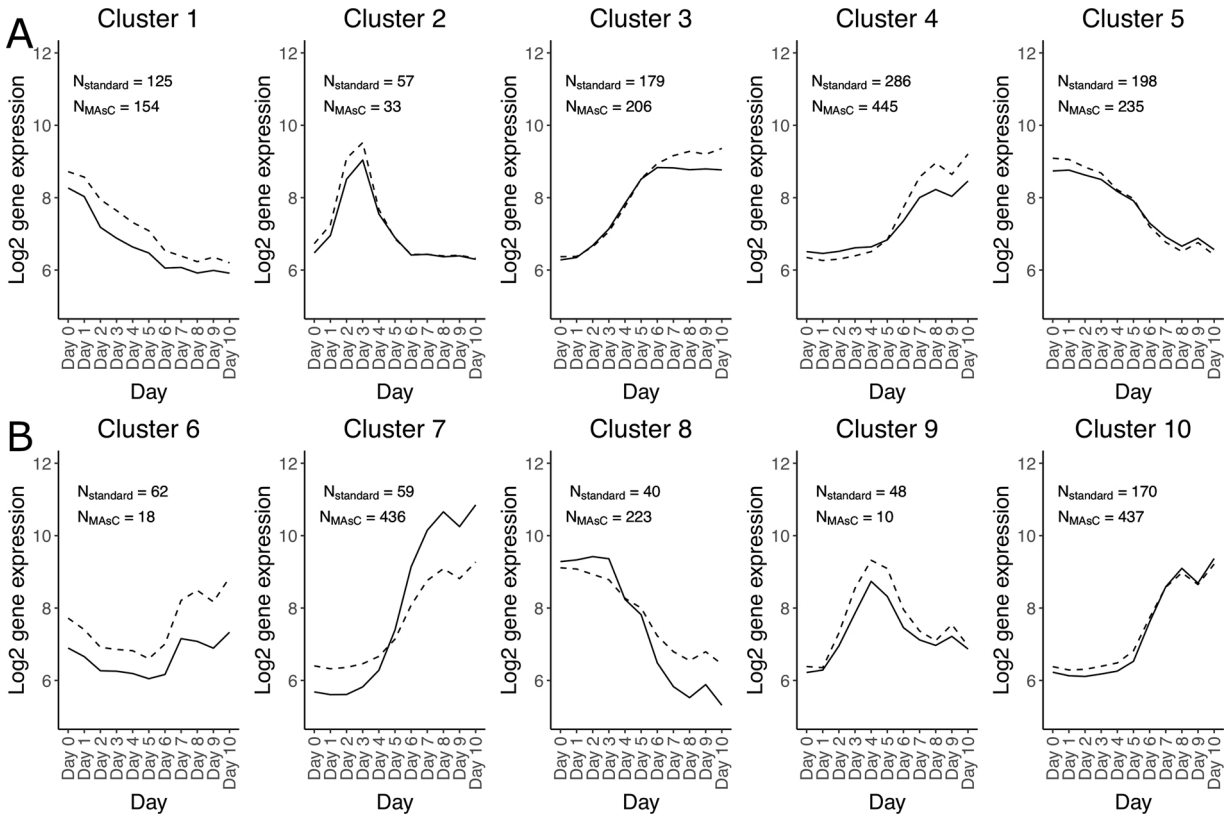

Fig. 3. Mesoderm cluster profiles.
The mean expression profile for the standard (solid line) and multi-assignment clustering ($\delta = 0.9$, dashed line) for the Mesoderm dataset. For each cluster, the number of genes assigned to that expression profile is shown above the profiles for the cluster.

number of pathways were identified as significantly enriched for both the investigated datasets (Fig. 5A and B).

3.1.1. *The mesoderm dataset*

The mean cluster size increased from 122 (standard K-means) to 220 genes (MAsC). Three clusters (2, 6 and 9) became smaller with a MAsC cluster sizes of 33, 18 and 10 genes, respectively. The remaining clusters became larger, with the largest increase observed for cluster 7 and 10, changing from 59 to 436 genes and from 170 to 437 genes, respectively. This concurrent increase is attributed to the high correlation (0.98) between the standard K-means centroids of these two clusters, resulting in a gene overlap of 85 % between these MAsC clusters. Out of the 1224 genes in the Mesoderm dataset, 287 (23 %) were not assigned to any clusters, since their correlation to all K-means centroids was below 0.9. In total, 219 (18 %), 246 (20 %), 402 (33 %) and 70 (6 %) genes were assigned to one, two, three or four clusters, respectively. No gene was assigned to more than four clusters.

3.1.2. *The hepatocyte dataset*

The mean cluster size increased from 322 to 446 genes when MAsC was applied, compared to the standard K-means clustering. Also, for this dataset, three clusters (1, 4 and 9) became smaller when using MAsC, with cluster sizes 258, 96 and 94 genes respectively. The other seven clusters increased in size and some of them considerably, e.g. cluster 6 that increased from 240 genes to 812 genes. As for the Mesoderm dataset, concurrent increase in cluster size can be attributed to the high correlation between the standard K-means centroids. For clusters 6 and 10 the correlation was 0.99, resulting in a gene overlap of 85 % between the corresponding MAsC clusters. The Hepatocyte dataset contained 3219 genes in total and of these 745 (23 %) were not assigned to any clusters, due to low correlations to all K-means centroids. As for the Mesoderm dataset, genes were assigned to between one and four clusters. Here, 758 (24 %), 1500 (47 %), 158 (5 %) and 58 (2 %) genes were assigned to one, two, three or four clusters, respectively. This shows that MAsC maintains an informative clustering partition, and avoids a situation where the majority of genes are assigned to the majority of clusters.

3.2. *Comparison of biological relevance between clustering approaches*

To evaluate whether higher biological relevance can be achieved by allowing genes to participate in multiple clusters, pathway enrichment analysis was performed on each of the ten clusters from the standard K-means clustering and MAsC, respectively, and the results compared. In this work, higher biological relevance was defined as an increase in the number of enriched pathways directly involved in the differentiation process behind the Mesoderm and Hepatocyte datasets.

3.2.1. *The mesoderm dataset*

As shown in Fig. 6 and 7, MAsC increased the number of significant pathways identified in four of the five clusters, for which enriched pathways were identified in the standard K-means results. Moreover, several pathways were also found with MAsC in two clusters (3 and 4) for which no any significant pathways were identified using the standard K-means. In total, eight pathways were identified as enriched in the K-means clusters. The corresponding number when using the MAsC was 30 (Fig. 5A). One reason for the higher sensitivity of the pathway analysis for MAsC clusters is the larger cluster sizes achieved using this approach. The elimination of genes with low correlation to all cluster centroids may also contribute, as these genes have different expression profiles from the other genes and may not be involved in the same processes. Consequently, the MAsC clusters have higher intracorrelations, and likely a larger fraction of genes involved in the same pathways.

In addition to the higher sensitivity in the pathway enrichment

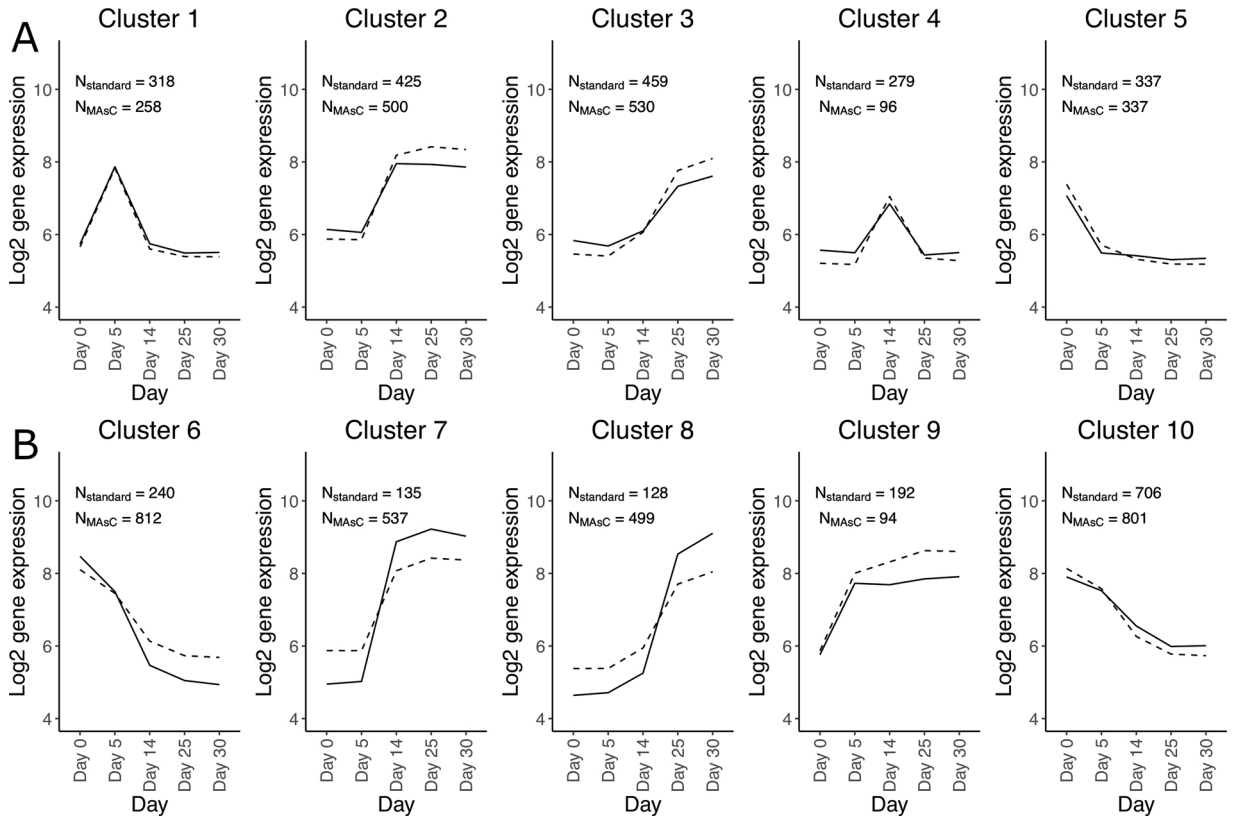

Fig. 4. Hepatocyte cluster profiles.
The mean expression profile for the standard (solid line) and multi-assignment clustering (δ = 0.9, dashed line) for the Hepatocyte dataset. For each cluster, the number of genes assigned to that expression profile is shown in the graphs for the cluster.

A

Cluster	Threshold.0.5	Threshold.0.6	Threshold.0.7	Threshold.0.8	Threshold.0.9	Original
1	3	3	3	0	0	0
2	4	5	3	1	1	1
3	12	9	11	0	3	0
4	8	11	7	7	6	0
5	5	1	4	4	3	1
6	6	6	2	0	0	0
7	8	11	9	7	6	3
8	6	1	4	3	0	0
9	2	1	1	1	5	1
10	10	9	6	8	6	2
Total	64	57	50	31	30	8

B

Cluster	Threshold.0.5	Threshold.0.6	Threshold.0.7	Threshold.0.8	Threshold.0.9	Original
1	7	4	2	2	2	2
2	16	9	6	4	4	2
3	20	20	20	22	17	6
4	1	0	0	0	0	1
5	6	6	6	3	1	0
6	6	8	8	8	8	9
7	21	10	9	4	4	3
8	20	20	20	21	15	6
9	10	6	1	0	0	11
10	6	7	8	9	9	6
Total	113	90	80	73	60	46

Fig. 5. Number of significant pathways for different thresholds. δ.
Heatmaps showing the number of significant pathways identified for different values of δ. Panel A represents the Mesoderm dataset and panel B represents the Hepatocyte dataset.

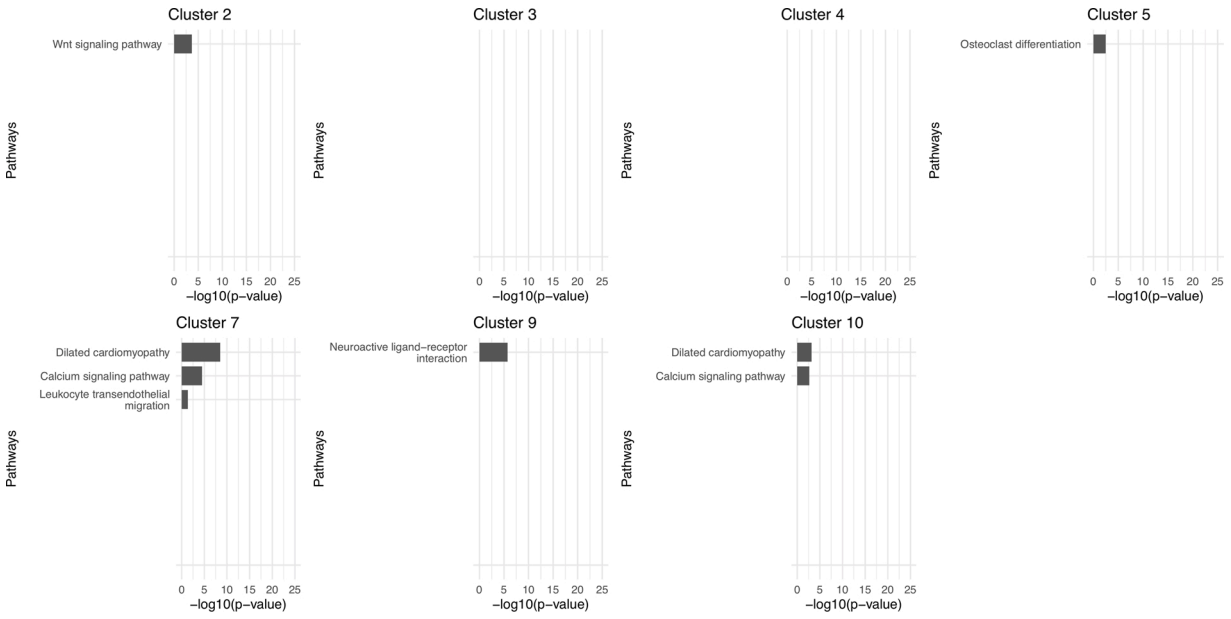

Fig. 6. Significantly enriched pathways for the Mesoderm dataset using K-means.
The top ten significant pathways for the Mesoderm dataset are shown for the standard K-means clustering. Only clusters with at least one significant pathway in one of the clustering analyses are shown.

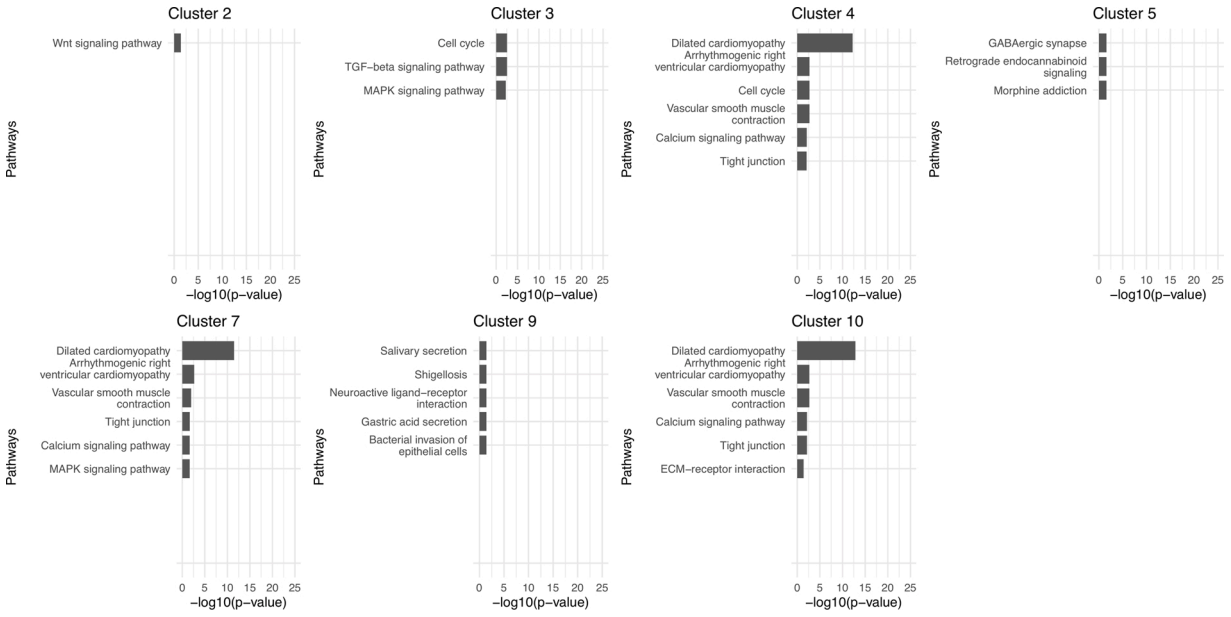

Fig. 7. Significantly enriched pathways for the Mesoderm dataset using MAsC.
The top ten significant pathways for the Mesoderm dataset are shown for the MAsC ($\delta = 0.9$). Only clusters with at least one significant pathway in one of the clustering analyses are shown.

analysis, the identified pathways also showed a higher biological relevance to the investigated differentiation process when MAsC was applied. This was particularly prominent for Mesoderm dataset cluster 3, 4, 7 and 10 (Figs. 6 and 7). Fig. 6 shows all pathways identified as significantly enriched using the standard K-means clustering and Fig. 7 shows the corresponding results based on MAsC. For cluster 3, no pathway was identified as enriched with standard K-means clustering, while with MAsC three relevant pathways were identified. Among these are 'TGF-beta signaling pathway' (Gordeeva, 2019) and the 'MAPK signaling pathway' (Zhang and Liu, 2002), which are highly involved in cell proliferation and cardiac differentiation. Similarly, for cluster 4 no significant pathways were identified using standard K-means. When MAsC was applied, five pathways were identified as significantly enriched and two of those; 'Dilated cardiomyopathy' and 'Arrythmogenic right ventricular cardiomyopathy' are directly connected to cardiac disease (Cojan-Minzat et al., 2020; Paul and Schulze-Bahr, 2020). 'Vascular smooth muscle contraction', 'Calcium signaling pathway' and 'Tight junction' are all coupled to cardiac functionality (Adesse et al., 2011; Winslow et al., 2016). Cluster 7 showed interesting results and three pathways were identified using the standard K-means clustering, two of which are tightly coupled to cardiac functionality, disease development and progression. Interestingly, when MAsC was applied, the number of enriched pathways increased from three to six, and the 'Leukocyte transendothelial migration' pathway (found with standard

K-means) was no longer significant. This pathway has no obvious association to mesoderm or cardiac tissue. Furthermore, all of the other six pathways identified using MAsC had high cardiac relevance. Finally, for cluster 10 there were two significantly enriched and highly relevant pathways identified using the standard K-means. With MAsC, four additional biologically relevant pathways for mesoderm and cardiac development were detected. Among these is the 'ECM-receptor interaction', which was not identified as enriched in any of the other clusters. It has a direct or indirect control of cellular activities such as adhesion, migration, differentiation, proliferation, and apoptosis, which are important during mesoderm and cardiac differentiation (Bosman and Stamenkovic, 2003).

3.2.2. The hepatocyte dataset

This dataset represents five different time points during differentiation of human pluripotent stem cells towards the hepatic lineage. Clustering results showed higher biological relevance when MAsC was applied compared to the standard K-means clustering (see Fig. 8 and 9). The Hepatocyte dataset is more than twice as large as the Mesoderm dataset and produces larger clusters; therefore, the number of enriched pathways is higher. In total, 46 pathways were identified as significant with the standard K-means, compared to 60 with MAsC (Fig. 5B). The results revealed that MAsC in general contributed to identify a larger number of pathways relevant to hepatocyte development than the standard K-means. With MAsC, three pairs of clusters show high overlap in terms of significant pathways: cluster 2 and 7, cluster 3 and 8, and cluster 6 and 10. This is explained by the high pairwise correlations between the cluster centroids (Fig. 4), resulting in a large gene overlap. For clusters 2 and 7, the number of identified pathways increased with two and one, respectively, when using MAsC. Of these, PPAR signaling pathway is of high importance for hepatocyte functionality and linked to metabolic disorders and non-alcoholic fatty liver disease (NAFLD) (Gomaraschi et al., 2019). The Lysosome pathway was also identified and, interestingly, lysosomes were in a recent report highlighted as sophisticated signaling centers that govern cell growth, division and differentiation, which are important for hepatocyte differentiation (Lawrence and Zoncu, 2019). Both PPAR and Lysosome were identified in standard K-means clusters 7 and 2, respectively, but appeared in both clusters with the MAsC. This highlights that the cluster profiles have actually captured the same underlying biological mechanisms, and that they could be interpreted jointly. This reduces the number of unique clusters to interpret, thus simplifying biological analysis.

For clusters 3 and 8, the number of identified pathways increased from 6 to 17 and 6–15 with MAsC, respectively. Here the standard K-means clusters had non-overlapping pathways, whereas the MAsC clusters had considerable overlap. These include several immune response and disease-related pathways, but of specific interest for hepatocytes is the Bile secretion pathway, as bile secretion is an important function in this cell type (Boyer, 2013). For clusters 6 and 10, the number of identified pathways changed from 9 to 8 and 6–9 with MAsC, respectively. The overlap between the pathways is stronger with MAsC, though the biological relevance of the result requires further investigation. One pathway of interest is Alcoholism (also identified with MAsC in cluster 5), which is of high importance in alcohol-induced liver disease (ALD), including liver cirrhosis severe steatohepatitis (Tilg et al., 2011).

For two clusters, 4 and 9, no pathways were identified as significant with MAsC, whereas several pathways were found with standard K-means. However, the pathways identified by K-means have no obvious association with hepatocyte development or liver functionality, suggesting that they were identified as a consequence that some genes may have been forced into clusters 4 and 9, as all genes are assigned to a cluster in standard K-means. With MAsC, the pathways were not found because these genes had lower correlations to the cluster centroids and were dropped from the analysis. Indeed, the cluster sizes for cluster 4 and 9 dropped from 279 and 192 genes to 96 and 92 genes, respectively.

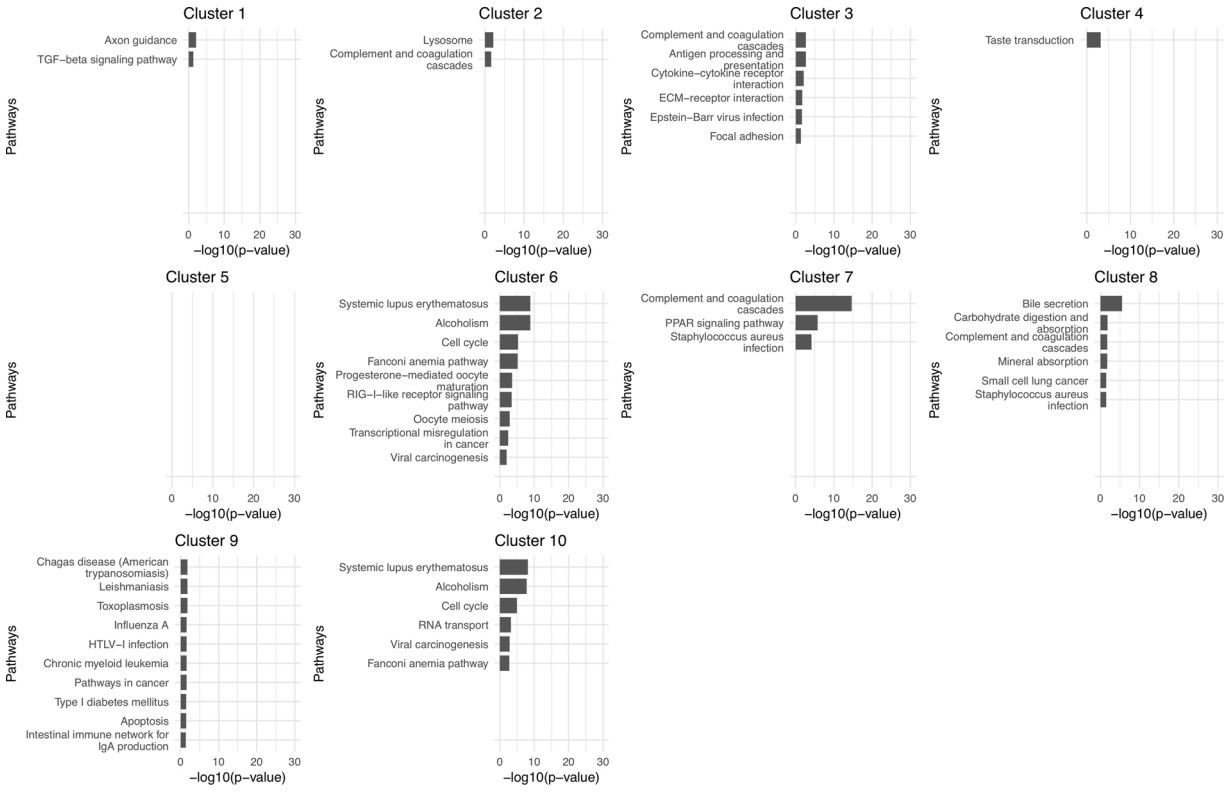

Fig. 8. Significantly enriched pathways for the Hepatocyte dataset using K-means.
The top ten significant pathways for the Hepatocyte dataset are shown for the standard K-means clustering. Only clusters with at least one significant pathway in one of the clustering analyses are shown.

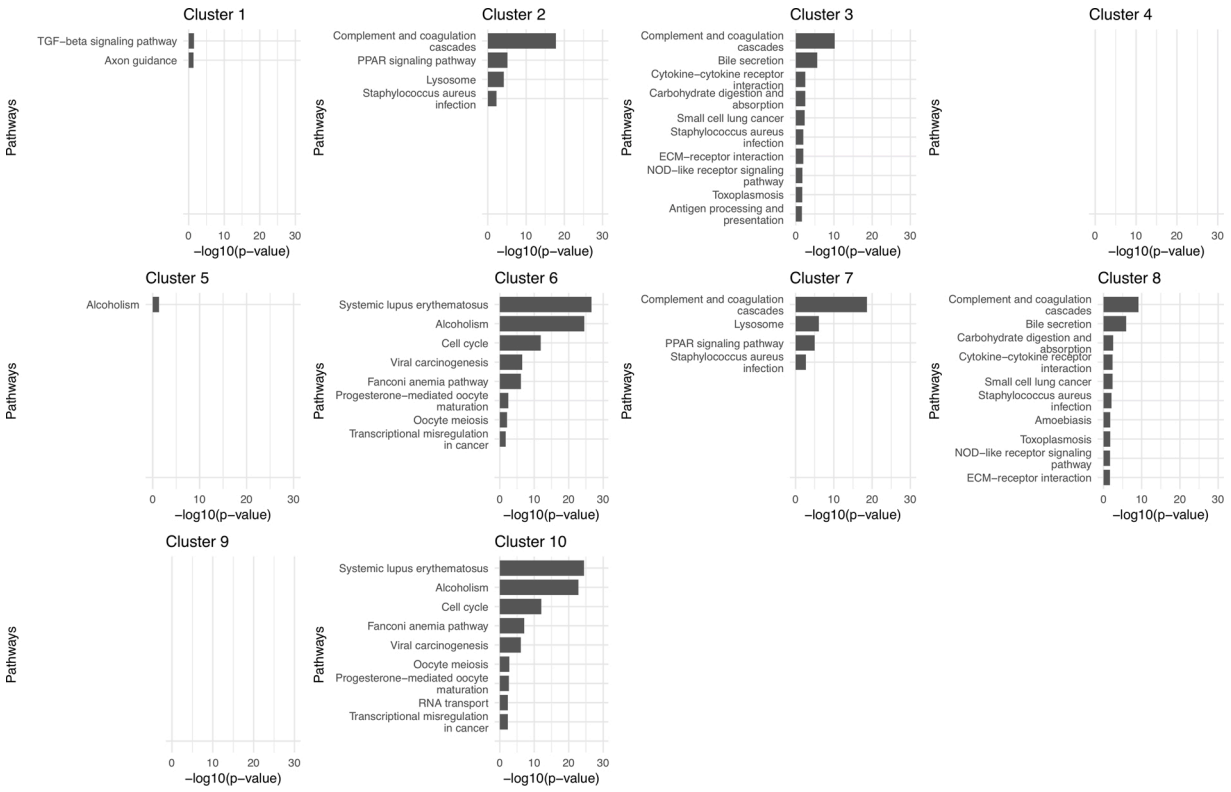

Fig. 9. Significantly enriched pathways for the Hepatocyte dataset using MAsC.
The top ten significant pathways for the Hepatocyte dataset are shown for MAsC ($\delta = 0.9$). Only clusters with at least one significant pathway in one of the clustering analyses are shown.

Since this results in elimination of irrelevant pathways, it clearly demonstrates that MAsC also has a noise reduction effect on the clustering analysis.

4. Discussion

Most genes are involved in a large number of functions and generally participate in many biological processes and different molecular pathways. Notably, this phenomenon is generally not reflected in commonly used clustering approaches. The MAsC algorithm has been developed to address this limitation, and the study presented here demonstrates how multi-assignment clustering produces partitions that better capture the processes of the underlying biological system. The allowance of multiple assignments of genes to more than one cluster produced larger cluster sizes, result of higher relevance to the studied system and contributed to improved biological interpretation of clustering. Related methods with multiple class membership have been elaborated on in previous research, but their solution included a two-step clustering, which did not allow assignment of genes to multiple groups (Bandyopadhyay et al., 2007). Cluster results can be evaluated in multiple ways with quantitative measurements, using visualization tools and various statistical tests. However, when it comes to the assessment of the utility of the results and the biological interpretation of the clustering partitions, a qualitative evaluation of the biological relevance is more informative. We argue that our multi-assignment clustering approach generates partitions of higher biological relevance and should be considered a powerful complement to standard K-means clustering for high-throughput transcriptomics data. A direction for future research is to evaluate this approach on other modalities of omics data, such as proteomics and epigenomics.

A strength of MAsC is the increased power of the downstream functional annotation analysis that is commonly the basis for biological interpretation. Since enrichment analysis generally requires around 200 genes (Huang et al., 2009), the benefits from MAsC will be greater when applied to initially smaller clustering partitions. We also observed improvements for larger partitions, but the effects were not as pronounced. Another advantage with MAsC is that it performs noise reduction on the clustering partition. In the standard K-means clustering, all genes in the dataset are forced into a cluster even when they have low similarity to all the centroids. With MAsC, genes are only assigned to a cluster if the correlation to the centroid is $\geq \delta$, otherwise the gene is excluded from the clustering analysis. This property may be greater when working with larger datasets, and provides an easy way to increase the signal-to-noise ratio, which contributes to generating biologically meaningful partitions.

A limitation with enrichment analysis is its dependency on information stored in annotation databases, and there is a risk that incompleteness of information negatively impacts the results. For example, well-studied processes and pathways that have many annotated genes will be overrepresented, and analyses may become biased to report these as significant. Other biologically important processes are missed due to lack of annotations. It is therefore challenging to apply this kind of analysis to systems that are less studied. A shortcoming with MAsC, inherited from K-means, is the random initialization of centroids, which means that the final clustering partitions will differ slightly between repeated analyses of the same data. This renders interpretation more difficult and there is a risk that biologically important genes are removed in the analysis if they have low correlation to all centroids. Selection of an appropriate correlation threshold δ is challenging and has a large impact on the resulting partitioning. One way to address this is to use a step-wise parameter search and set a threshold that gives satisfactory cluster sizes. The choice will depend on the properties of the data, and we therefore recommend a parameter search for each dataset. Another consequence of the random initialization of centroids is a risk that two or more of them have a high correlation, which will result in high overlap of genes between the clusters. These clusters may be

considered redundant and could be merged into a single cluster to facilitate interpretation. However, this was not considered in the present work since it would limit the possibility to compare the results from standard K-means with MAsC. Merging redundant clusters represents an avenue for future development of this method.

5. Conclusions

Here we present a multi-assignment extension to the K-means clustering method that allows genes to be assigned into more than one cluster. MAsC produced clustering partitions with higher biological relevance compared to the original K-means clusters, as evident from the larger number of statistically significant pathways related to the cellular systems studied. The improved sensitivity in the analysis can be attributed to larger cluster sizes and that all genes highly correlated to a cluster centroid will be assigned to that cluster. The noise reduction property of MAsC further improves the resulting partitions, by removal of genes with low similarity to all of the clusters, rather than forcing them into the closest cluster. We believe MAsC is a powerful complement to existing clustering methods and can support data interpretation in future high-throughput profiling studies.

Data statement

The datasets analyzed during the current study are available in the ArrayExpress repository, namely the Mesoderm dataset with accession number E-MTAB-5219 (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-5219/) and the Hepatocyte dataset with accession number E-MTAB-5367 (https://www.ebi.ac.uk/arrayexpress/experiments/ E-MTAB-5367/).

CRediT authorship contribution statement

Benjamin Ulfenborg: Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing - original draft, Writing - review & editing. **Alexander Karlsson:** Investigation, Methodology, Formal analysis, Visualization, Writing - original draft, Writing - review & editing. **Maria Riveiro:** Investigation, Methodology, Formal analysis, Visualization, Writing - original draft, Writing - review & editing. **Christian X. Andersson:** Writing - original draft, Writing - review & editing. **Peter Sartipy:** Writing - original draft, Writing - review & editing. **Jane Synnergren:** Investigation, Methodology, Formal analysis, Funding acquisition, Project administration, Resources, Visualization, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the University of Skövde, Sweden, the Knowledge Foundation [2014/0301 and 2017/0302], and Takara Bio Europe, Gothenburg. The Knowledge Foundation had no role in the study. Takara Bio Europe contributed with data interpretation and writing of the manuscript.

References

Adesse, D., Goldenberg, R.C., Fortes, F.S., Iacobas, D.A., Iacobas, S., de Carvalho, A.C.C., de Narareth Meirelles, M., Huang, H., Soares, M.B., Tanowitz, H.B., et al., 2011. Gap junctions and chagas disease. Advances in Parasitology. Elsevier, pp. 63–81.
Aggarwal, C.C., 2014. Data Classification: Algorithms and Applications. CRC press.
Bandyopadhyay, S., Mukhopadhyay, A., Maulik, U., 2007. An improved algorithm for clustering gene expression data. Bioinformatics 23, 2859–2865.
Bosman, F.T., Stamenkovic, I., 2003. Functional structure and composition of the extracellular matrix. J. Pathol. A J. Pathol. Soc. Gt. Britain Irel. 200, 423–428.
Boyer, J.L., 2013. Bile formation and secretion. Compr. Physiol. 3, 1035–1078.
Carvalho, B.S., Irizarry, R.A., 2010. A framework for oligonucleotide microarray preprocessing. Bioinformatics 26, 2363–2367.
Caussinus, H., Fekri, M., Hakam, S., Ruiz-Gazen, A., 2003. A monitoring display of multivariate outliers. Comput. Stat. Data Anal. 44, 237–252.
Cojan-Minzat, B.O., Zlibut, A., Agoston-Coldea, L., 2020. Non-ischemic dilated cardiomyopathy and cardiac fibrosis. Heart Fail. Rev. 1–21.
Devkota, P., Wuchty, S., 2020. Controllability analysis of molecular pathways points to proteins that control the entire interaction network. Sci. Rep. 10, 1–9.
Fallah, A., Sadeghinia, A., Kahroba, H., Samadi, A., Heidari, H.R., Bradaran, B., Zeinali, S., Molavi, O., 2019. Therapeutic targeting of angiogenesis molecular pathways in angiogenesis-dependent diseases. Biomed. Pharmacother. 110, 775–785.
Ferraro, M.B., Giordani, P., 2015. A toolbox for fuzzy clustering using the R programming language. Fuzzy Sets Syst. 279, 1–16.
Fu, L., Medico, E., 2007. FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. BMC Bioinformatics 8, 3.
Gan, G., Ma, C., Wu, J., 2007. Data clustering: theory, algorithms, and applications. Siam.
Ghosheh, N., Küppers-Munther, B., Asplund, A., Edsbagge, J., Ulfenborg, B., Andersson, T.B., Björquist, P., Andersson, C.X., Carén, H., Simonsson, S., Sartipy, P., Synnergren, J., 2017. Comparative transcriptomics of hepatic differentiation of human pluripotent stem cells and adult human liver tissue. Physiol. Genomics 49. https://doi.org/10.1152/physiolgenomics.00007.2017.
Gomaraschi, M., Fracanzani, A.L., Dongiovanni, P., Pavanello, C., Giorgio, E., Da Dalt, L., Norata, G.D., Calabresi, L., Consonni, D., Lombardi, R., et al., 2019. Lipid accumulation impairs lysosomal acid lipase activity in hepatocytes: evidence in NAFLD patients and cell cultures. Biochim. Biophys. Acta (BBA)-Molecular Cell Biol. Lipids 1864, 158523.
Gordeeva, O., 2019. TGFβ family signaling pathways in pluripotent and teratocarcinoma stem cells' fate decisions: balancing between self-renewal, differentiation, and cancer. Cells 8, 1500.
Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: a k-means clustering algorithm. J. R. Stat. Soc. Ser. C (Applied Stat. 28, 100–108.
Huang, D.W., Sherman, B.T., Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4, 44.
Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recognit. Lett. 31, 651–666.
Lawrence, R.E., Zoncu, R., 2019. The lysosome as a cellular centre for signalling, metabolism and quality control. Nat. Cell Biol. 21, 133–142.
Li, Y., Chen, L., 2014. Big biological data: challenges and opportunities. Genomics Proteomics Bioinformatics 12, 187.
Liu, Z., Pan, Q., Dezert, J., Mercier, G., 2015. Credal c-means clustering method based on belief functions. Knowledge-based Syst. 74, 119–132.
Meng, C., Zeleznik, O.A., Thallinger, G.G., Kuster, B., Gholami, A.M., Culhane, A.C., 2016. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief. Bioinform. 17, 628–641.
Mitchell, T.M., 1999. Machine learning and data mining. Commun. ACM 42, 30–36.
Nikolaou, K., Sarris, M., Talianidis, I., 2013. Molecular pathways: the complex roles of inflammation pathways in the development and treatment of liver cancer. Clin. Cancer Res. 19, 2810–2816.
Paul, M., Schulze-Bahr, E., 2020. Arrhythmogenic right ventricular cardiomyopathy: evolving from unique clinical features to a complex pathophysiological concept. Herz.
Rodriguez, M.Z., Comin, C.H., Casanova, D., Bruno, O.M., Amancio, D.R., Costa, L., da, F., Rodrigues, F.A., 2019. Clustering algorithms: a comparative approach. PLoS One 14.
Si, Y., Liu, P., Li, P., Brutnell, T.P., 2014. Model-based clustering for RNA-seq data. Bioinformatics 30, 197–205.
Sulakhe, D., Balasubramanian, S., Xie, B., Berrocal, E., Feng, B., Taylor, A., Chitturi, B., Dave, U., Agam, G., Xu, J., et al., 2014. High-throughput translational medicine: challenges and solutions. Systems Analysis of Human Multigene Disorders. Springer, pp. 39–67.
Tarca, A.L., Draghici, S., Khatri, P., Hassan, S.S., Mittal, P., Kim, J., Kim, C.J., Kusanovic, J.P., Romero, R., 2009. A novel signaling pathway impact analysis. Bioinformatics 25, 75–82.
Tilg, H., Moschen, A.R., Kaneider, N.C., 2011. Pathways of liver injury in alcoholic liver disease. J. Hepatol. 55, 1159–1161.
Ulfenborg, B., Karlsson, A., Riveiro, M., Améen, C., Åkesson, K., Andersson, C.X., Sartipy, P., Synnergren, J., 2017. A data analysis framework for biomedical big data: application on mesoderm differentiation of human pluripotent stem cells. PLoS One 12. https://doi.org/10.1371/journal.pone.0179613.
Wenskovitch, J., Crandell, I., Ramakrishnan, N., House, L., North, C., 2017. Towards a systematic combination of dimension reduction and clustering in visual analytics. IEEE Trans. Vis. Comput. Graph. 24, 131–141.
Winslow, R.L., Walker, M.A., Greenstein, J.L., 2016. Modeling calcium regulation of contraction, energetics, signaling, and transcription in the cardiac myocyte. Wiley Interdiscip. Rev. Syst. Biol. Med. 8, 37–67.
Wu, G.P.K., Chan, K.C.C., Wong, A.K.C., 2011. Unsupervised fuzzy pattern discovery in gene expression data. in: BMC Bioinformatics S5.

Xu, R., Wunsch, D.C., 2010. Clustering algorithms in biomedical research: a review. IEEE Rev. Biomed. Eng. 3, 120–154.
Zhang, W., Liu, H.T., 2002. MAPK signal pathways in the regulation of cell proliferation in mammalian cells. Cell Res. 12, 9–18.
Zhang, M., Adamu, B., Lin, C.-C., Yang, P., 2011. Gene expression analysis with integrated fuzzy C-means and pathway analysis. 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society 936–939.