Title

Clust: automatic extraction of optimal co-expressed gene clusters from gene expression data

Authors

Basel Abu-Jamous and Steven Kelly

Affiliations

Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, OX1 3RB, UK

Corresponding Author

Email: steven.kelly@plants.ox.ac.uk; Telephone: +44 (0)1865 275123

Abstract

Identification of co-expressed genes within a given experimental or biological context can provide evidence for genetic or physical interactions between genes. Thus, detection of co-expression has become a routine step in large-scale analyses of gene expression data. In this work, we show that application of the most commonly used methods to identify co-expressed gene clusters produce results that do not match the biological expectations of co-expressed gene clusters. Specifically, clusters generated using these methods are not discrete and can contain up to 50% unreliably assigned genes. Consequently, downstream analyses on these clusters, such as functional term enrichment analysis, suffer from high error rates. We present *clust*, an automated method that solves this problem by extracting clusters from gene expression datasets that match the biological expectations of co-expressed genes. Using 100 gene expression datasets from five model organisms we demonstrate that the statistical properties of clusters generated by *clust* are better than those produced by other methods. We further show that this improvement results in a concomitant improvement in detection of enriched functional terms.

Keywords: clustering

Introduction

Gene transcription is dynamically and coordinately regulated in all living organisms. Such coordinate regulation is manifest as concordant changes in the transcript abundance of genes in time series and perturbation-response datasets. Gene transcription is regulated by the binding of transcription factors to DNA/chromatin elements located in promoter or enhancer regions of genes. Typically, transcription factors comprise ~10% of the total number of genes in a genome, and complex spatio-temporal patterns of transcription are achieved through the combinatorial action of these genes in regulatory networks. The combinatorial nature of these networks means that their behaviour is inherently conditional. That is, genes that appear co-expressed under one condition are not necessarily co-expressed under all conditions. A corollary of this is that within any one experimental context (e.g. time series spanning some biological process or perturbation-response experiment) not all genes will be behaving coordinately. Instead, subsets of genes have the right combination of regulators to behave coordinately during the experimental context while others are following patterns of regulation that are independent of the experimental design. Thus, within a given observation window (i.e. experimental context) it is not expected that all genes can be assigned to a limited set of coordinate behaviours (Nilsson, et al., 2009; Pierson, et al., 2015).

Given that only subsets of genes are likely to be co-expressed within a particular context, it follows that identification of these subsets is a data extraction problem and not a data partitioning problem. That is, the aim is to identify and extract the cohorts of genes that are behaving coordinately from the complete set of genes that are detected within a particular context, and is not to partition the complete set of genes into a set of gene clusters. In practice, clustering methods have been widely applied to gene expression data with the expectation that they will identify the complete set of discrete cohorts of genes that have co-ordinated behaviours (i.e. the clusters of co-expressed genes), and that the all of genes that exhibit those behaviours will be assigned to the correct cluster. However, the vast majority of methods that aim to identify cohorts of co-expressed genes are based on data partitioning (e.g.

Markov clustering (Enright, et al., 2002), k-means (MacQueen, 1967), hierarchical clustering (Eisen, et al., 1998), and self-organising maps (Kohonen, 1982)). These approaches attempt to assign all detected genes to a finite set of clusters, with the number of clusters determined by numerical optimisation of a data partitioning metric (Ronan, et al., 2016). Thus, genes that are not co-expressed in the context under investigation are also assigned to their "best-fitting" cluster such that the majority of clusters will contain both co-expressed and non-co-expressed genes. This result does not adhere to the expectation of the biological properties of a co-expressed gene cluster. i.e. that each cluster contains only those genes that exhibit co-ordinate behaviour in the experimental or biological context under question and that no two clusters should have an identical profile.

Here we show through analysis of 100 real biological datasets from five model organisms that application of data partitioning-based clustering methods to gene expression data generates clusters that include substantial numbers of unreliably assigned genes, i.e. genes that should have been excluded. Such unreliable content comprises up to about 50% of these clusters. To address this problem we provide a novel method called *Clust* for cluster extraction from gene expression data. *Clust* is designed to produce co-expressed clusters of genes that satisfy the biological expectations of a co-expressed gene cluster. We show that *Clust* satisfies these expectations while producing co-expressed clusters with lower levels of dispersion than any data partitioning method. We also show that the clusters produced by *Clust* do not contain unreliably clustered genes typical of data partitioning methods. Furthermore, through functional term enrichment analysis we show that application of *Clust* results in clusters that contain the largest number of reliable enriched functional terms.

Results

Problem definition, aim and approach

Gene expression datasets (RNAseq and microarray) contain quantitative estimates (observations) of mRNA abundance for a set of genes at multiple experimentally, spatially, or temporally discrete

conditions. Across these conditions, it is expected that the mRNA abundance of transcriptionally co-regulated genes will exhibit coordinate behaviour. These co-regulated cohorts of genes include those that are inherent modules of the system being studied, as well as those that may be conditional on applied experimental perturbations. The observations also include transcript abundance estimates for genes that are behaving independently in the experimental series. Furthermore., for genes that are transcriptionally co-regulated, variance in RNA processing and half-life cause fluctuations in transcript abundance such that abundance estimates for are inherently noisy. Thus, the goal of gene expression clustering is to identify and extract the discrete cohorts of genes whose transcripts are behaving coordinately (albeit with biological noise) across the observations under consideration.

Fig. 1 presents simulated gene expression data to illustrate the problem of extracting distinct cohorts of co-expressed genes. Each simulated dataset contains 500 genes, with 100 genes in each of three distinct clusters and 200 genes that do not belong to any cluster. Fig. 1a shows the same clusters simulated with increasing levels of biological noise (D1 to D4) and Fig. 1b shows the desired results. That is, to extract three distinct clusters of genes (C1 to C3) while discarding the genes that behave independently. In conflict with the desired goal, data partitioning methods require all genes to be included in one of the clusters. For example, application of *k*-means (the most commonly used method for analysing gene expression datasets) recovers the three simulated profiles. However, each cluster also contains a large cohort of genes that do not share the same expression profile. This inclusion results in clusters with high levels of dispersion and high levels of inter-cluster similarity, such that many genes that are assigned to a given cluster fit entirely within the profile of, and thus can be justifiably assigned to, one or more other clusters (Fig. 1c). This result violates the expectations of co-expressed gene clusters, and produces clusters whose gene assignment is unreliable. *Clust* is designed to address this problem by extracting the largest and least dispersed set of clusters whose profiles are distinct and exclude those genes that do not belong to these clusters. That is, to identify

and extract the complete set of genes that are exhibiting coordinate behaviour in the experimental series under consideration.

Data sources and comparative methods

To demonstrate the performance characteristics of *clust* on real biological datasets, the method was applied to 100 different gene expression datasets (Supplementary Table S1). These datasets comprised ten microarray datasets and ten RNAseq datasets from each of five different model organisms; *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Arabidopsis thaliana*, and *Saccharomyces cerevisiae*. To put these performance characteristics in context, five of the most commonly used co-expression clustering methods (*k*-means, hierarchical clustering (HC), self-organising maps (SOMs), Markov clustering (MCL), and WGCNA) were also applied to these datasets. For each of these additional methods, the best-practice operating procedures were followed as described in Methods.

Clust robustly produces tight and non-overlapping clusters

As *clust* is a cluster extraction method, it does not necessarily assign all genes to clusters. On average across the 100 test datasets *clust* assigned 50% of the input genes to clusters (Fig. 2a), and produced sets of clusters that have significantly lower dispersion than those produced by MCL (p-value 2.5×10^{-31}), k-means (p-value 4.6×10^{-7}), HC (p-value 2.4×10^{-13}), WGCNA (p-value 3.1×10^{-64}), or SOMs (p-value 4.8×10^{-24}) (Fig. 2b). Clusters produced by *clust* are discrete, such that genes assigned to one cluster do not fit within the profile boundaries of any other cluster (JI = 0 for all clusters, Fig. 2c). This is not the case for data partitioning methods, where 10% to 50% of the genes that are included in a given cluster also fit within the boundaries of at least one other cluster (Fig. 2c). Thus, application of data partitioning methods to gene expression data produces clusters that are not discrete and contain between 10 and 50% unreliably assigned genes (Supplementary Table S2).

In addition to producing datasets that are both lower in dispersion and discrete, the distribution of the properties of *clust* clusters is also unimodal (Fig. 2d-f and Supplementary Table S3). In contrast,

cluster dispersion is multimodal for MCL, HC, WCGNA, and SOMs (Fig. 2d); cluster overlap is multi-modal or uniformly distributed for MCL, KM, HC and SOMs (Fig. 2e); cluster size is biased towards small clusters for MCL and WCGNA (Fig. 2f). Thus, the individual clusters returned by data partitioning methods are inconsistent and vary considerably in their quality. In contrast, the quality of clusters produced by *clust* is unimodal such that all clusters can be considered equally valid. This property of *clust* clusters is independent of the number of genes in a given cluster (Supplementary Figures S1 and S2). In contrast, the properties of clusters returned by data partitioning methods display a significant dependency on cluster size, such that larger clusters have both higher dispersion and a higher proportion of genes that fit within the boundaries of other clusters (Supplementary Figure S1).

None of the six methods, including *clust*, behaves differently on datasets from different species (Supplementary Figures S3 and S4). Therefore, the species from which the data was produced is not a factor that affects the performance of any of these clustering methods. However, the dispersion of clusters produced by *k*-means, HC, and SOMs, is dependent on the number conditions under consideration such that the more conditions being considered the worse the results of the clustering (Supplementary Figure S5). In contrast, the behaviour of *clust* is unaffected by the number of genes or the number of conditions that are under consideration (Supplementary Figures S5, S6, S7, and S8). Thus, unlike all of the other tested methods the behaviour of *clust* is consistent for different types or quantities or input data.

Clust produces clusters which are rich in robustly enriched functional terms

One of the most commonly applied tests to co-expressed clusters of genes is functional term enrichment, as a cluster of co-expressed genes is expected to be enriched with genes that have related biological roles. As *clust* assigns on average 50% of genes to clusters, it was investigated how this reduction in gene number affects the detection of functional term enrichment. To do this, each of the methods were evaluated for their ability to detect enrichment of GO terms in the *Arabidopsis thaliana* and *Saccharomyces cerevisiae* gene expression datasets. These datasets were selected

because of the rich GO term annotation of these two species genomes, and the consistent gene name annotation of between datasets. To simplify data visualisation, the GO term enrichment results from WCGNA, the worst performing method, was excluded from Venn diagrams and can be found in the Supplementary Tables S4 and S5.

Of the 6,321 significantly enriched GO terms detected by all methods (excluding WCGNA), 3,288 (52%) were detected by two or more methods and 861 GO terms (14%) were detected by all methods (Fig. 3a). Given that the results were considerably different between methods, it is reasonable to assume that GO terms detected by two or more methods for a given input dataset are those that are more likely to be correct, however there was no significant difference between methods in their ability to recover these GO terms (Fig. 3b). Thus, while *clust* only assigns 50% of genes to clusters, this reduction in gene content does not reduce the number of detected enriched functional terms compared to other methods.

To determine whether the qualitative difference in GO term discovery between methods was due to unreliably assigned genes, the clusters produced by the data partitioning methods we further analysed both with all of the unreliably assigned genes removed (the stringent set), and with all genes that fit within the boundaries of the cluster included (the expanded set) (Supplementary Figure S9). Those functional terms that remained significantly associated with a cluster irrespective of whether the stringent or expanded version of the cluster was analysed were deemed robust to unreliable gene content. On average for the data partitioning methods, between 20% and 95% of GO terms are unreliably assigned to clusters (Fig. 3c). As before, those GO terms that were detected by two or more methods were considered more likely to be correct, and comparison between the methods revealed that *clust* recovers largest proportion of these terms (Fig. 3d). Thus, the enriched functional terms associated with co-expressed gene clusters generated by *clust* are more likely to be correct than those associated with clusters produced by other methods.

Discussion

Co-expression clustering is a routinely used step in data exploration for gene expression analysis. Here we show that the most commonly used methods for conducting co-expression analysis do not match the biological expectations of a co-expressed cluster of genes, producing clusters that are highly dispersed and contain large proportions of genes that could be equally assigned to other clusters within the same dataset. Moreover, the methods behave inconsistently, with substantial differences in clustering performance attributable to differences in datatype or data quantity. We present *clust,* as a method designed to solve all of these problems. *Clust* was compared with five commonly used clustering methods (MCL, k-means, HC, WGCNA, and SOMs) by application to 100 different microarray and RNA-seq gene expression datasets from five model species. In contrast to the other tested methods, *clust* behaviour is consistent and is unaffected by species, datatype, number of genes, or number of conditions. Thus, *clust* is also robust to increases in data quantity without sacrificing the quality of the results.

The most commonly conducted post-clustering analysis is to detect enrichment of functional terms within clustered sets of co-expressed genes. We show that conducting such analyses on clusters produced using the most commonly used methods for co-expressed gene clustering results in 20% and 95% of enriched functional terms being unreliably assigned to clusters. This observation has implications for the utility of downstream analysis conducted on these clusters. For example, putative regulatory relationships are often inferred by identifying regulatory genes that occur in clusters that are enriched for specific functional terms. Thus, unreliability of enriched functional term assignment likely contributes to the high false positive discovery rate in the discovery rate of regulatory interactions from co-expression data (Faith, et al., 2007). As *clust* is designed to solve the problem of reliability of gene assignment to clusters, it does not suffer from such unreliability in enriched functional terms, and therefore represents a solution to this problem. Moreover, *clust* detects the largest number of reliably assigned enriched functional terms of all the methods tested.

Methods

Overview of the clust cluster extraction method

Clust has a pipeline of steps that extract final optimised clusters of co-expressed genes from a gene expression dataset. First, *clust* employs a number of base clustering methods (e.g. k-means clustering, hierarchical clustering, and self-organising maps) to produce initial sets of clusters. Each method is subject to a wide parameter sweep so that multiple clustering results from each method are generated. These initial clusters are provided as input to construct consensus "seed" clusters using Bi-CoPaM (Abu-Jamous, et al., 2013). All seed clusters are evaluated by the M-N distance (MND) metric (Abu-Jamous, et al., 2015) which consider both within-cluster dispersion and cluster size, and the set of non-overlapping clusters that minimise MND and maximise cluster size are selected as elite seed clusters. The final step of the algorithm removes outliers from elite seed clusters using a Tukey filter, defines the cluster profile based on the range of expression values observed within the cluster, and then assigns all genes from the input dataset that fit within this cluster profile. The full details of *clust* are provided in Supplementary Text S1, a standalone Python implementation of *Clust* is available at https://github.com/BaselAbujamous/clust.

Selection of 100 gene expression datasets

The 100 gene expression datasets were downloaded from the Gene Expression Omnibus (GEO) repository on 2nd of July 2017 (NCBI Resource Coordinators, 2017). For each one of the five model species, ten microarray datasets and ten RNAseq datasets were downloaded. In all cases the most recently published datasets for each of these species was selected, given that the dataset had at least 4 different conditions (time-points or treatments) and no more than 50 samples including replicates. RNAseq datasets were chosen only if the resulting TPM, RPKM, FPKM, or CPM quantitation files were available from the GEO repository. Microarray datasets were a mix of both one-colour or two-colour microarrays. The complete list of the 100 datasets and their properties is available in Supplementary Table S1.

Implementation of clust and comparative methods

k-means was run using the Python *sklearn.cluster* implementation. The Python *mcl* package was used to run MCL (van Dongen, 2001). The Python *scipy.cluster.hierarchy* package was used to run HC clustering. The *blockwiseModules* module of the R *WGCNA* library was used to run WGCNA. The Python *sompy* package was used to run SOMs. The Python package *clust 1.2.0* was used to run *clust*. Each of the methods was run using their default parameters. However, all of them, except for MCL, require pre-setting the number of clusters (*k*). When *k* is unknown, it is a common practice to test a range of *k* values and to choose the one which optimises some cluster validation index. Therefore, we applied each of methods with all *k* values from 2 to 50 and evaluated them by using the widely used cluster validation index, the Davies–Bouldin (DB) index (Davies & Bouldin, 1979). For each dataset, the cluster set generated from the k value that minimises this index was selected for analysis.

Cluster dispersion metric (MSE)

The mean squared error (MSE) metric is used to measure within-cluster dispersion. If the cluster has *N* genes and the dataset has *D* dimensions, the MSE value for that cluster will be:

$$MSE = \frac{1}{D \times N} \sum_{g=1}^{N} \left\| \vec{x}_g - \vec{z} \right\|^2,$$

where $\vec{x}_g$ is a vector of the gene expression profile of the g^{th} gene in this cluster, $\vec{z}$ is a vector of the average expression profile of all genes in this cluster, and $\left\| \vec{x}_g - \vec{z} \right\|$ is the Euclidean distance between these two vectors. Note that the MSE value here is normalised by the number of genes in the cluster. When calculating the MSE value for a whole clustering result (a set of clusters), it is calculated as the weighted average of the MSE values of the each of the clusters, where the weight is the size (number of genes) in each of the clusters.

Cluster similarity metric (JI)

The Jaccard Index (JI) metric is used to measure the similarity amongst the clusters in a clustering result (Jaccard, 1901). JI is calculated as the ratio between the number of “overlap genes” and the number of all genes in clusters. “Overlap genes” are those genes which are included in a cluster while

their expression profiles also fit within the boundaries of at least one other cluster. The upper and the lower boundaries of a cluster at any given dimension (condition) are respectively calculated as the maximum and the minimum expression values of all genes in that cluster after trimming the most extreme 1% values to reduce the effect of outliers.

GO term enrichment analysis

The GO term annotations for *Arabidopsis thaliana* and *Saccharomyces cerevisiae* were downloaded from the Gene Ontology Consortium's online repository at http://www.geneontology.org (Ashburner, et al., 2000; The Gene Ontology Consortium, 2017). Significantly enriched GO terms were taken as those that obtained an adjusted hypergeometric test p-value ≤ 0.001.

Figure legends

Figure 1. Expectations and outcomes for application of data-partitioning methods to co-expression clustering.

(**a & b**) Simulated gene expression data for 500 genes with increasing noise (D1 – D4). (**a**) All genes. (**b**) profiles of the genes in each of the three simulated clusters as well as the extra unclustered genes at each one of the four levels of dispersion. The horizontal axis of each plot represents the six different conditions/samples, while the vertical axis represents gene expression values. (**c**) The results of applying a partitioning method (*k*-means in this case) to the same simulated datasets. (**d**) Heat-maps that show the percentage of genes in a cluster that also fit well within each one of the other clusters.

Figure 2. Evaluation of the performance of clustering methods.

(**a**-**c**) Evaluation of clustering performance over all 100 datasets. (**a**) the percentage of input genes that were included in clusters; (**b**) the average dispersion of clusters measured by weighted-averaging of individual cluster MSE values; (**c**) percentage of the overlap amongst clusters, as measured by JI index. (**d**-**e**) Distributions of individual cluster properties for all 100 datasets. MSE values (**d**), JI values (**e**), and cluster sizes (**f**). (Supplemental Tables S2 and S3 and Supplemental Figures S1 to S8).

Figure 3. Evaluation of GO term enrichment in the results of the clustering methods.

(**a**) Venn diagram demonstrating substantial differences in enriched GO terms between methods. The numbers on this diagram represent the number of GO terms detected as significantly enriched across

the 20 selected datasets. The union of these sets includes 6321 terms, 3033 of which (48%) are exclusive to a single method. (**b**) The percentage of reliable GO terms (i.e. those detected by two or more methods) detected by each method. There is no significant difference between methods (Supplementary Table S4). (**c**) F-scores quantifying the similarity between the set of GO terms detected as enriched in the original clusters and the set of GO terms detected as enriched in clusters after taking into account unreliably assigned genes (Supplementary Table S4). (**d**) as in (b) but considering robust GO terms. *Clust* shows significantly higher values than HC ($p = 0.009$) and SOMs ($p = 0.0009$). Although *clust* also shows higher average values than MCL and k-means, but they are not significantly higher at a p-value threshold of 0.01.

Figures

Figure 1

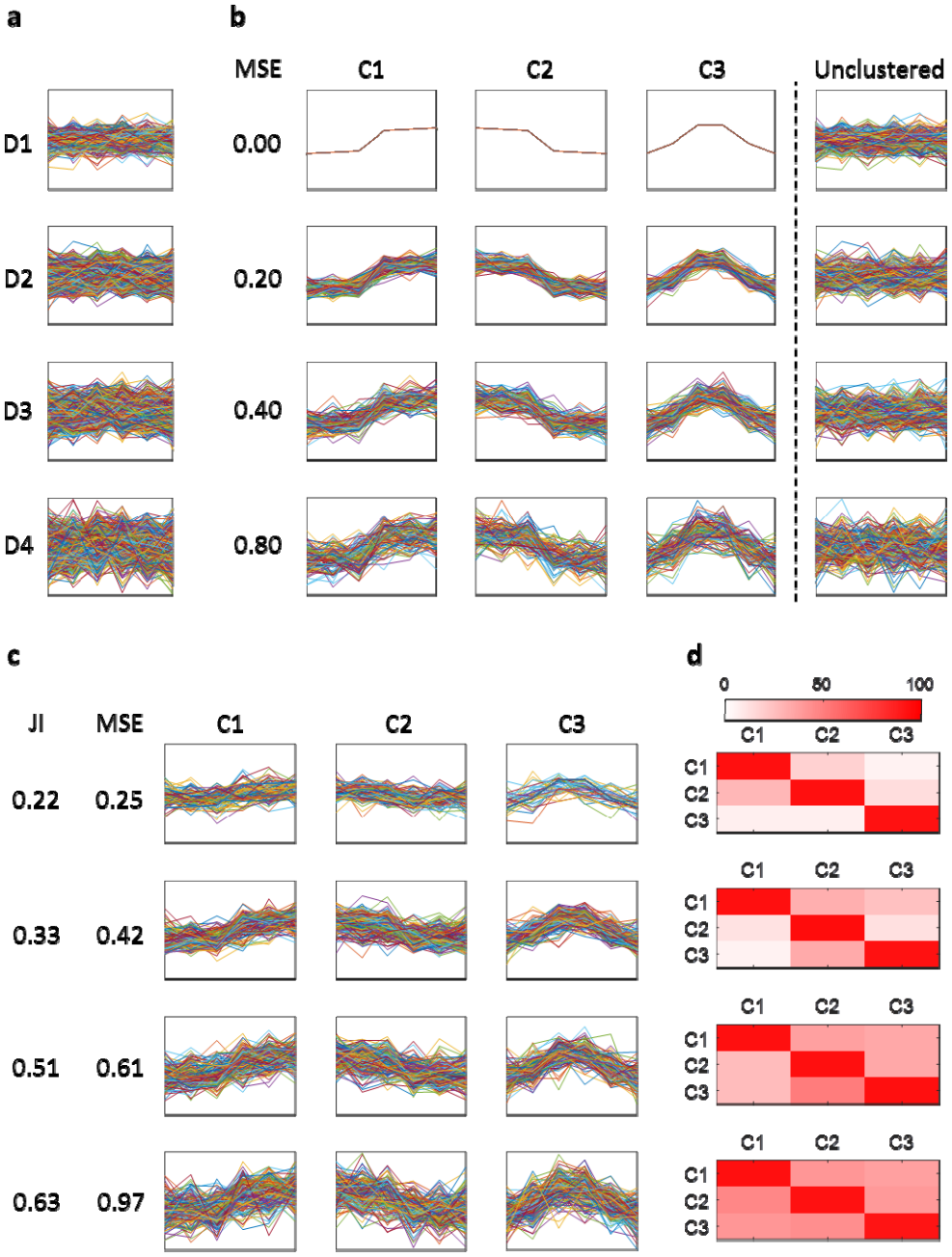


Figure 2

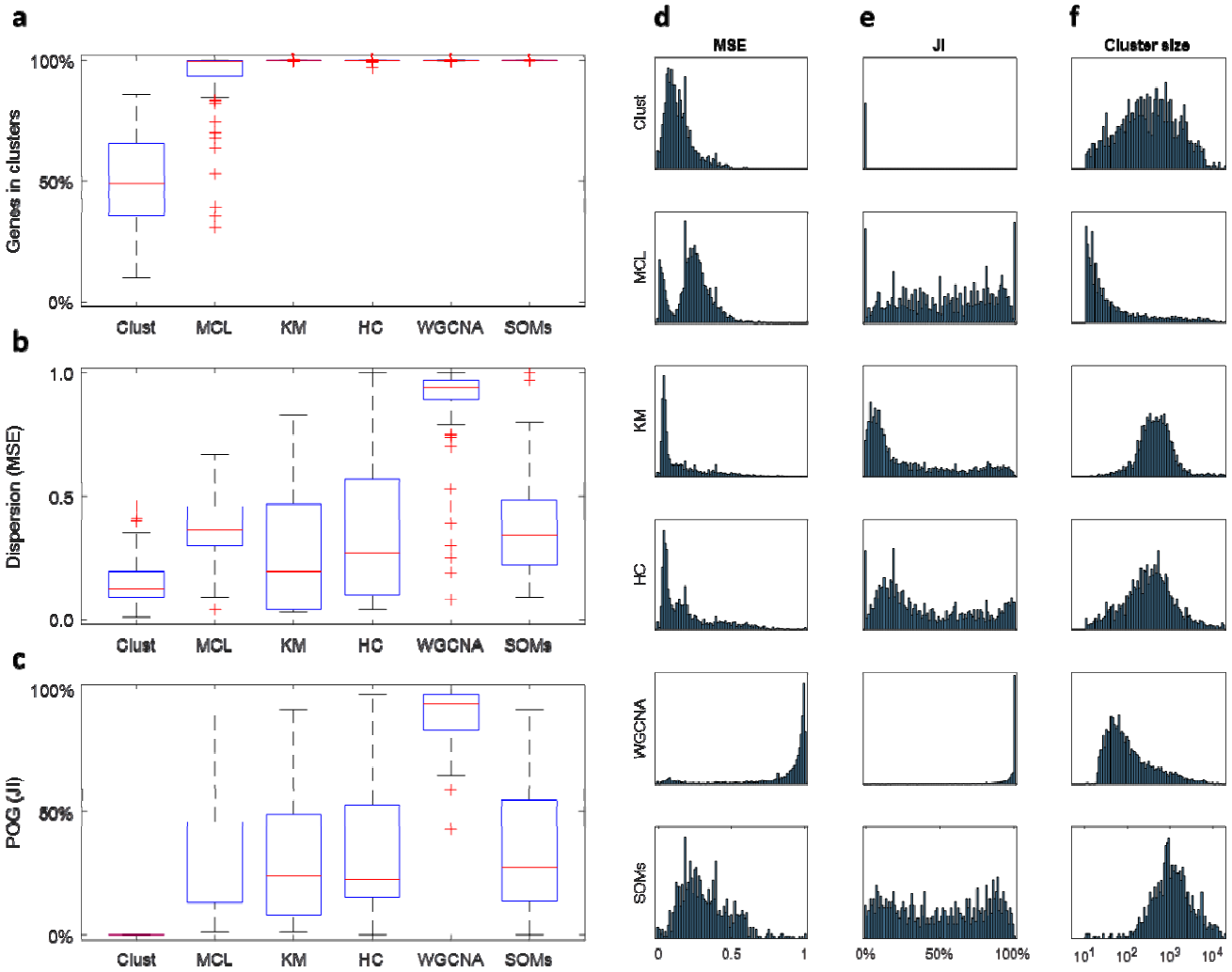
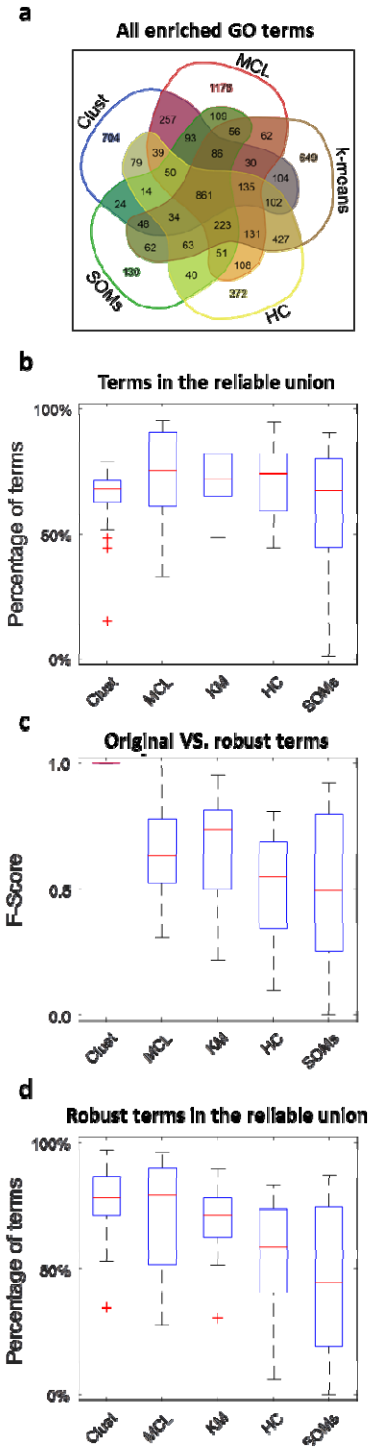


Figure 3



References

Abu-Jamous, B., Fa, R., Roberts, D. J. & Nandi, A. K., 2013. Paradigm of Tunable Clustering using Binarization of Consensus Partition Matrices (Bi-CoPaM) for Gene Discovery. *PLOS ONE,* 8(2), p. e56432.

Abu-Jamous, B., Fa, R., Roberts, D. J. & Nandi, A. K., 2015. UNCLES: method for the identification of genes differentially consistently co-expressed in a specific subset of datasets. *BMC Bioinformatics,* Volume 16, p. 184.

Ashburner, M. et al., 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics,* Volume 25, pp. 25-29.

Davies, D. L. & Bouldin, D. W., 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* PAMI-1(2), pp. 224-227.

Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. *PNAS,* 95(25), p. 14863–14868.

Enright, A. J., Van Dongen, S. & Ouzounis, C. A., 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research,* 30(7), pp. 1575-1584.

Faith, J. J. et al., 2007. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLOS Biology,* 5(1), p. e8.

Jaccard, P., 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles,* Volume 37, pp. 547-579.

Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics,* 43(1), p. 59–69.

MacQueen, J., 1967. *Some methods for classification and analysis of multivariate observations.* s.l., University of California Press, p. 281–297.

NCBI Resource Coordinators, 2017. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research,* 45(Database), pp. D12-D17.

Nilsson, R. et al., 2009. Discovery of genes essential for heme biosynthesis through large-scale gene expression analysis. *Cell Metabolism,* 10(2), p. 119–130.

Pierson, E. et al., 2015. Sharing and specificity of co-expression networks across 35 human tissues. *PLOS Computational Biology,* 11(5), p. e1004220.

Pirim, H., Ekşioğlu, B., Perkins, A. D. & Yüceer, Ç., 2012. Clustering of high throughput gene expression data. *Computers & Operations Research,* 39(12), pp. 3046-3061.

Ronan, T., Qi, Z. & Naegle, K. M., 2016. Avoiding common pitfalls when clustering biological data. *Science Signalling,* 9(432), p. re6.

The Gene Ontology Consortium, 2017. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research,* 45(D1), p. D331–D338.

van Dongen, S., 2001. *Graph clustering by flow simulation [PhD Thesis].* Utrecht: Utrecht University Repository.