Supplementary Materials for

Robust clustering of noisy high-dimensional gene expression data for patients subtyping

Pietro Coretto*[1], Angela Serra†[2], and Roberto Tagliaferri‡[2]

[1]DISES, University of Salerno, Fisciano (SA), Italy
[2]NeuRoNe Lab, DISA-MIS, University of Salerno, Fisciano (SA), Italy

*pcoretto@unisa.it - equal contributions
†aserra@unisa.it - equal contributions
‡robtag@unisa.it

Contents

1 Data set collection and preparation

The experiments were performed on the same 5 real cancer data sets coming from the TCGA database available throught the Genomic Data Commons portal (https://portal.gdc.cancer.gov/) used in the SNF paper [6] (see Table 1). The data sets are the TCGA curated level-3 data of the GBM, BIC, LSCC, KRCCC and COAD cancer on which the SNF authors performed three steps of preprocessing: outlier removal, missing-data imputation and normalization. As a further step, features with low variance were eliminated.

Data set	NGe	NPat	NClusters
GLIO	2409	205/215	3
BREAST	3563	89/105	5
LUNG	2409	96/215	4
KIDNEY	3580	89/122	3
COLON	3563	92/92	3

Table 1: For each data set the number of genes (NGe) and patients (NPat) and the number of clusters (K) used in the analyses are reported.

2 Similarity Network Fusion

SNF [6] is an intermediate integration network fusion methodology able to integrate multiple genomic data (e.g., mRNA expression, DNA methylation and microRNA expression data) to identify relevant patient subtypes. The method first constructs a patient similarity network for each view. Then, it iteratively updates the network with the information coming from other networks to make them more similar at each step. At the end, this iterative process converges to a final fused network. The authors tested the method combining mRNA expression, microRNA expression and DNA methylation from five cancer data sets. They showed that the similarity networks of each view have different characteristics related to patient similarity while the fused network gives a clearer picture of the patient clusters. They compared the proposed methodology with iClust and the clustering on concatenated views. Results were evaluated with the silhouette score for clustering coherence, Cox log-rank test p-value for survival analysis for each subtype and the running time of the algorithms.

The data used in this study are the same on which the SNF algorithm was tested in the original paper. The difference is that our methodology only uses the gene expression data. SNF, which is a more general procedure, able to integrate different data layers, in this comparison, was applied on the same multi-view data sets and by using the same parameters identified by the authors in their original work.

3 TMIX

The TMIX approach was introduced in [5] and extensively treated in [4]. The notation here is as in Section 2.2 of the main paper if not otherwise stated. The data generating process is represented as finite mixture

of Student's t-distribution

$$m(\boldsymbol{y};\boldsymbol{\eta}) = \sum_{j=1}^{k} \pi_j f(\boldsymbol{y};\nu_j,\boldsymbol{\mu}_j,\boldsymbol{S}_j),$$

where $f(\cdot)$ is the multivariate non-central Student's t-distribution with ν_j degrees of freedom, mean vector $\boldsymbol{\mu}_j$ and scale matrix $\boldsymbol{S}_j$. Scale matrices are scaled version of cluster's covariance matrices. The parameters π_j are interpreted as expected clusters' sizes as usual. The unknown parameter vector $\boldsymbol{\eta}$ includes the quartets $(\pi_j,\nu_j,\boldsymbol{\mu}_j,\boldsymbol{S}_j)$ for each j. The parameter ν_j controls the tail behavior of the jth clusters, and therefore allows to accommodate heavy tails and outliers. $\boldsymbol{\eta}$ is fitted based on maximum likelihood estimation (MLE) which can be computed by EM-type algorithms [see 4]. Once $\boldsymbol{\eta}$ is estimated, points are assigned based on the optimal Bayes assignment rule (equation (3) in the main paper). In this paper TMIX clustering is performed using the `EMMIXskew` software of [7].

Although the method cannot cope with arbitrary data contamination [3], it is robust in practice in most situations. Although both TMIX and OTRIMLE are robust model-based clustering methods, they are conceptually and methodologically different.

- TMIX does not treat the noise on its own, because noise and outliers are captured by the tails of each clusters. Therefore, for the jth clusters the parameters $(\nu_j,\boldsymbol{\mu}_j,\boldsymbol{S}_j)$ capture both regular and non regular points.
- TMIX does not have a clear rule to identify the noise component of the data set. [5] proposed a rule that requires some strong distributional assumptions for the estimated Mahalanobis distances to clusters' centers.
- TMIX does not use the eigenratio constraint to prevent degeneracy of the scale matrices. The `EMMIXskew` discards degenerate solutions, which means that solutions on the border of the parameter space cannot be achieved.

`EMMIXskew` allows for setting 5 different covariance models [see 7]. The more general full covariance model (adopted by the OTRIMLE), which is the default choice in `EMMIXskew`, did not always produce a solution in our experiments. That happened particularly for large m, where a full parameterisation would require necessarily some form of regularization. Furthermore, in our experiments we found that selecting the right covariance model was important in terms of RLED_{min}. Several strategies have been tried. The best approach was to pick the covariance specification optimizing the Bayes Information Criterion (BIC), a popular strategy for model selection in model-based clustering. Therefore, for each m, we compute the 5 TMIX clusterings according to the 5 covariance models and we retain the best solution according to the BIC.

It's well known that initialization is a particularly difficult task in clustering. `EMMIXskew` implements several strategies: fully random initials, k-means initialization, etc. In the experiments these methods produced somewhat unstable results. In order to make a fair comparison, both OTRIMLE and TMIX have been initialized using the robust initialization proposed in [1] and implemented in the OTRIMLE package of [2].

4 Survival Curves results

The survival curves for the clusters obtained with the OTRIMLE, SNF and TMIX algorithms are reported. Here we show the figures for all the datasets by using coloured lines. In fact, also the figures for the BREAST

cancer and LUNG cancer present in the main text are shown here.

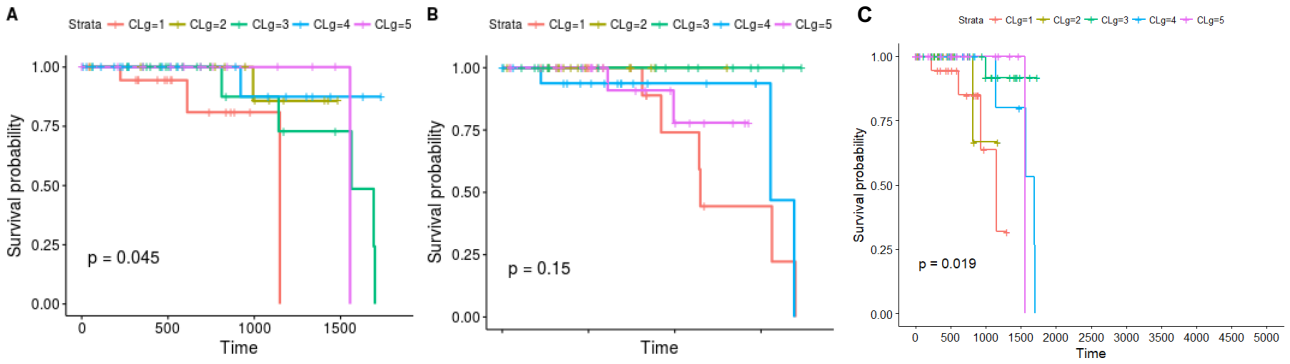


Figure 1: Survival Curves of the BREAST dataset. (A) Survival curves of the clusters obtained with the OTRIMLE algorithm with $m^* = 2$ and $\gamma = 5$. (B) survival curves obtained with the SNF algorithm with $\alpha^* = 0.5$. (C) survival curves obtained with the TMIX algorithm with $m^* = 2$

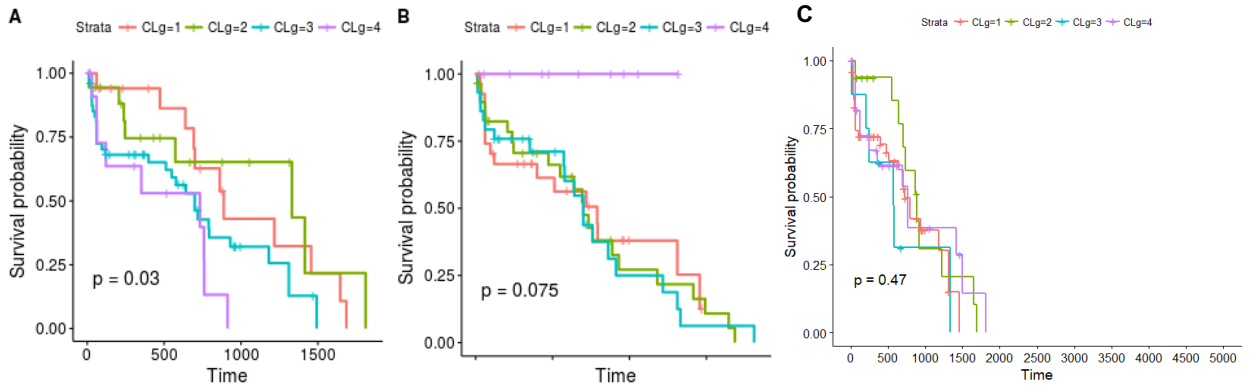


Figure 2: Survival Curves LUNG dataset.(A) survival curves of the clusters obtained with the OTRIMLE algorithm by using $m^* = 11$ and $\gamma^* = 10$. (B) survival curves obtained with the SNF algorithm with $\alpha^* = 0.5$. (C) survival curves obtained with the TMIX algorithm with $m^* = 7$.

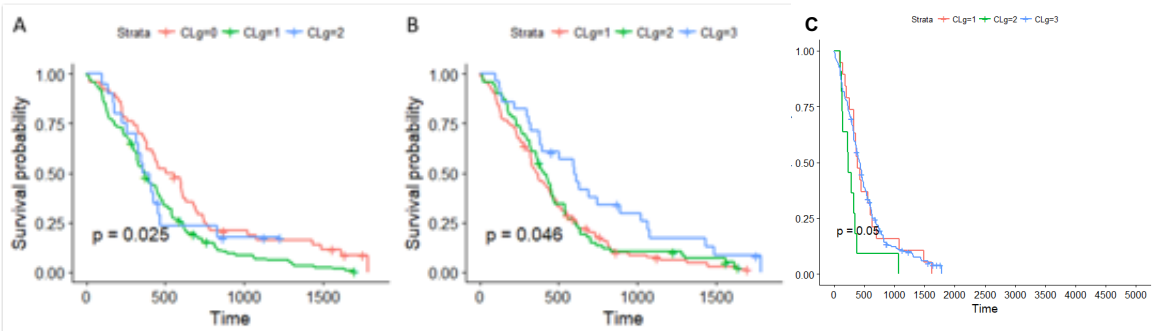


Figure 3: Survival Curves of the GLIO dataset. (A) Survival curves of the clusters obtained with the OTRIMLE algorithm with $m^* = 17$ and $\gamma^* = \text{Inf}$. (B) survival curves obtained with the SNF algorithm with $\alpha^* = 0.4$. (C) survival curves obtained with the TMIX algorithm with $m^* = 21$.

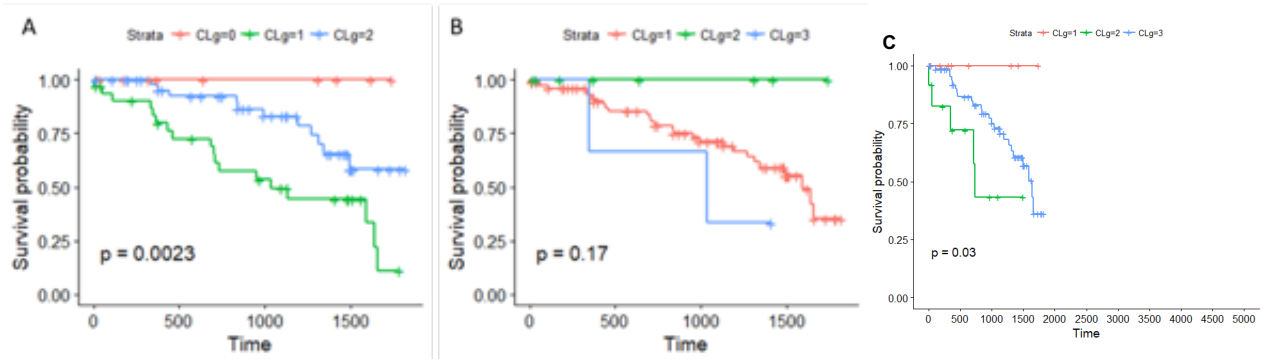


Figure 4: Survival Curves of the KIDNEY dataset. (A) survival curves of the clusters obtained with the OTRIMLE algorithm by using $m^* = 7$ and $\gamma^* = 20$. (B) survival curves obtained with the SNF algorithm with $\alpha^* = 0.4$. (C) survival curves obtained with the TMIX algorithm with $m^* = 11$.

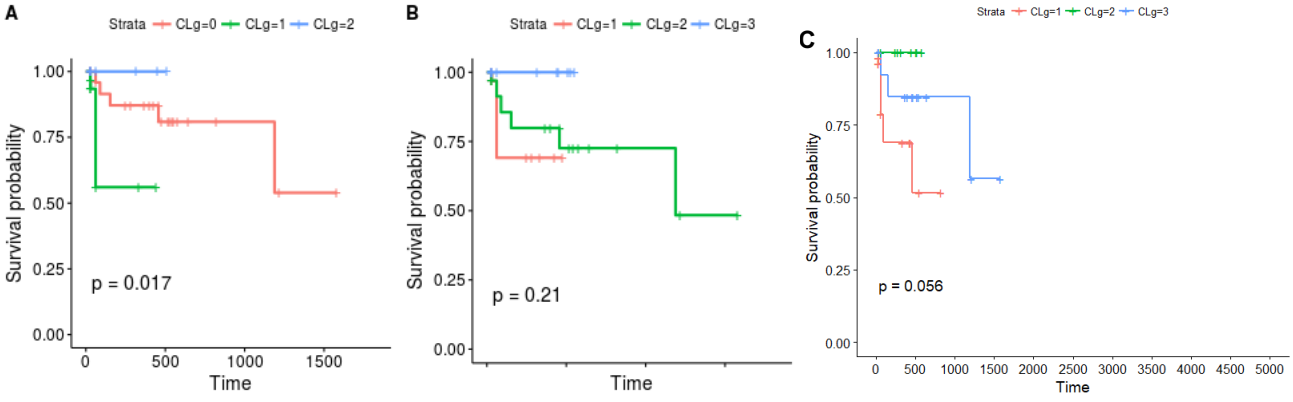


Figure 5: Survival Curves of the COLON dataset. (A) survival curves of the clusters obtained with the OTRIMLE algorithm by using $m^* = 28$ and $\gamma^* = 3$. (B) survival curves obtained with the SNF algorithm with $\alpha^* = 0.7$. (C) survival curves obtained with the TMIX algorithm with $m^* = 3$.

5 Over-represented KEGG pathways

Here we report the results of the KEGG pathways over-representation analysis performed on the clusters obtained with the OTRIMLE, SNF and TMIX algorithms. For each cluster the most relevant pathways over-represented by the lists of down-regulated (left column) and up-regulated (right column) genes are reported. The darker are the points in the figure the higher is their relevance, in terms of p-values, and of the association of the pathways to the up/down regulated genes.

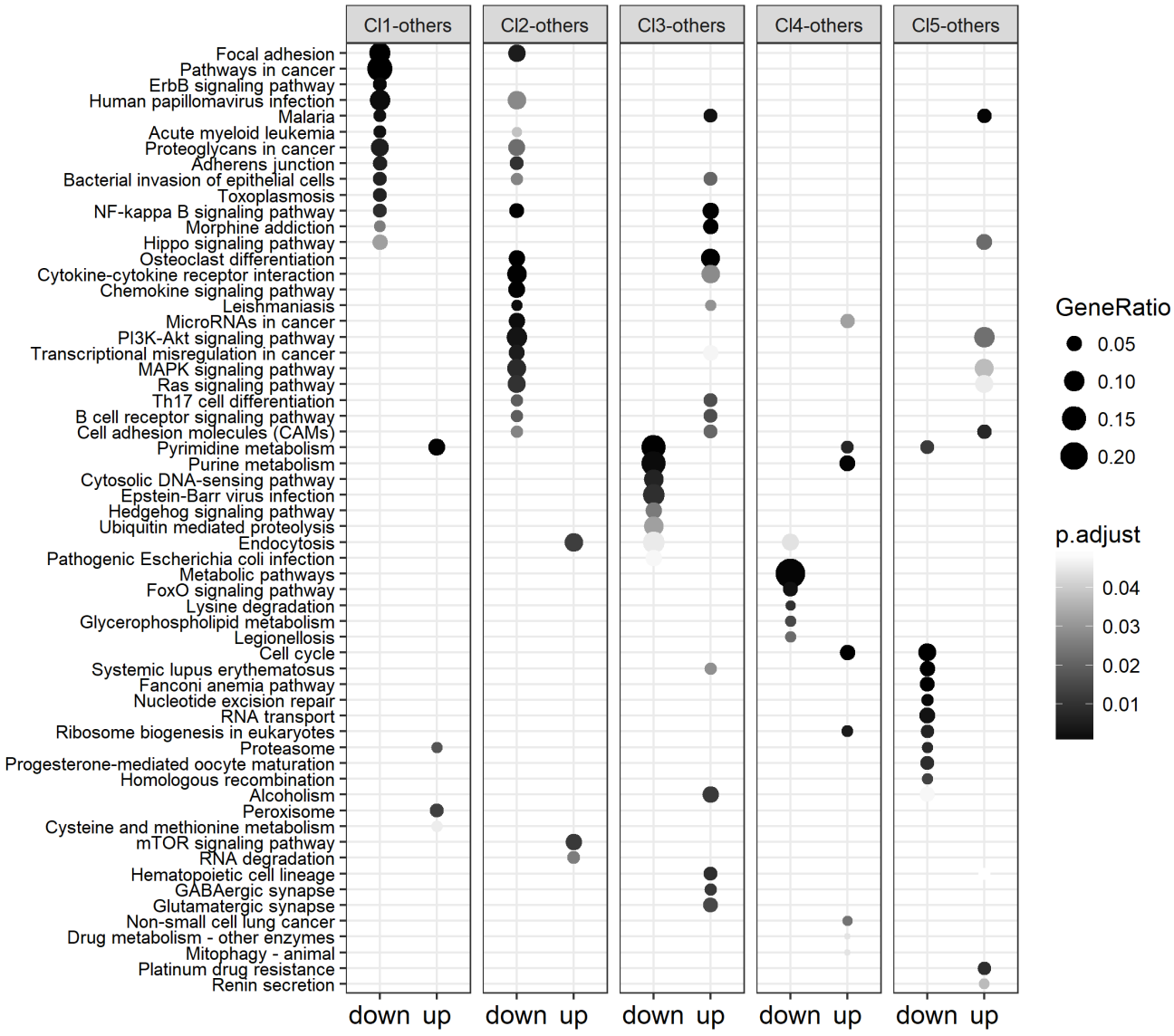


Figure 6: Results of the KEGG pathways over-representation analysis on the SNF clustering of the breast cancer patients.

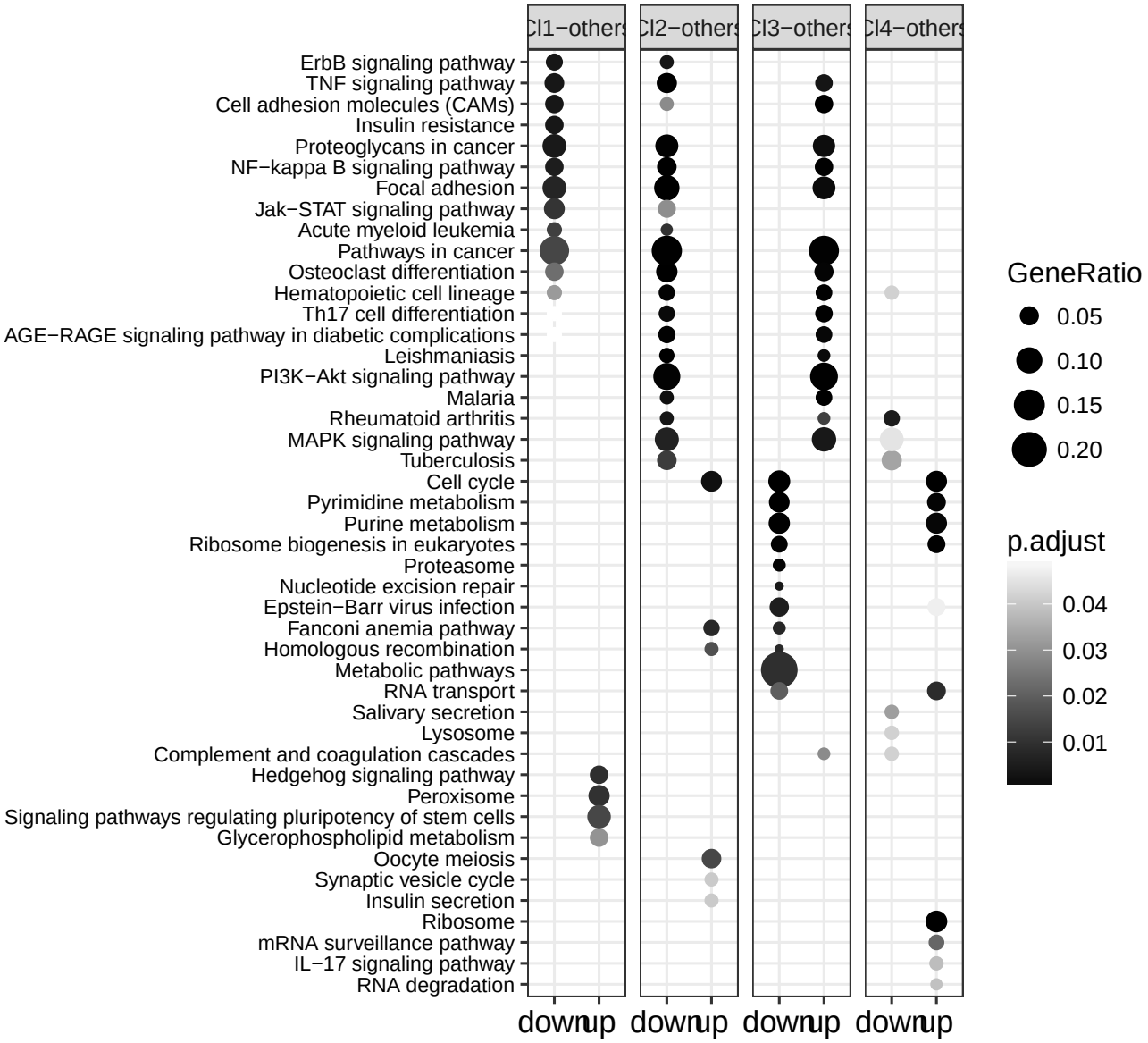


Figure 7: Results of the KEGG pathways over-representation analysis on the TMIX clustering of the breast cancer patients.

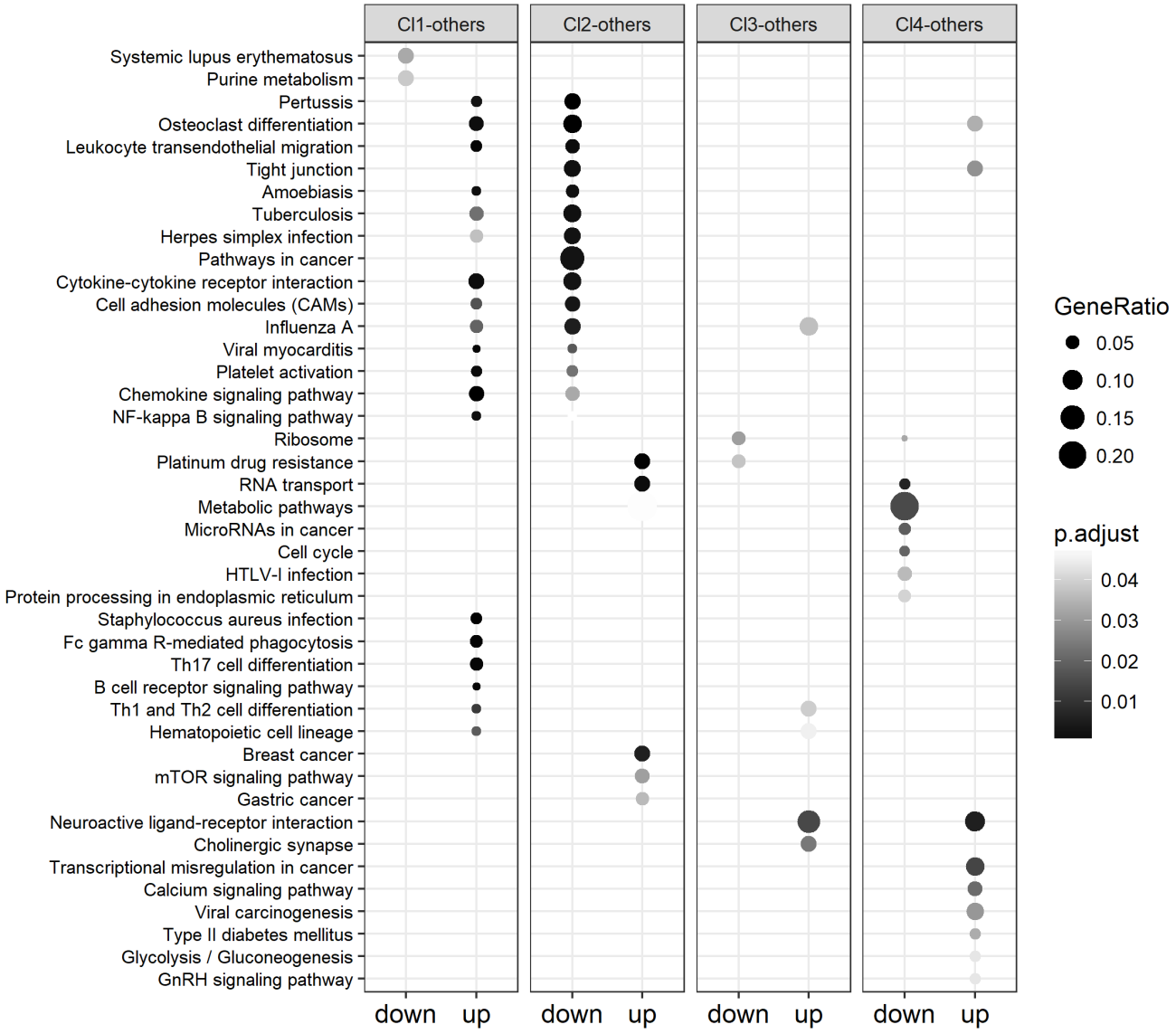


Figure 8: Results of the KEGG pathways over-representation analysis on the SNF clustering of the lung cancer patients.

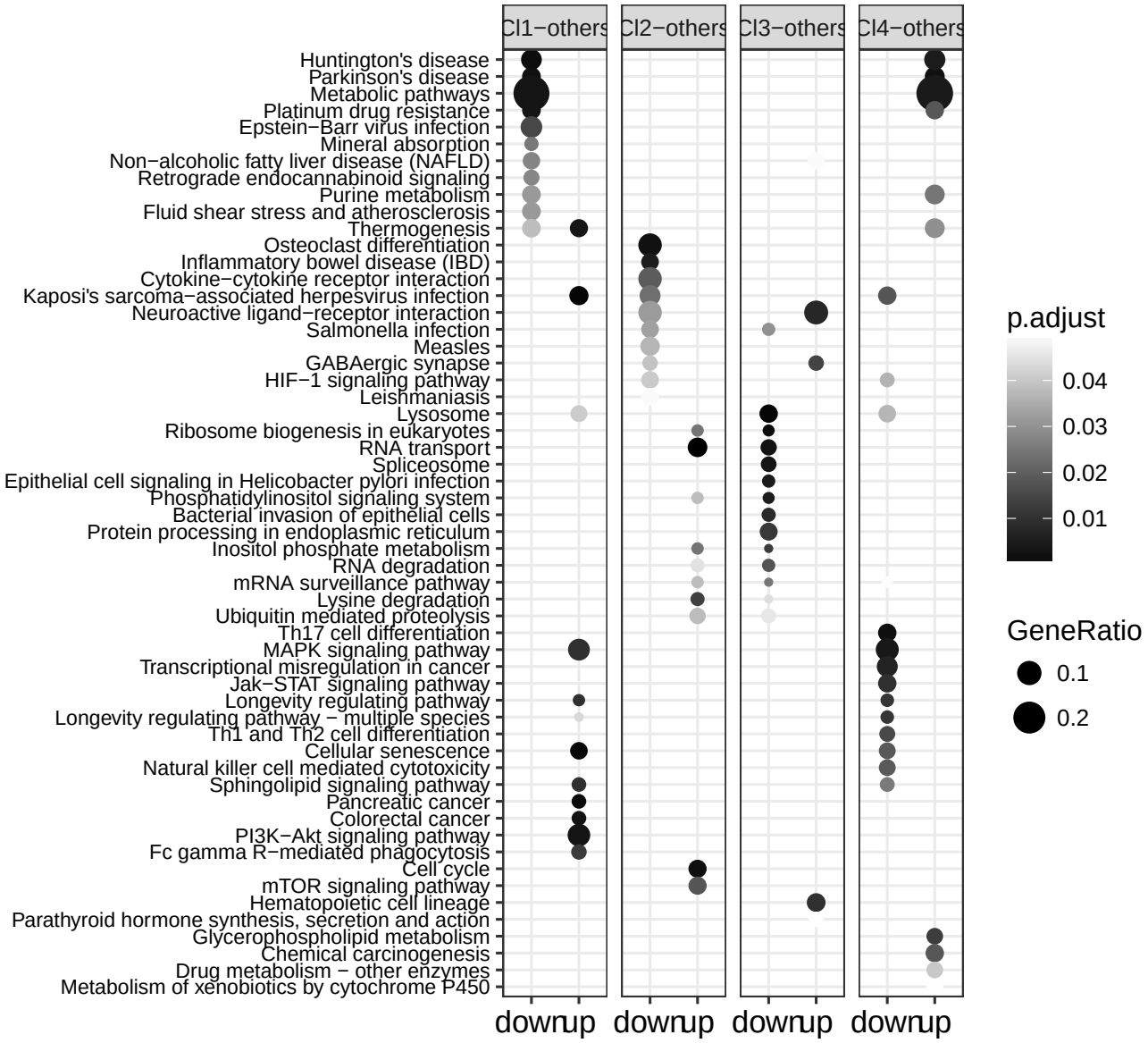


Figure 9: Results of the KEGG pathways over-representation analysis on the TMIX clustering of the lung cancer patients.

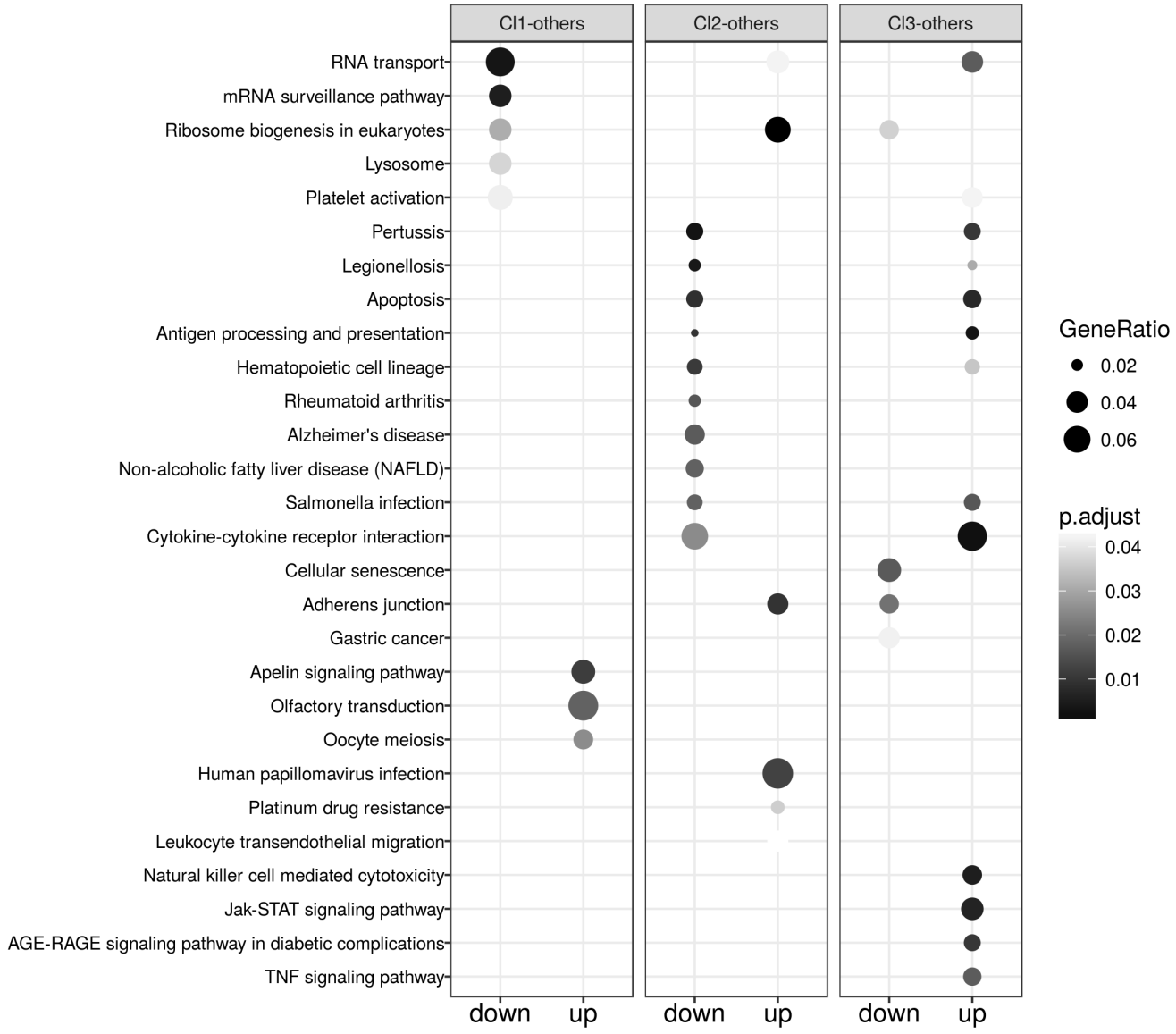


Figure 10: Results of the KEGG pathways over-representation analysis on the SNF clustering of the colon cancer patients.

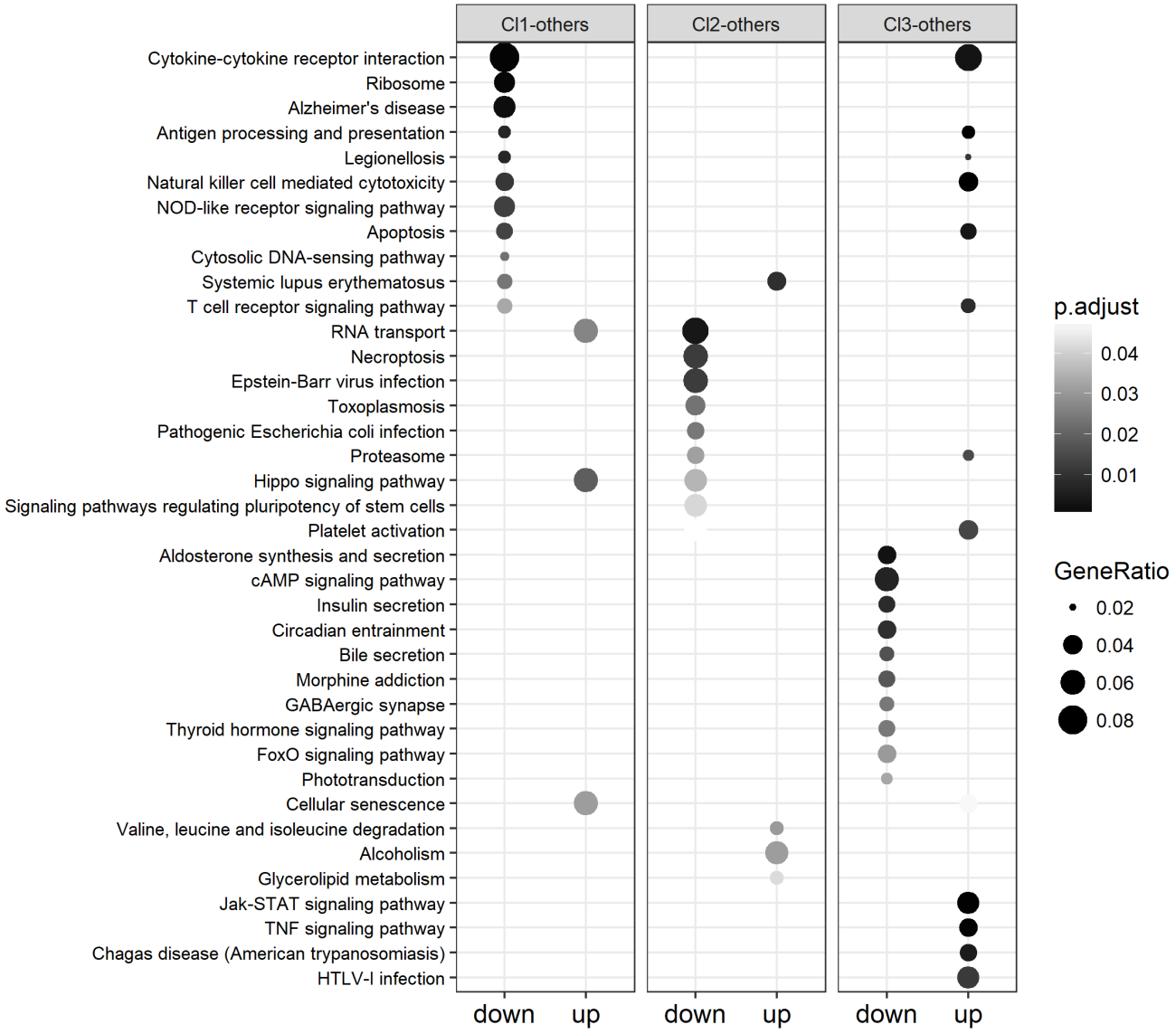


Figure 11: Results of the KEGG pathways over-representation analysis on the OTRIMLE clustering of the colon cancer patients.

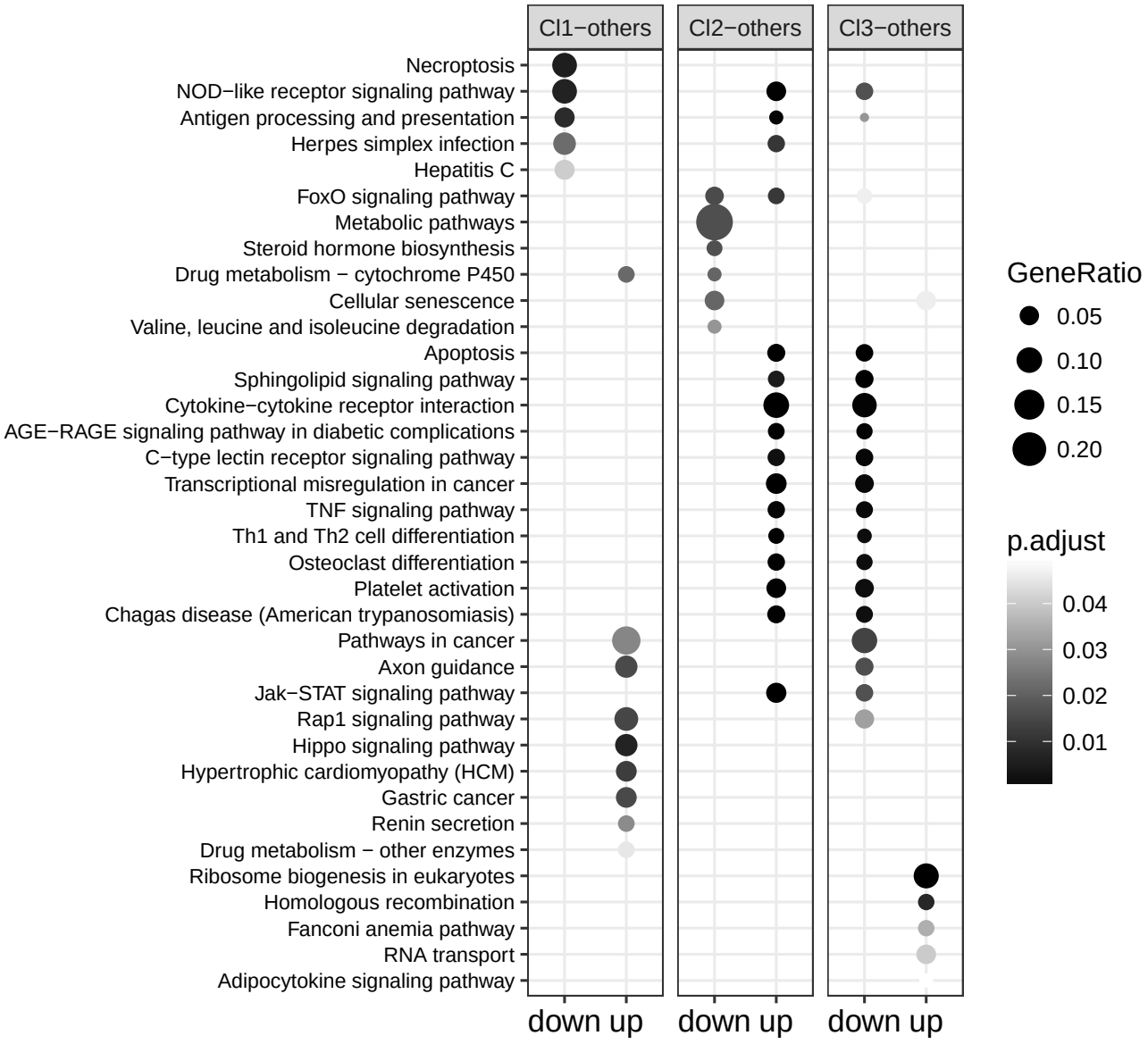


Figure 12: Results of the KEGG pathways over-representation analysis on the TMIX clustering of the colon cancer patients.

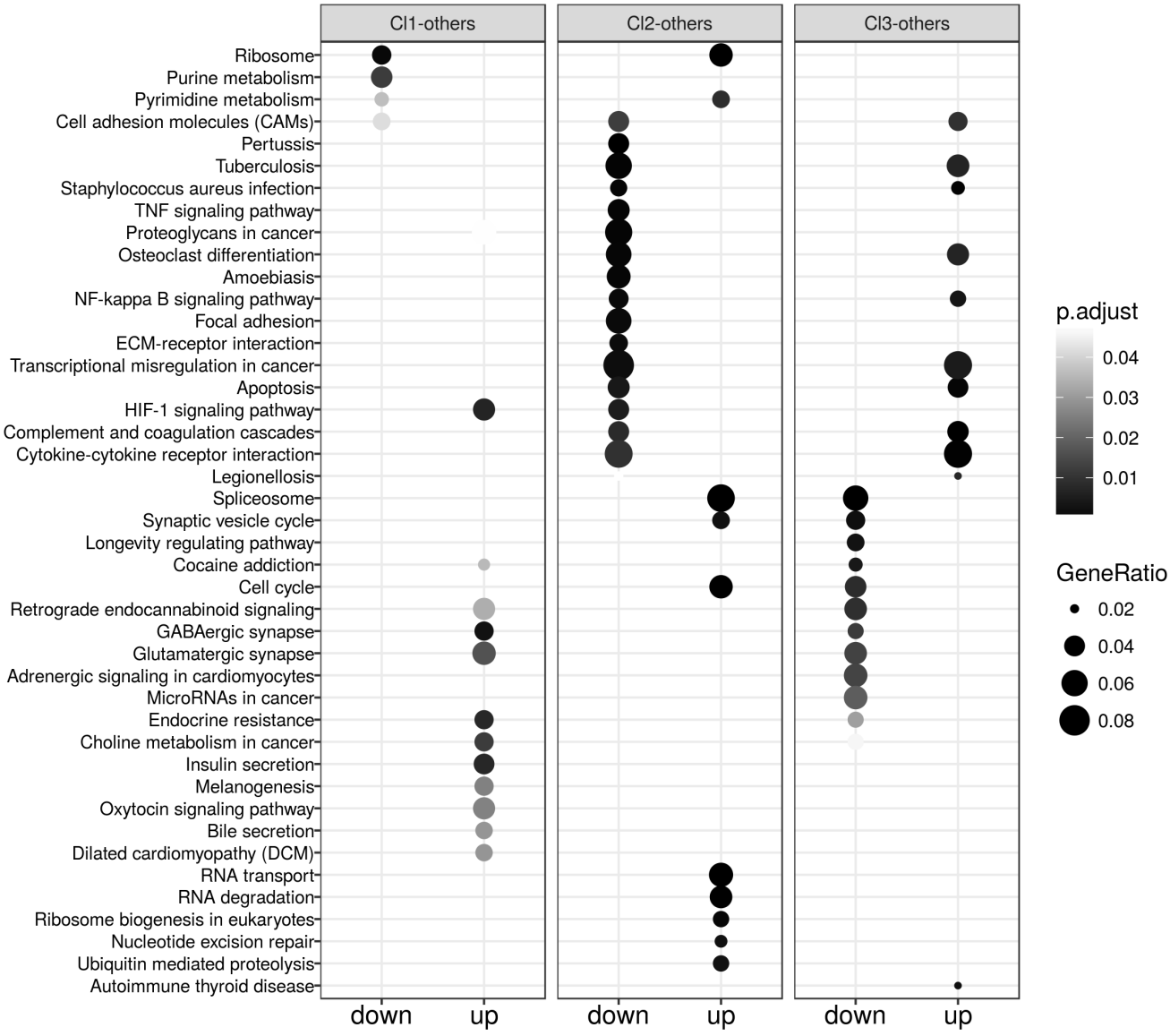


Figure 13: Results of the KEGG pathways over-representation analysis on the SNF clustering of the glioblastoma patients.

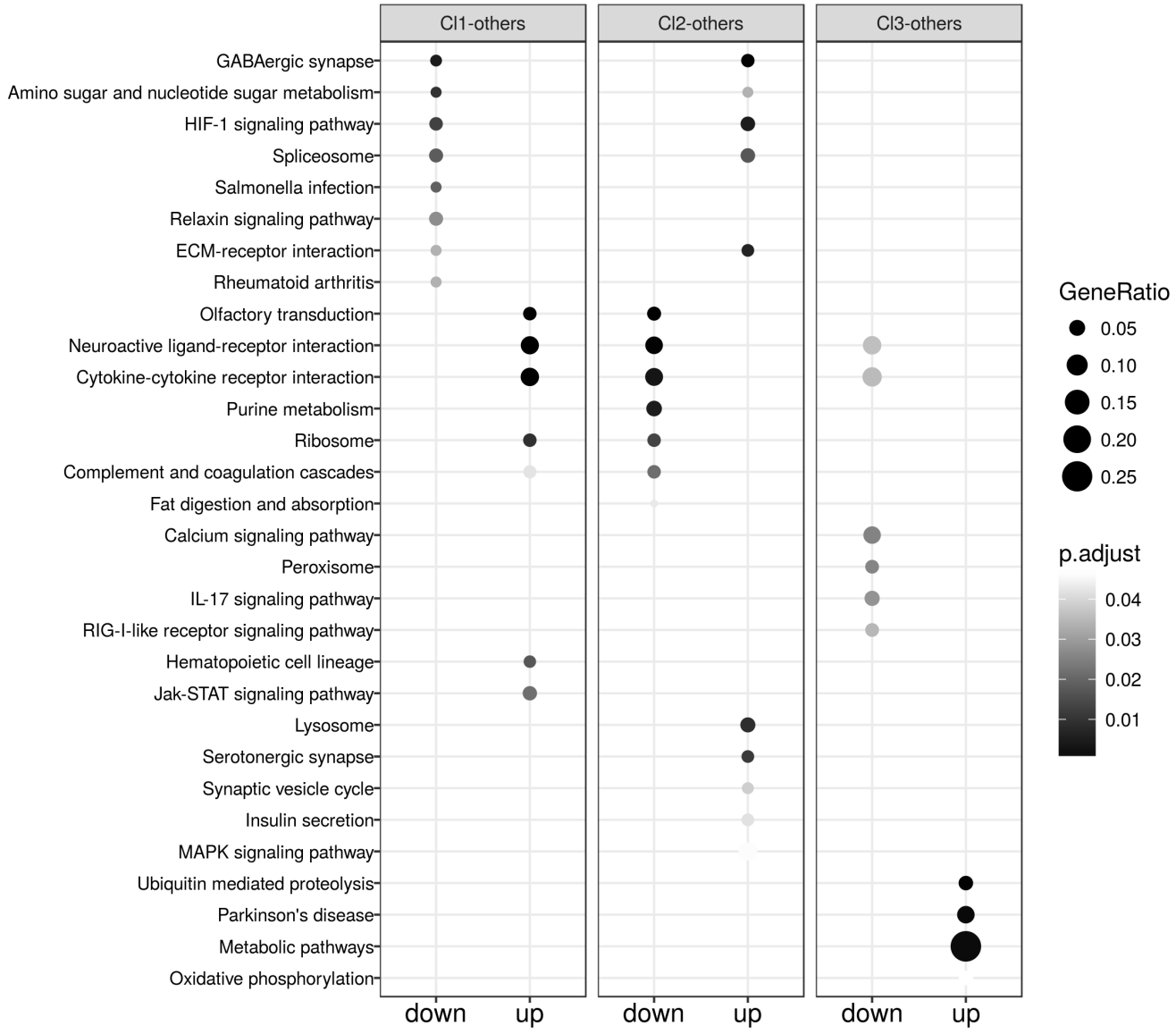


Figure 14: Results of the KEGG pathways over-representation analysis on the OTRIMLE clustering of the glioblastoma patients.

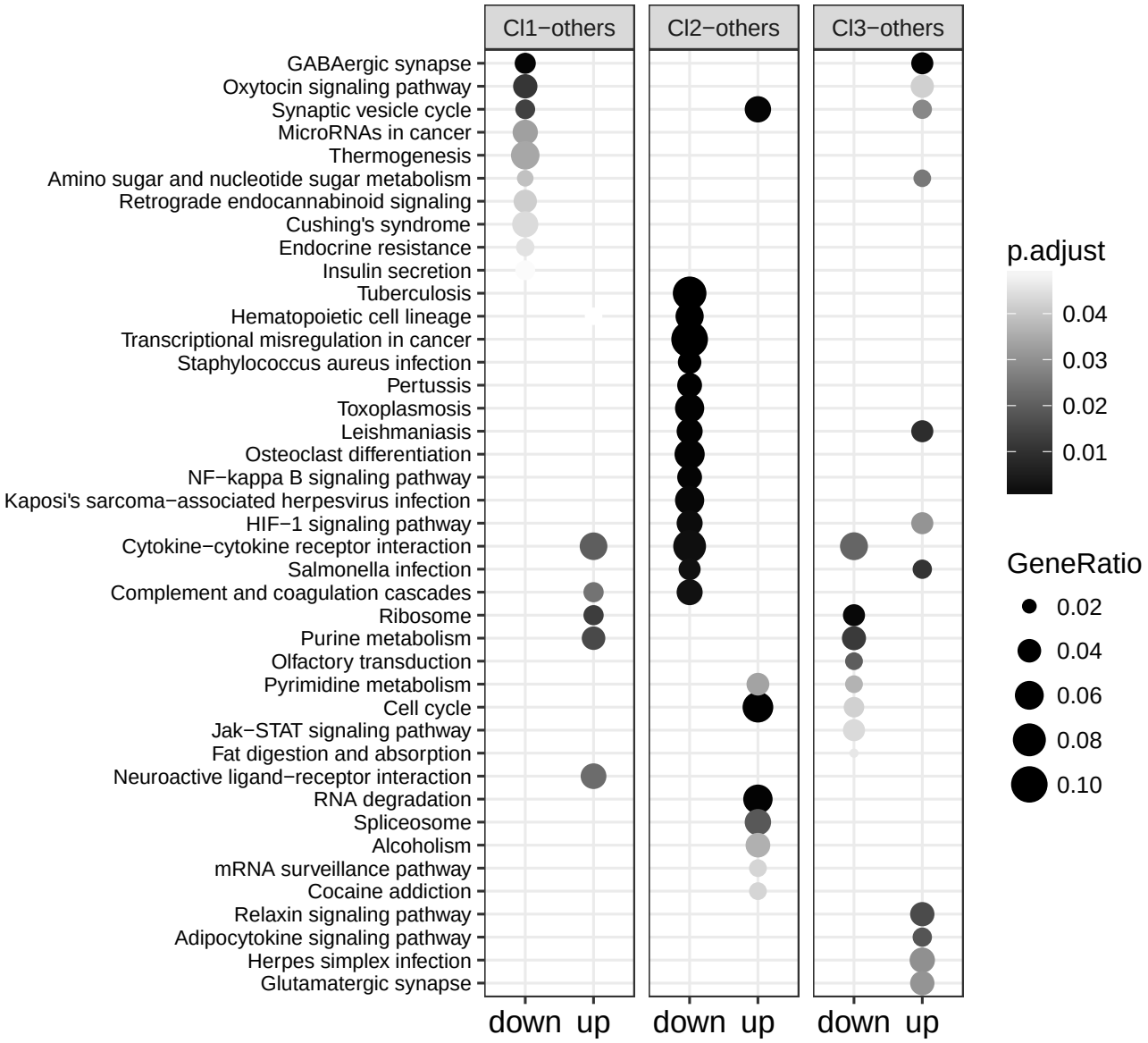


Figure 15: Results of the KEGG pathways over-representation analysis on the TMIX clustering of the glioblastoma patients.

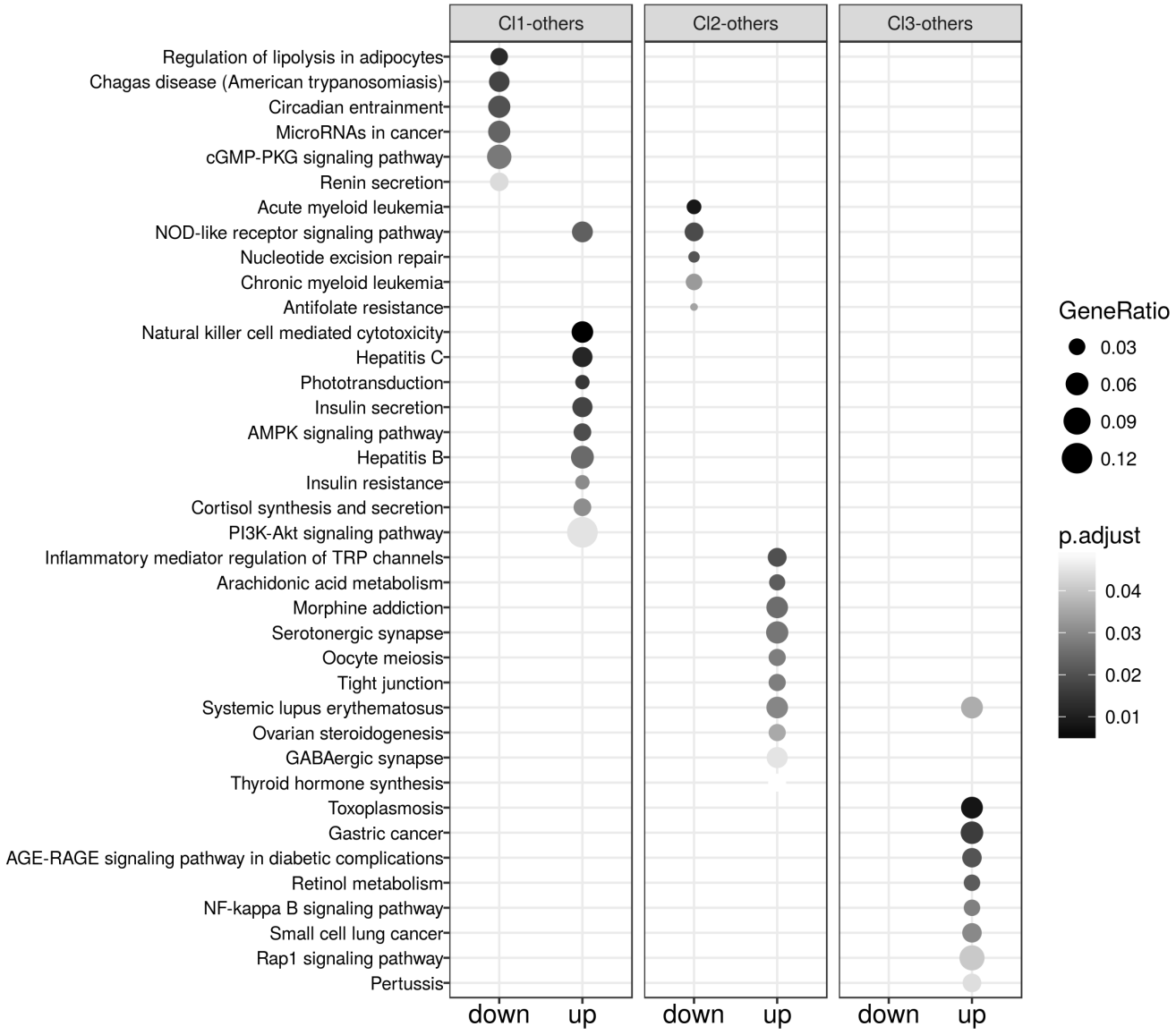


Figure 16: Results of the KEGG pathways over-representation analysis on the SNF clustering of the kidney cancer patients.

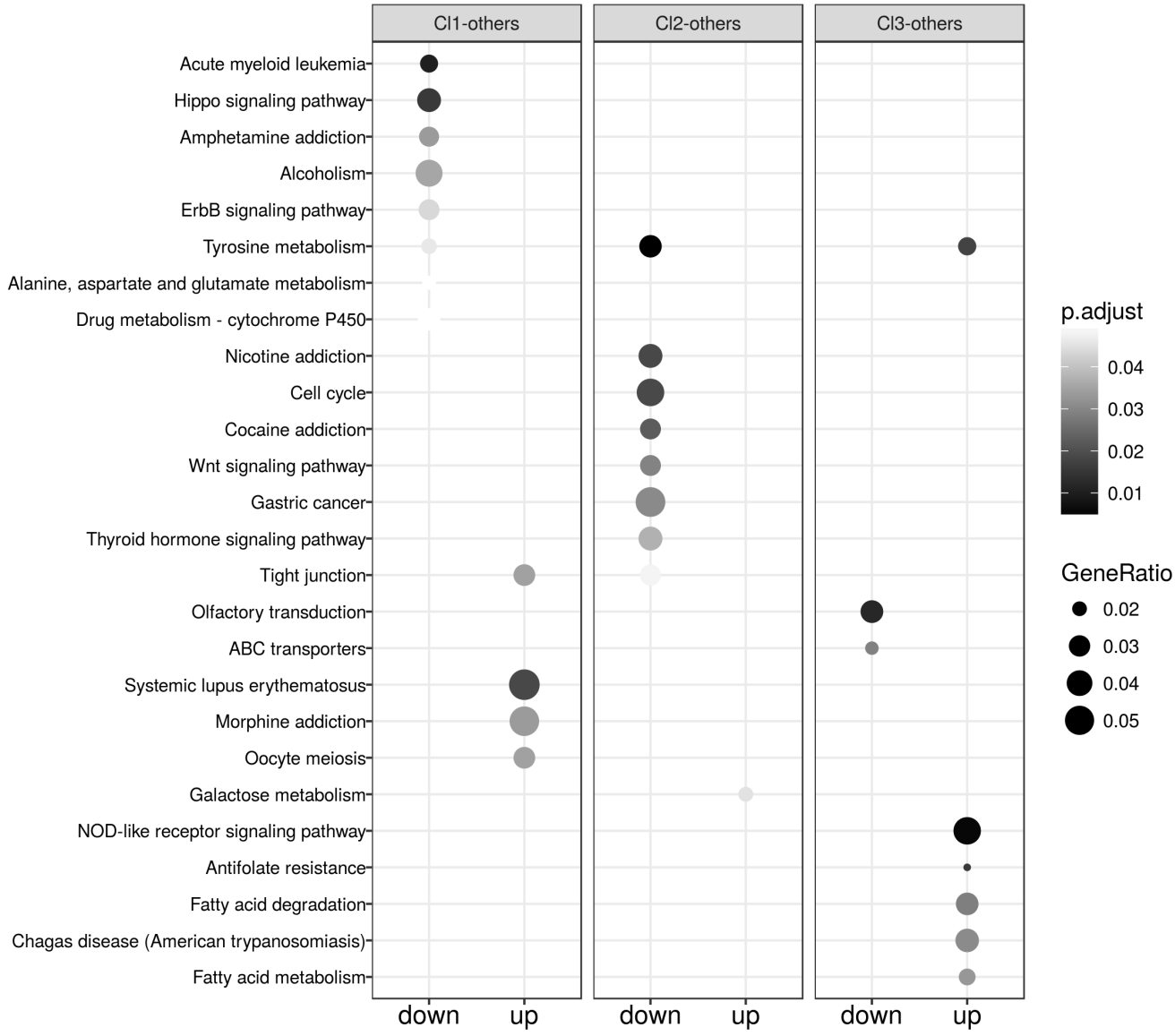


Figure 17: Results of the KEGG pathways over-representation analysis on the OTRIMLE clustering of the kidney cancer patients.

6 Selection of the m parameter

Here we report the optimal m^* decided by our algorithm compared to the distribution of the ordered eigenvalues of the RSC matrix. For scaling reasons the following plots report the largest 100 components only.

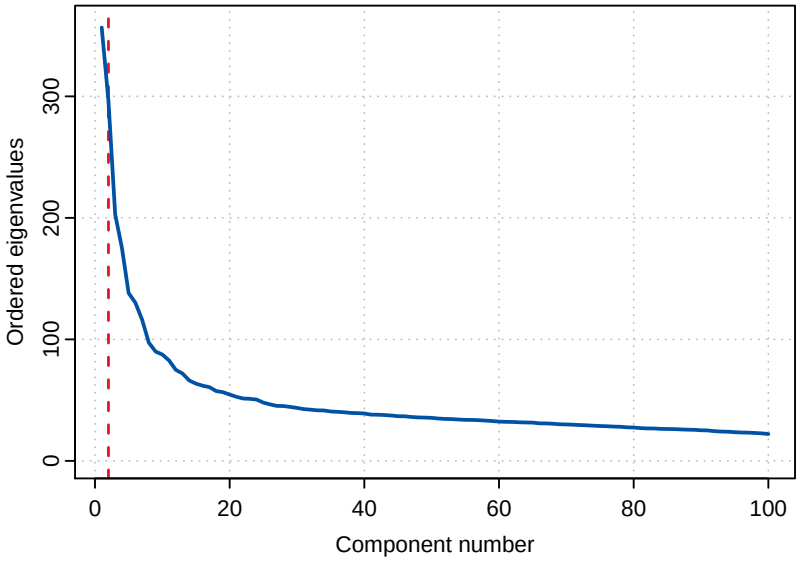


Figure 18: Ordered eigenvalues of the RSC matrix for the BREAST data set. The vertical red line corresponds to the optimal m^* decided by the algorithm.

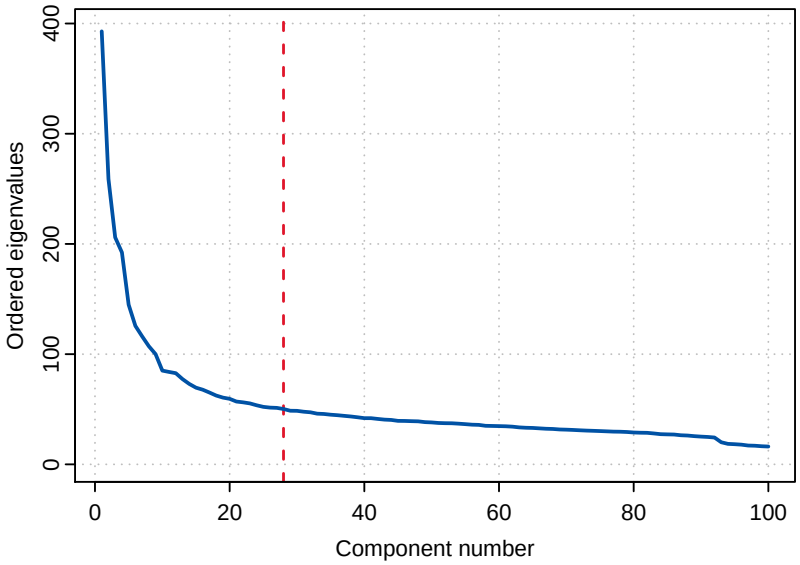


Figure 19: Ordered eigenvalues of the RSC matrix for the COLON data set. The vertical red line corresponds to the optimal m^* decided by the algorithm.

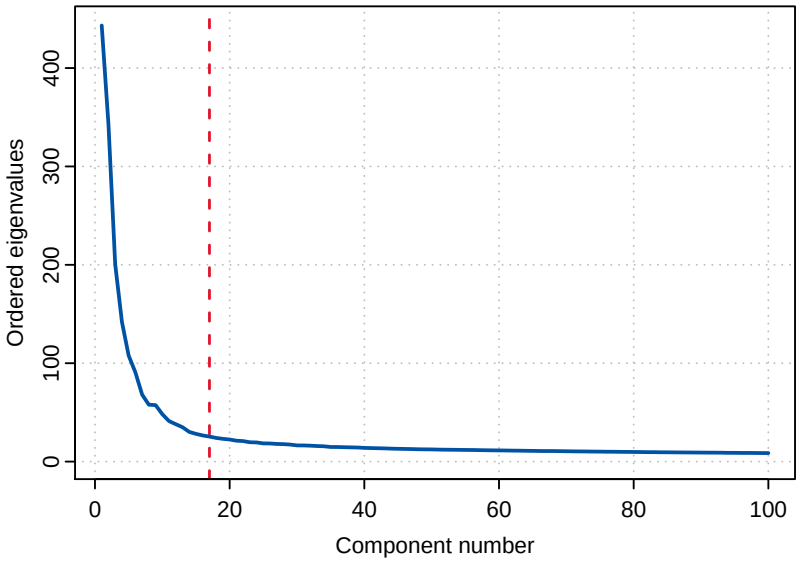


Figure 20: Ordered eigenvalues of the RSC matrix for the GLIO data set. The vertical red line corresponds to the optimal m^* decided by the algorithm.

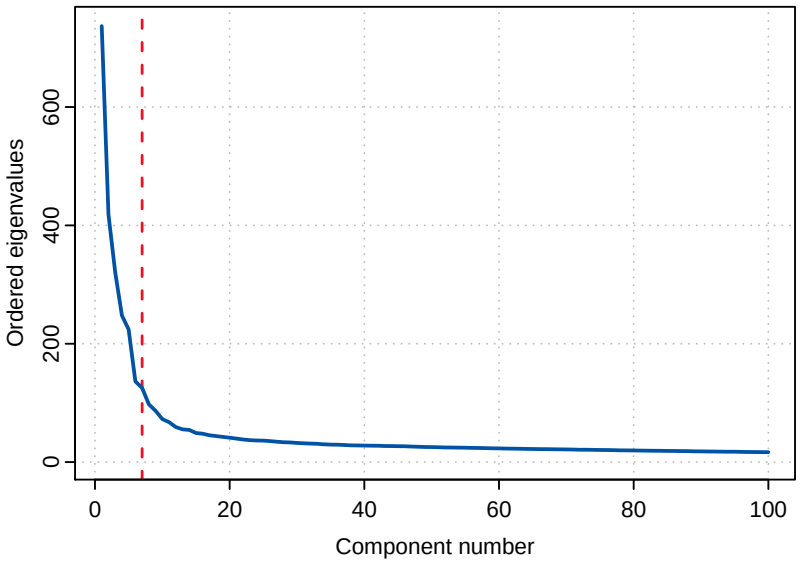


Figure 21: Ordered eigenvalues of the RSC matrix for the KIDNEY data set. The vertical red line corresponds to the optimal m^* decided by the algorithm.

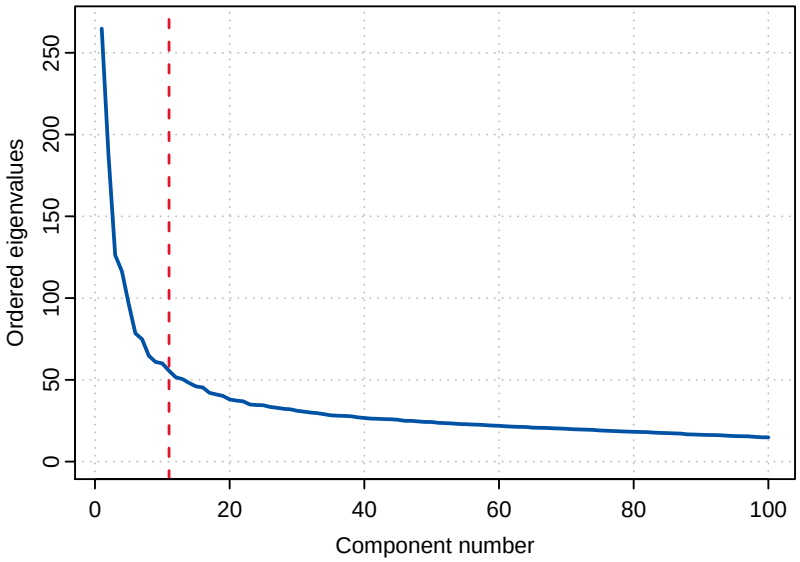


Figure 22: Ordered eigenvalues of the RSC matrix for the LUNG data set. The vertical red line corresponds to the optimal m^* decided by the algorithm.

7 Sensitivity of RLED_{min} value to the m and γ parameters for OTRIMLE method

The experiments with the OTRIMLE algorithm were executed by performing a grid search on the m and γ parameters. In particular, we consider m from 2 to 30 (by steps of 1), from 33 to 48 (by step of 3) and from 50 to 100 (by steps of 5). The grid for the γ is $\{1, 2, 3, 5, 10, 20, 50, 100, 500, 1000, 10000, +\infty\}$. In this section the behaviour of the $RLED_{min}$ measure for each combination of m and γ is reported in the following bubble plots. The figures report the $RLED_{min}$ values for both the solutions with and without noise. For each cluster, the p-value of the log rank text was also computed. Red dots identify solutions with no significant p-values, while red ones identify clusterings with significant p-values. Blu points identify solutions that give raise to small clusters (clusters with less than 21% of the patients in the dataset). Combinations that are not marked in the bubble plots are those obtaining a number of clusters different from the desired one. Both the blue and the empty solutions were discarded from the results.

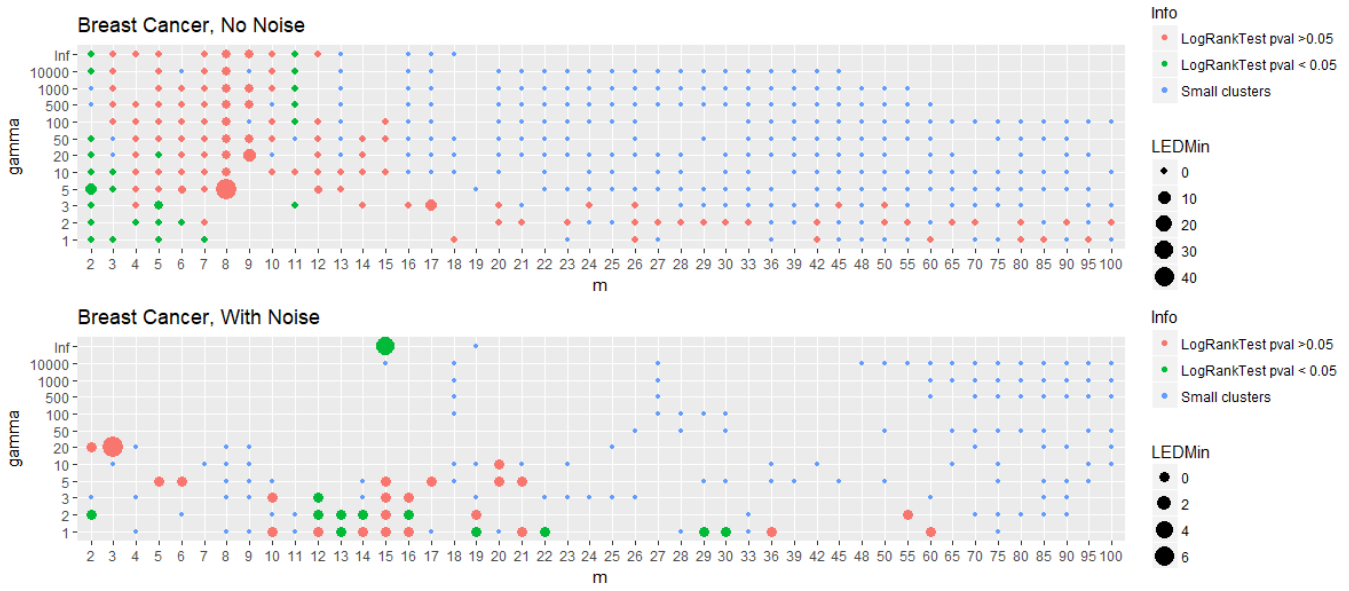


Figure 23: Distribution of the $RLED_{min}$ value across the grid of γ and m values of the cluster identified with the OTRIMLE algorithm in the BREAST data set.

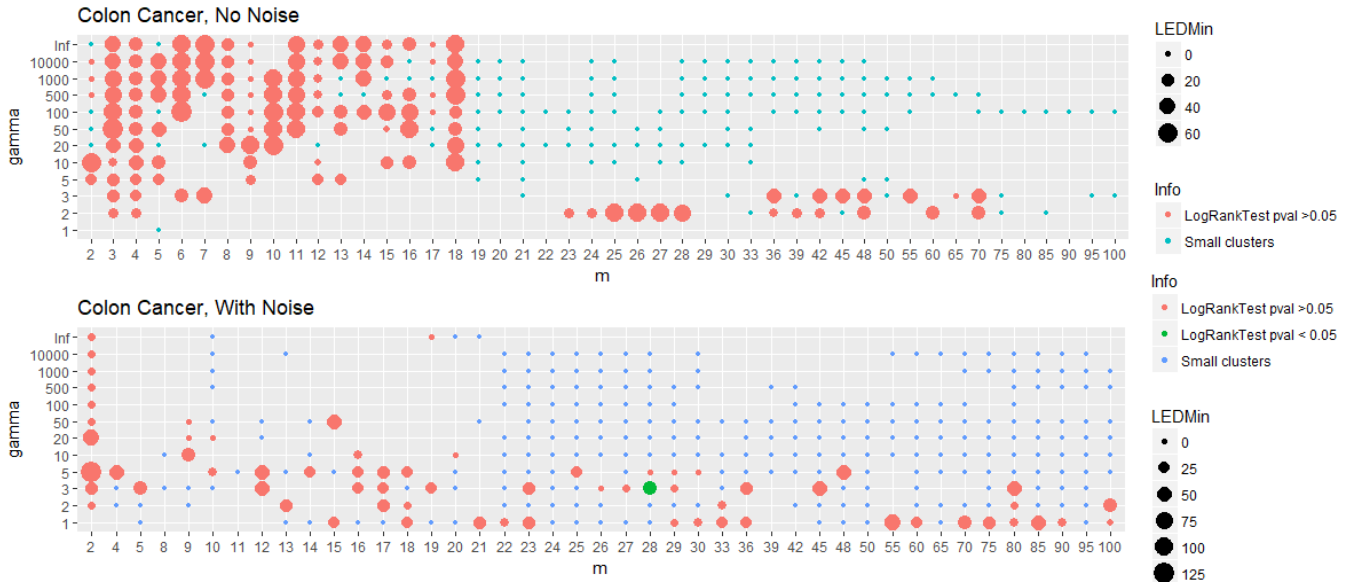


Figure 24: Distribution of the $RLED_{min}$ value across the grid of γ and m values of the cluster identified with the OTRIMLE algorithm in the COLON data set.

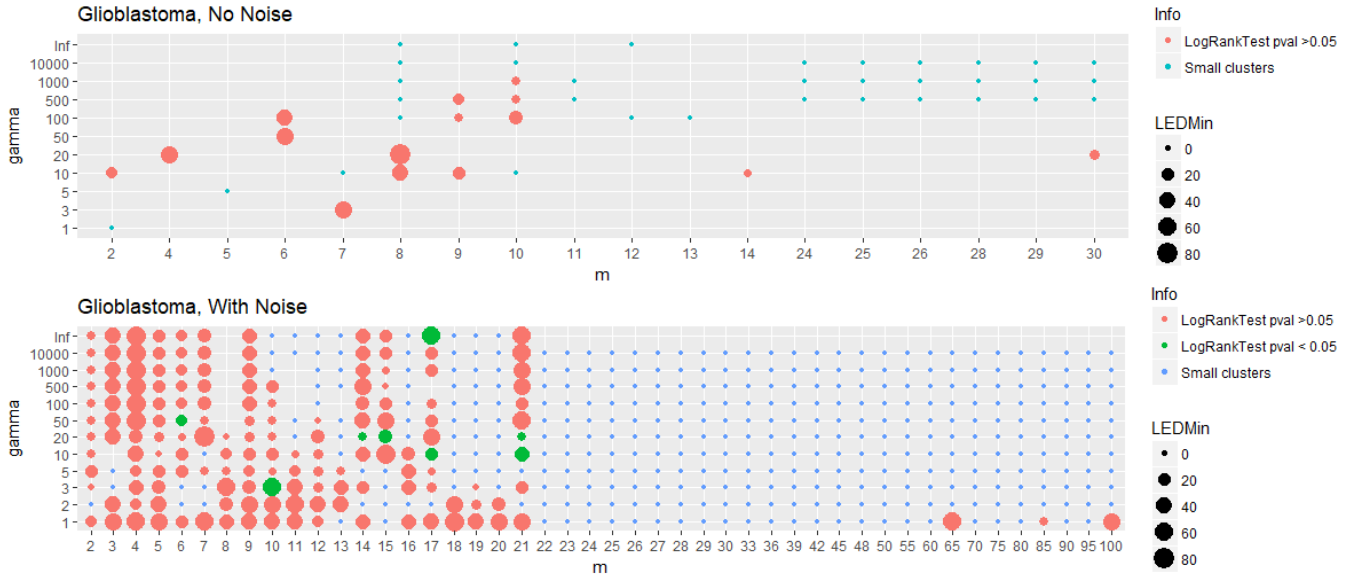


Figure 25: Distribution of the $RLED_{min}$ value across the grid of γ and m values of the cluster identified with the OTRIMLE algorithm in the GLIO data set.

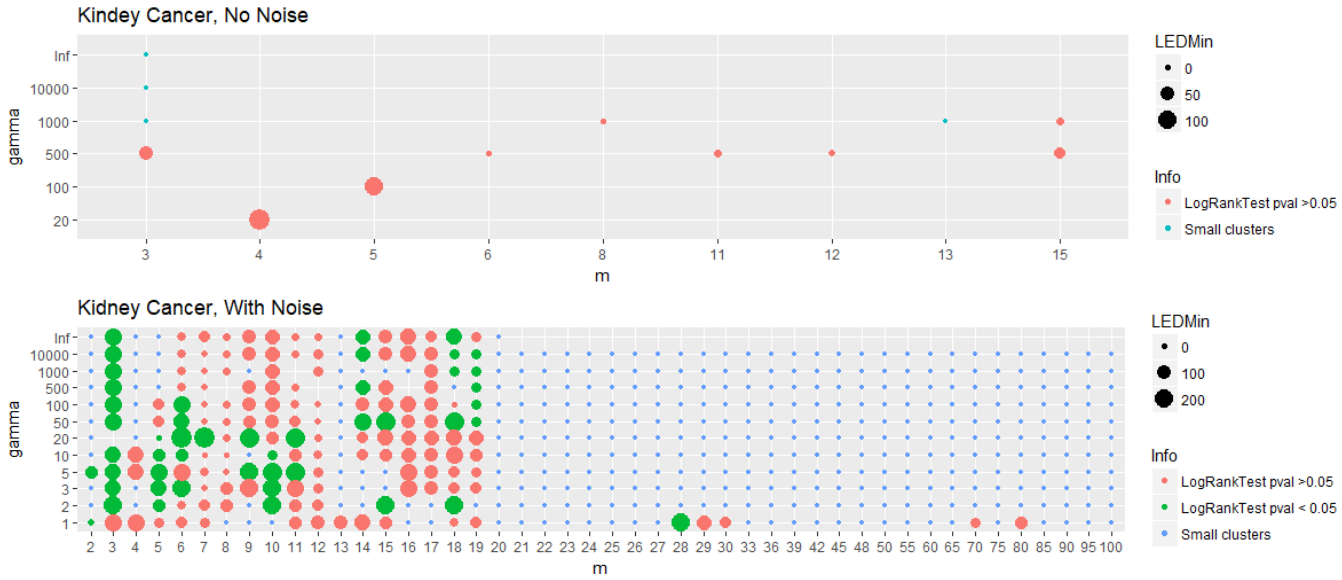


Figure 26: Distribution of the $RLED_{min}$ value across the grid of γ and m values of the cluster identified with the OTRIMLE algorithm in the KIDNEY data set.

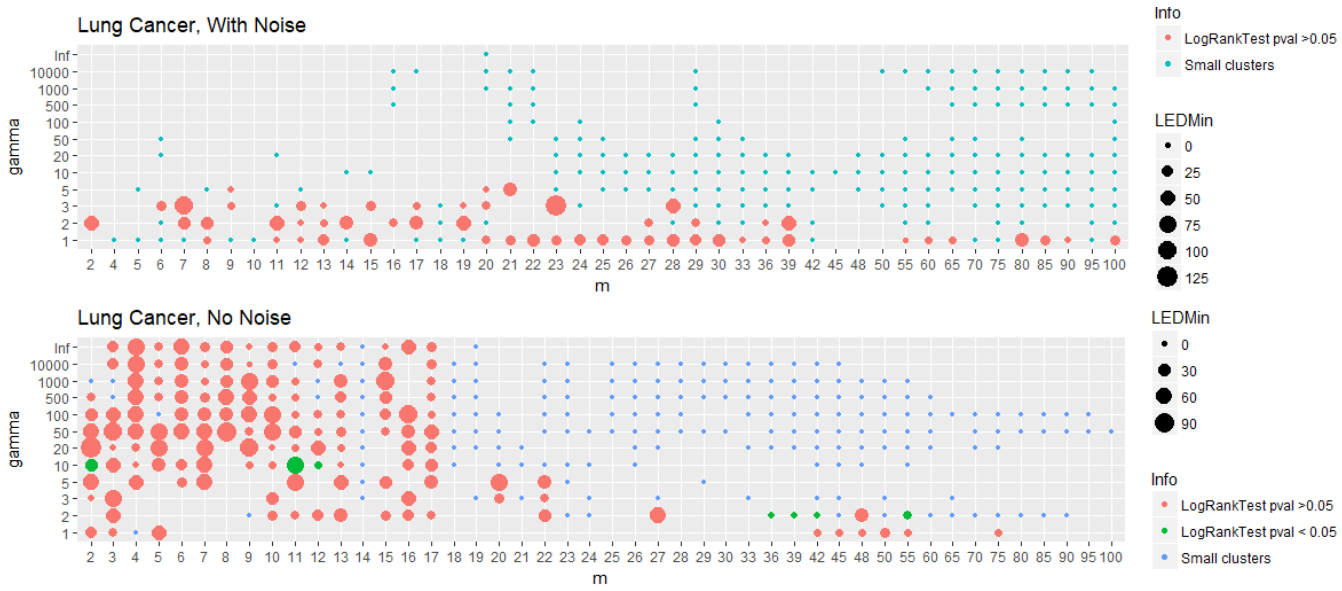


Figure 27: Distribution of the $RLED_{min}$ value across the grid of γ and m values of the cluster identified with the OTRIMLE algorithm in the LUNG data set.

8 Sensitivity of RLED_{min} value to the α parameters for SNF method

The SNF algorithm was executed by different values of α in the interval going from 0.3 to 1 by step of 0.1. Here we report the $RLED_{min}$ value distribution for all the clustering results in all the datasets. For each cluster the p-value of the log rank text was also computed. Red dots identify solutions with no significant p-values, while red ones identify clustering with significant p-value. Blu points identifies solutions that give raise to small clusters (clusters with less that 21% of the patients in the dataset).

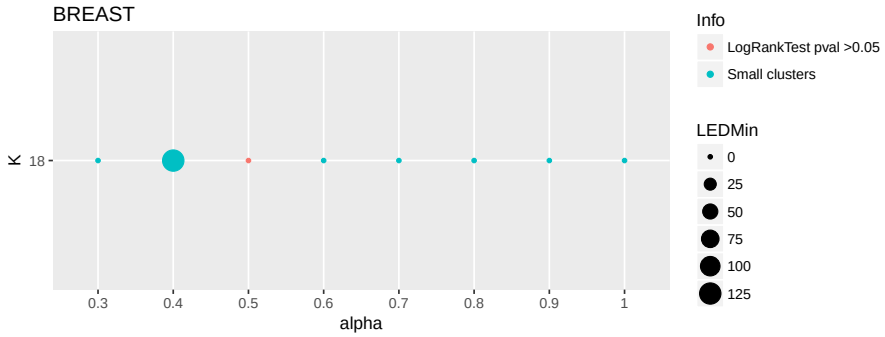


Figure 28: Distribution of the $RLED_{min}$ value across the grid of α values of the cluster identified with the SNF algorithm in the BREAST data set.

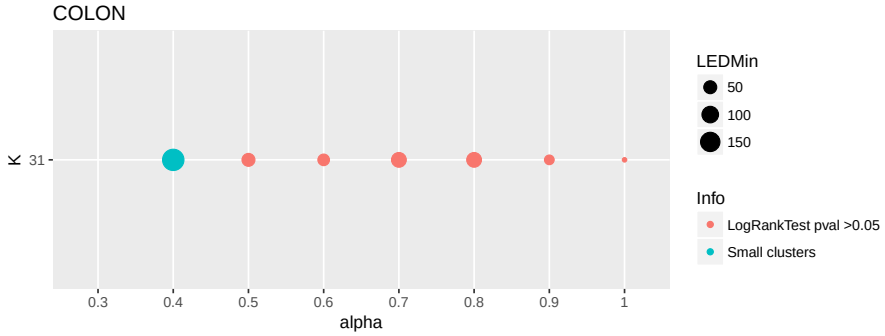


Figure 29: Distribution of the $RLED_{min}$ value across the grid of α values of the cluster identified with the SNF algorithm in the COLON data set.

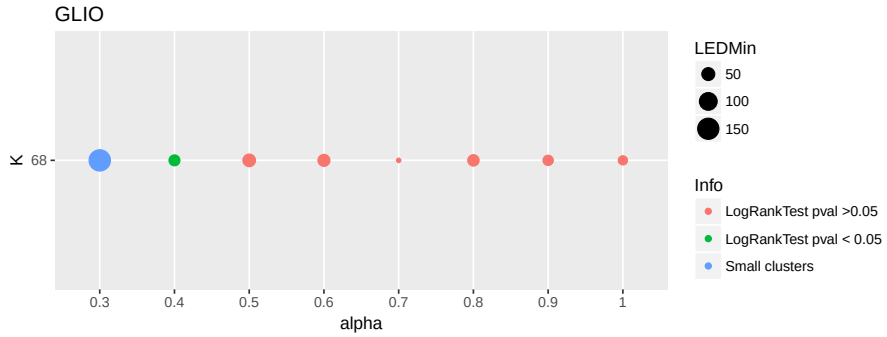


Figure 30: Distribution of the $RLED_{min}$ value across the grid of α values of the cluster identified with the SNF algorithm in the GLIO data set.

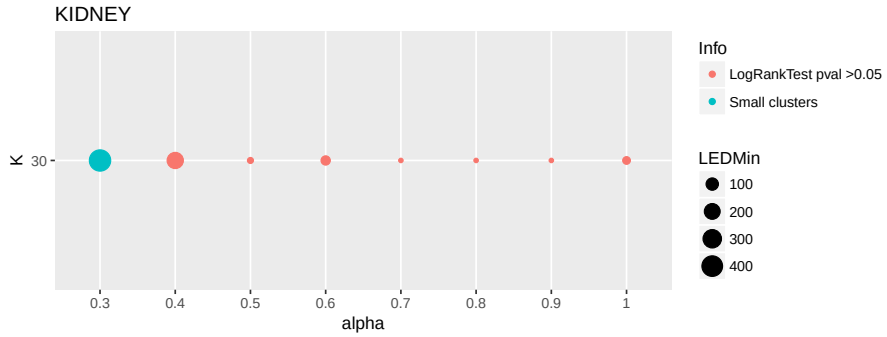


Figure 31: Distribution of the $RLED_{min}$ value across the grid of α values of the cluster identified with the SNF algorithm in the KIDNEY data set.

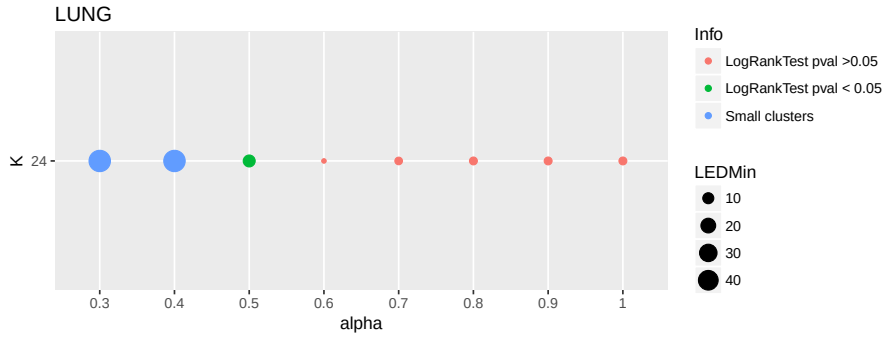


Figure 32: Distribution of the $RLED_{min}$ value across the grid of α values of the cluster identified with the SNF algorithm in the LUNG data set.

9 Sensitivity of RLED_{min} value to the m parameters for TMIX method

The TMIX algorithm was executed by different values of m from 2 to 30 (by steps of 1), from 33 to 48 (by step of 3) and from 50 to 100 (by steps of 5). Here we report the $RLED_{min}$ value distribution for all the clustering results in all the datasets. For each cluster the p-value of the log rank text was also computed. Red dots identify solutions with no significant p-values, while red ones identify clustering with significant p-value. Only the m values for which a solution was obtained are reported.

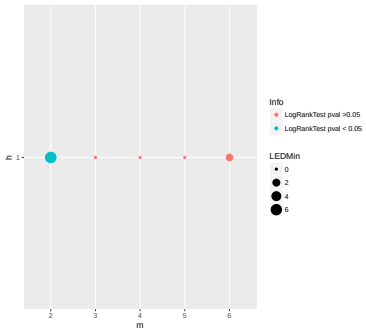


Figure 33: Distribution of the $RLED_{min}$ value across the grid of α values of the cluster identified with the TMIX algorithm in the BREAST data set.

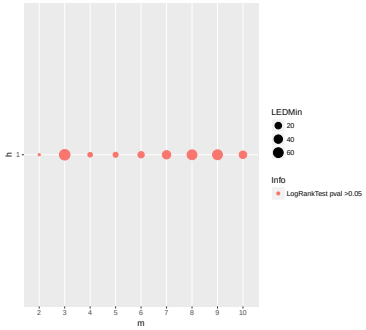


Figure 34: Distribution of the $RLED_{min}$ value across the grid of m values of the cluster identified with the TMIX algorithm in the COLON data set.

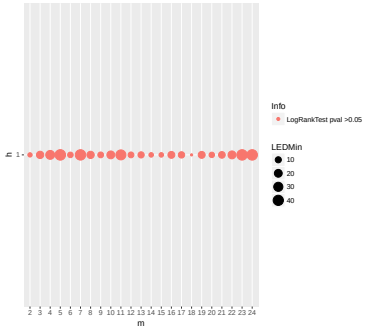


Figure 35: Distribution of the $RLED_{min}$ value across the grid of m values of the cluster identified with the TMIX algorithm in the GLIO data set.

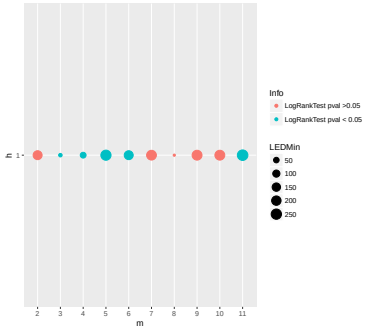


Figure 36: Distribution of the $RLED_{min}$ value across the grid of m values of the cluster identified with the TMIX algorithm in the KIDNEY data set.

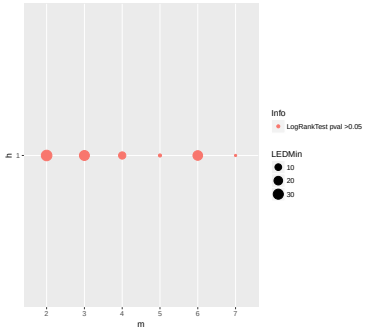


Figure 37: Distribution of the $RLED_{min}$ value across the grid of m values of the cluster identified with the TMIX algorithm in the LUNG data set.

10 Execution time

In this section the execution time of the SNF, TMIX and of the proposed methods are reported for each data set. The tests were performed on a windows system installed on a Intel(R) Core(TM) i7-4790K CPU(4GHz) machine. In Table 2 the timing is reported for a single α value for the SNF algorithm. Table 3 reports the OTRIMLE timing for $m = 100$ and $\gamma = 1$, which is the most demanding pair of inputs for the algorithm. Table 4reports the OTRIMLE timing for $m = 100$, which is the most demanding pair of inputs for the algorithm.In our study we considered 552 of these pairs, although, as highlighted in Section 3 of the paper, the number of m values can be greatly reduced by concentrating the search around the elbow region of the distribution of the eigenvalues of the RSC matrix (see Section 6 in this Supplement).

Table 2: SNF execution time (in seconds). The first column refers to the time needed to build the integrate affinity matrix (W). The second column refers to the time needed to compute spectral clustering on W.

	W	Clustering
KIDNEY	0.13	0.01
GLIO	0.72	0.02
BREAST	0.87	0.02
LUNG	0.16	0.01
COLON	0.10	0.01

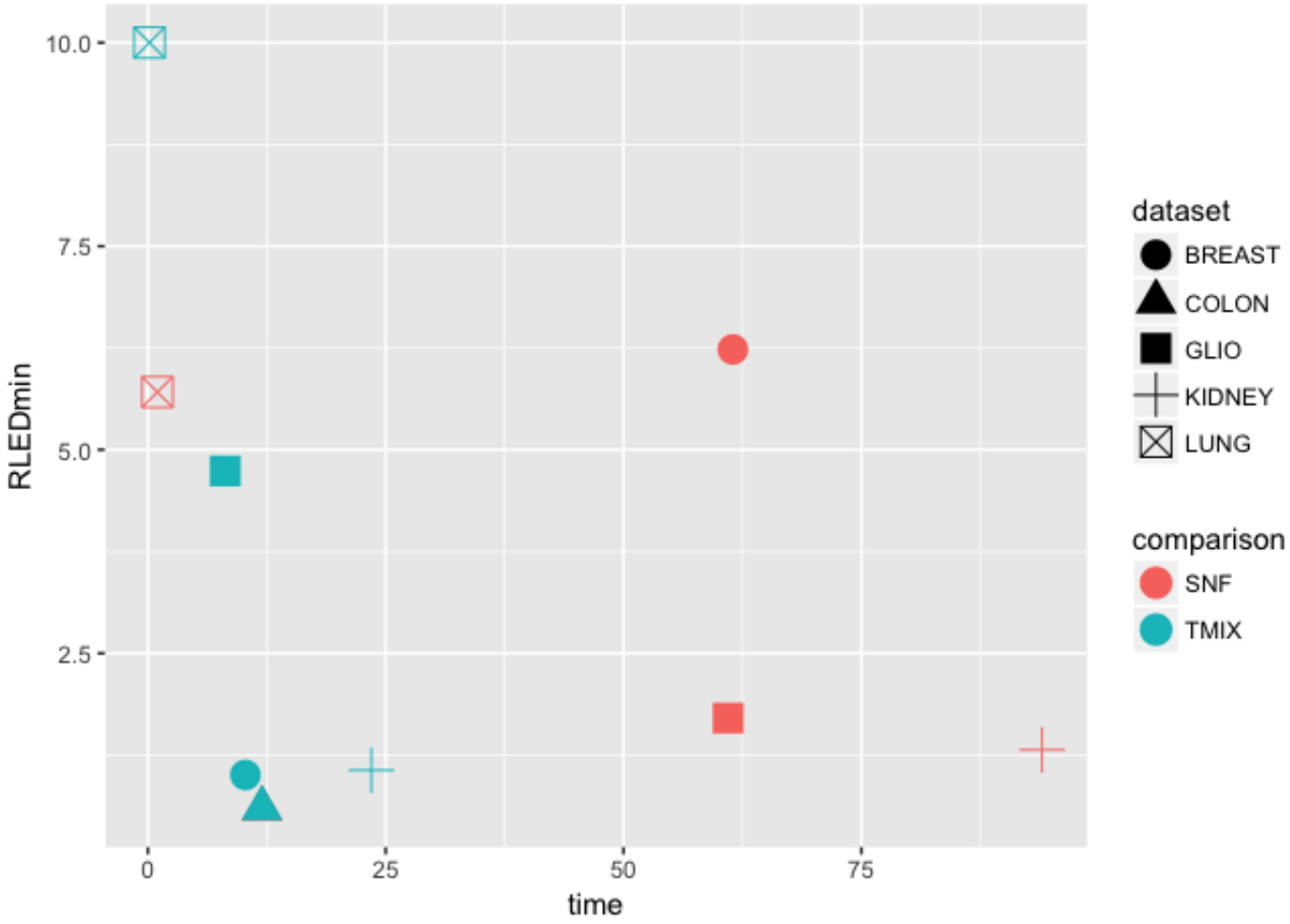
Table 3: OTRIMLE execution time (in seconds). The first column refers to the time needed to compute the decomposition of the RSC matrix. The second column refers to the time needed perform OTRIMLE clustering.

	RSC	OTRIMLE
KIDNEY	7.67	0.94
GLIO	2.48	1.22
BREAST	6.22	1.23
LUNG	2.25	1.01
COLON	6.25	1.12

Table 4: TMIX execution time (in seconds). The first column refers to the time needed to compute the decomposition of the RSC matrix. The second column refers to the time needed perform TMIX clustering.

	RSC	TMIX
KIDNEY	7.67	0.04
GLIO	2.48	0.15
BREAST	6.22	0.12
LUNG	2.25	0.06
COLON	6.25	0.011

Figure 38: The plot shows the relation between the computational time of OTRIMLE method and its gain in term of $RLED_{min}$ for the 5 data sets analysed in this study. The x-axis shows how much bigger is the computational time of OTRIMLE with respect to SNF and TMIX and it is computed as the ration between OTRIMLE execution time sand SNF and TMX execution times respectively . The y-axis show how much bigger is the increase of $RLED_{min}$ of OTRIMLE algorithm with respect to SNF and TMIX and it is computed as the ration between OTRIMLE $RLED_{min}$ values and SNF and TMIX $RLED_{min}$ values respectively



.

References

[1] Coretto, P. and Hennig, C. (2017a). Consistency, breakdown robustness, and algorithms for robust improper maximum likelihood clustering. *Journal of Machine Learning Research*, **18**(142), 1–39.

[2] Coretto, P. and Hennig, C. (2017b). otrimle: Robust model-based clustering. R package (version 1.1). Available at: https://CRAN.R-project.org/package=otrimle.

[3] Hennig, C. (2004). Breakdown points for maximum likelihood estimators of location?scale mixtures. *The Annals of Statistics*, **32**(4), 1313–1340.

[4] McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

[5] Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, **10**(4), 339–348.

[6] Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, **11**(3), 333–337.

[7] Wang, K., Ng, A., and McLachlan, G. (2018). *EMMIXskew: The EM Algorithm and Skew Mixture Distribution*. R package version 1.0.3.