Computational Biology

MEBIOM

Review of CancerSubtypes: an R/Bioconductor package for molecular cancer subtype identification, validation and visualization

Authors:

Ana Mendes (99641)
Inês Carvalho (99672)
Inês Santos(99676)
Mariana Costa (99703)

Group 10

2023/2024 – 1st Semester, P1

Contents

1 Introduction

Cancer is a genetic disease caused by molecular changes in genes that control the growth and proliferation of cells. This disease is highly complex mainly due to the heterogeneity of cancer-associated genes resulting in a massive number of cancer genotypes. Thus, despite some general alterations that are common across cancers, the majority are cancer-specific. Even within each type of cancer, several different subtypes exist. [1] [3] [14]

Given the large number of phenotypes in cancer, which influence the progression of the disease and the response to treatment, it is imperative to optimize the detection of clinical subtypes in order to customise treatment. The development of high-throughput technologies in genomics and transcriptomics enabled the collection of multi-omics data. This data extensively covers the data generated from genome, proteome, transcriptome, metabolome, and epigenome [9], thus providing an integral view of the different processes that occur in a cell, revealing its complexity.

The Cancer Genome Atlas (TCGA) is a wide collection of cancer multi-omics data that integrate molecular and genetic information of various cancers and their subtypes [9] . This initiative as many others enabled a paradigm shift toward multi-omics approaches, which have an enormous potential to enhance the identification of molecular subtypes, and therefore to improve diagnosis and disease prognosis. [3]

Due to the multi-omics data revolution, it is now more achievable to obtain stable, reliable, and large-scale multi-omics data from cancer patients [3]. This allowed a huge step closer to personalized and targeted therapy practises, one of the main goals in cancer treatment.

In this regard, many computational methods for identifying cancer subtypes from multi-omics data have been developed, including the R/Bioconductor package CancerSubtypes. This package has two distinctive features compared to other subtype discovery methods: it enables analysis of a wide range of datasets and the input and output of each step in the framework have the same format. These features enable comparison with the results from different studies. [15]

In this report, the aim was to explore the package CancerSubtypes. We intend to describe the different methods used in feature selection, cancer subtype identification, and results evaluation and visualization. Also, we went through the different application scenarios presented in the Supplementary Material, and uncovered the datasets used, as well as the results obtained by the authors. Furthermore, we tested the program and critically analysed its performance and results. In the end, the program was tested on new data, the dataset BRCA.mRNA from the RTCGA library, and the results obtained agreed with the authors' findings.

2 Methods

2.1 Data pre-processing and feature selection

The CancerSubtypes package manages multi-omics datasets that are usually characterized for having high complexity, dimensionality and variability, as well as containing noise and missing values. Therefore, data pre-processing and feature selection methods are crucial to enhance the data quality and interpretability, along with reducing noise and removing irrelevant features [5]. Regarding data pre-processing, this package provides a distribution check function to examine statistical distribution and presence of patterns, an imputation function to substitute incomplete and missing information and a normalization function to uniformize scales within the datasets. These are operated sequentially to simplify the data preparation and further analysis.

The feature selection methods implemented in this package are Variance (Var), Median Absolute Deviation (MAD), Principal Component Analysis (PCA) and the COX Model (David, 1972). Var, MAD and PCA are commonly used tools to analyse cancer genomic data. Moreover, the Cox Model allows for the selection of specific cancer-related features, through survival analysis. All these techniques are greatly helpful to provide biological insights of the disease and overall implementation of the CancerSubtypes package.

The output of the previously described computational steps is a matrix, in the same data format as the inputs, that is immediately available for the downstream analysis, providing a standardized framework and cohesive workflow among all integrated methods of the package.

2.2 Cancer subtype identification methods

CancerSubtypes covers five frequently cited computational approaches and a unified method for identifying cancer subtypes:

2.2.1 Consensus Clustering (CC)

Consensus clustering (CC) (Monti et al., 2003) is a clustering methodology that estimates the number of clusters in a dataset, assesses the stability of these clusters, and visualizes the results. It uses resampling to determine clusterings of specified counts and calculates pairwise consensus values. This method is popular in cancer genomics, where it combines multiple clusters into a stable single cluster, leading to the discovery of molecular subclasses of disease. Iteratively, a Consensus Matrix is generated at each level, providing graphical displays for determining cluster number and membership. [7] [13]

2.2.2 Consensus non-negative matrix factorization (CNMF)

Consensus non-negative matrix factorization (CNMF) (Brunet et al., 2004) is a combination between NMF and Consensus learning methods, which is efficient for identifying distinct molecular patterns and class discovery.

NMF is an algorithm that reduces the size of expression data from thousands of genes to a few metagenes. NMF can recover meaningful biological information from cancer-related microarray data and is less sensitive to gene selection or initial conditions. It can detect alternative or context-dependent gene expression patterns in complex biological systems, making CNMF a general method for robust molecular pattern discovery. [2]

CC and CNMF are used for single-genomic data sets.

2.2.3 Integrative Clustering (iCluster)

Integrative clustering (iCluster) (Shen et al., 2009) method is a scalable approach that integrates multi-omics data to create a single integrated cluster assignment. This method aligns concordant patterns across multiple data types, revealing potentially novel subclasses by combining weak yet consistent evidence across data types. The iCluster model uses an expectation-maximization algorithm for parameter estimation and a soft-threshold method to divide samples into different subgroups based on latent variables. [8] [11]

2.2.4 Similarity Network Fusion (SNF)

Similarity network fusion (SNF) (Wang et al., 2014) is a method for aggregating multi-omics data to discover patient similarities. It constructs networks of samples for each available data type and efficiently fuses them into one network, using a nonlinear combination method. SNF outperforms single data type analysis and integrative approaches for identifying cancer subtypes and predicting survival. [12]

2.2.5 SNF-CC

SNF-CC is a method proposed by the authors that combines SNF and CC focusing on multi-omics data analysis.

2.2.6 Weighted SNF (WSNF)

Weighted SNF (WSNF) (Xu et al., 2016) is a method that uses the miRNA-TF-mRNA regulatory network to identify cancer subtypes. It uses interatomic databases to build a network representing features like microRNAs, transcription factors, and mRNAs. The weight of these features is calculated using the network information and expression data. The feature weight is then integrated into a network fusion approach to cluster samples and identify cancer subtypes. This approach is modified to consider feature weight when clustering patients for cancer diagnosis. [16]

All of these approaches are bundled to have the same input and output formats for simple comparison.

2.3 Results Validation and Visualization

After obtaining the results, it is crucial to validate and comprehensively analyze the determined subtype of cancer. This process helps grasping the implications of the discoveries and offers valuable insights into the underlying reasons behind these subtypes. To effectively assess and represent these findings, the CancerSubtypes platform offers a set of four robust statistical techniques. These methods serve to ensure a reliable and meaningful interpretation, bridging the gap between computational analysis and biological understanding.

2.3.1 Survival Analysis

It is a statistical method used to analyze the time it takes for an event of interest to occur. This event could be anything that marks a change in status, like death, recovery, relapse, or any other significant outcome. Survival analysis deals with what is known as "time-to-event" data, meaning that instead of looking at a binary outcome (like success or failure), it examines the time taken for an event to occur. The survival function describes the probability that an event has not yet occurred at time 't'. The log-rank test is a hypothesis test used to compare the survival distributions of two or more groups. It helps determine whether there are significant differences in survival rates between groups. The lower the p-value obtained, the higher the difference between groups. [4]

2.3.2 Differential expression

Differential expression refers to the process of comparing the levels of gene or protein expression between two or more different conditions, such as healthy and diseased tissues, or control and treatment groups. The goal is to identify genes or proteins that exhibit significant changes in expression levels in response to a specific experimental condition. This analysis is crucial in fields like genomics and molecular biology as it helps to understand the underlying biological mechanisms driving differences between conditions. By analyzing the expression profiles of a set of genes or proteins, researchers can classify tumors into distinct subtypes based on their molecular characteristics. This classification helps in understanding the heterogeneity of cancer and tailoring treatment strategies. [10] It can be represented through Venn Diagrams (Figure 1). In each individual set there are genes that are differentially expressed in one type of cancer. The intersection between sets represents genes that are differentially expressed in both types of cancer. When a certain gene is present in only one set, it means that the gene is specific to that cancer type. It can be applied to identify genes that are specifically upregulated or downregulated in each type or identify genes that are differentially expressed in different subtypes.

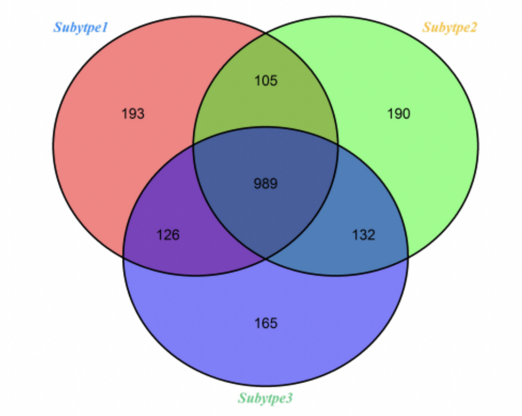


Figure 1: Venn Diagram

2.3.3 Statistical significance of clustering

The article by Y. Liu et al. (2008) [6] addresses the specific challenge of determining whether observed clusters in high-dimensional, small-sample datasets are statistically significant. This is particularly important because, in such data settings, it's easy to obtain clusters by chance alone. High-dimension, low-sample size data refers to datasets where there are a large number of features or variables relative to the number of observations or samples. This scenario is common in fields like genomics, where researchers may have data on thousands of genes but only a limited number of individuals. Clustering is a technique used in data analysis to group similar data points together based on certain characteristics or features they share. It's often employed to identify patterns or subgroups within the data. In statistics, "statistical significance" refers to the likelihood that an observed relationship or difference in data is not due to chance. It's an indication of the strength of evidence against the null hypothesis (which suggests that any observed effect is merely due to random variability). This type of analysis can be made through SigClust summary plots (Figure 2), which evaluates the significance of clustering by comparing the observed clustering pattern in the data to a simulated null distribution. A significant deviation from the null distribution indicates that the observed clustering is unlikely to have occurred by random chance alone.

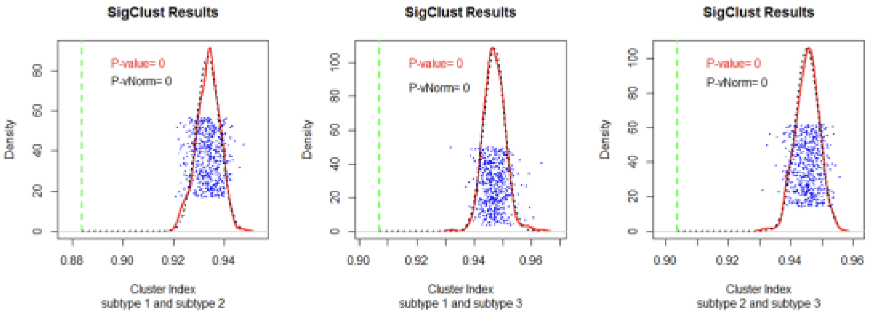


Figure 2: SigClust summary plot

2.3.4 Silhouette Width

Cluster analysis is a technique used in data analysis to group similar data points together. It identifies natural groupings or clusters within a dataset based on the similarity or proximity of data points. After performing a cluster analysis, it's crucial to evaluate the quality of the resulting clusters. This involves assessing how well data points within a cluster are grouped together and how distinct the clusters are from each other.Silhouette width is a metric that quantifies how similar an object is to its own cluster (cohesion) compared to other clusters (separation). It's a measure that ranges from -1 to 1, where: a high value (close to 1) indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters, implying a good clustering; a low value (close to -1) suggests that the object is poorly matched to its own cluster and well matched to neighboring clusters, indicating a questionable clustering; and a value near 0 means that the object is on or very close to the decision boundary between two neighboring clusters.

3 Datasets

The CancerSubtypes package takes **multi-omics** data as input, including gene expression, miRNA expression and DNA methylation. To access cancer data for analysis, the authors propose the use of TCGAbiolinks, TCGA Workflow or RTCGA packages. These packages allows you to search, download and prepare data from The National Cancer Institute (NCI) Genomic Data Commons (GDC) repository for use in R. The data accessed belonged to Level 3 TCGA data, which corresponds to open access data that has already undergone some processing techniques such as normalization and aggregation.

In the application scenarios 2, 3 and 4 presented in the Supplementary Material, the authors processed a glioblastoma multiforme (GBM) multi-omics dataset from TCGA that included gene expression, DNA methylation, miRNA expression data and survival data.

4 Results

4.1 Author's Results

4.1.1 Scenario 1: Using CancerSubtypes with TCGA data to discover cancer subtypes

One of the most typical applications of the CancerSubtypes package is to identify the cancer subtypes using a single genomic data type (gene expression data). As a generic example of this package' potentialities, an analysis was carried out with a level 3 TCGA dataset, which refers to data that was subjected to a long and complex process of analysis to extract biological information and simplify interpretation for researchers.

In order to discover the cancer subtypes, the first step consisted on retrieving clinical and survival data about GBM (Glioblastoma Multiforme, an aggressive type of brain cancer) as well as gene expression data, that were submitted to pre-processing techniques. The samples in the gene expression dataset were then compared to the samples in the survival dataset to check for potential matches. To identify cancer subtypes, the Consensus Clustering method was applied. These results were visualized and validated with survival analysis and silhouette width, that determined that the performance of the Consensus Clustering method was not significant. In fact, the p-value (tool to measure statistical significance) obtained was higher than 0.1 which emphasizes that this identification method is not suitable for the cancer subtype analysis, in these conditions.

A statistical significance of clustering (SigClust) test was then conducted as well as a Differently Expression Analysis for the identified cancer subtypes. All the results obtained reveal that the subtypes selected by the Consensus Clustering Method share similar data points and features, hindering the evaluation of a correct identification.

4.1.2 Scenario 2: Investigating the impact of different feature selection methods in cancer subtype identification

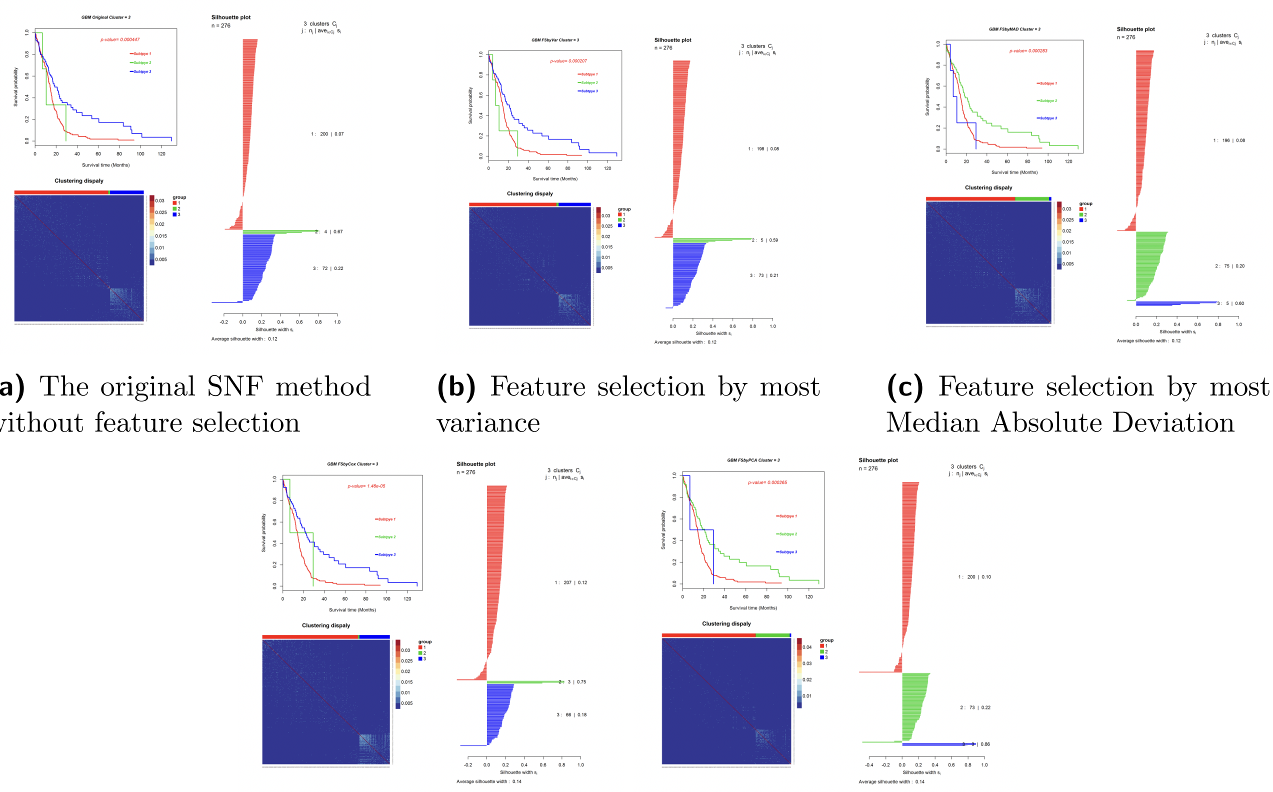
In this scenario, the researchers investigate the impact of different feature selection methods in cancer subtype identification.

For this, they used the similarity fusion network method (SNF) for subtype identification, with data from a glioblastoma multiforme (GBM) multi-omics dataset from TCGA.

The results were analysed for the SNF method with no feature selection (Figure 3a), using the most variance method (Figure 3b), using the Median Absolute Deviation (MAD) (Figure 3c), the COX model (Figure 3d), and finally PCA (Figure 3e).

Based on the survival analysis and the silhouette width plot, the researchers performed an analysis to find the feature selection method that allowed a better performance. Based on the results, it was determined that the Cox model outperformed all the other methods.

Taking the negative logarithm of a p-value is a way to transform it into a more interpretable scale. It compresses the range of p-values, making smaller values (more significant) stand out.



(a) The original SNF method without feature selection

(b) Feature selection by most variance

(c) Feature selection by most Median Absolute Deviation

(d) Feature selection by COX

(e) Feature selection by PCA

Figure 3: Survival curves and Silhouette plots for the identified cancer subtypes using different feature selection methods

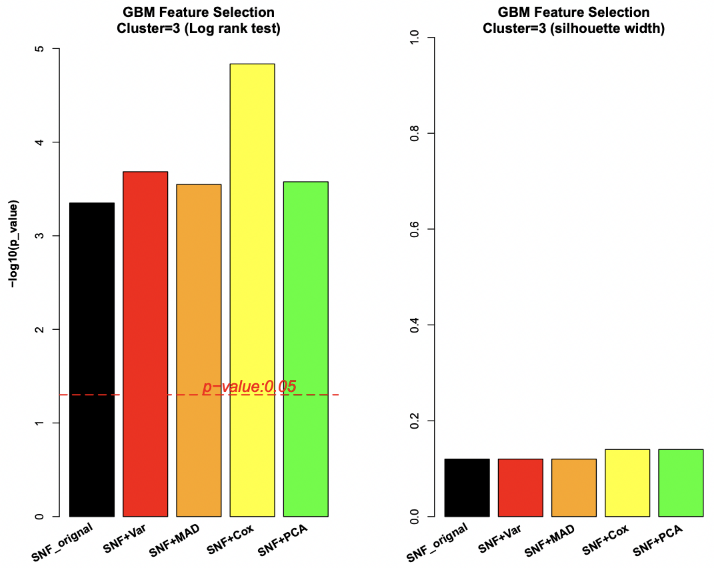


Figure 4: The barplot for the Log-rank test p-values and Silhouette width of each feature selection method

In the survival analysis, the lower the obtained p-value is, the better the features separate between subtypes. In this case, classification will be better the higher the values are in the barplot for the Log-rank test p-values in Figure 4.

Based on this, we can conclude that the Cox model performed best based on the survival analysis results. In the silhouette plots, wider silhouettes indicate data points that are well-clustered and have a strong affinity for their assigned cluster compared to neighboring clusters. This is a positive indicator of the quality of the clustering results. Based on this, according to the barplot for the Silhouette width of each feature selection method in Figure 4, Cox and PCA produce higher widths of the silhouette plot, which means the clustering results are better.

Joining the information from both barplots in Figure 4, we get the conclusion that the Cox model performed the best.

4.1.3 Scenario 3: Comparing the performance of different cancer subtype identification methods.

This scenario analyses the performance of the six previously mentioned clustering algorithms on cancer subtype identification, using the same input dataset for each clustering method and GBM gene expression and miRNA expression datasets for experiment analysis.

As input data for CC and CNMF, the gene expression data and miRNA expression data for each patient were concatenated.

The Log-rank test p-value was used for identifying cancer subtypes. The Silhouette width is not listed for comparison due to different numerical levels of similarity matrix for each method, despite providing crucial information for insight investigation.

Therefore, and bearing in mind Figure 5, SNF and its variants (SNF.CC and WSNF) perform the best in this dataset, due to their higher values in the barplot.

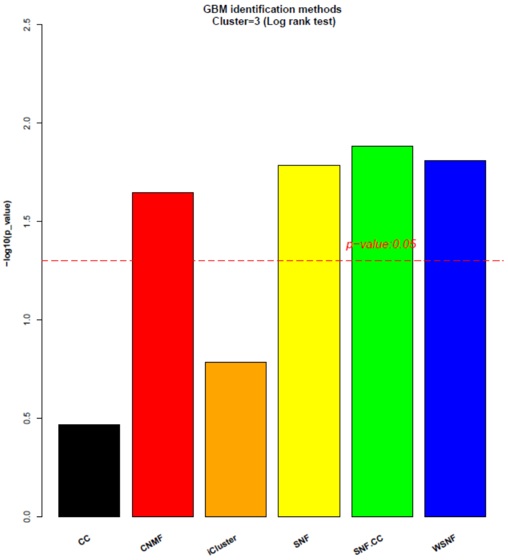
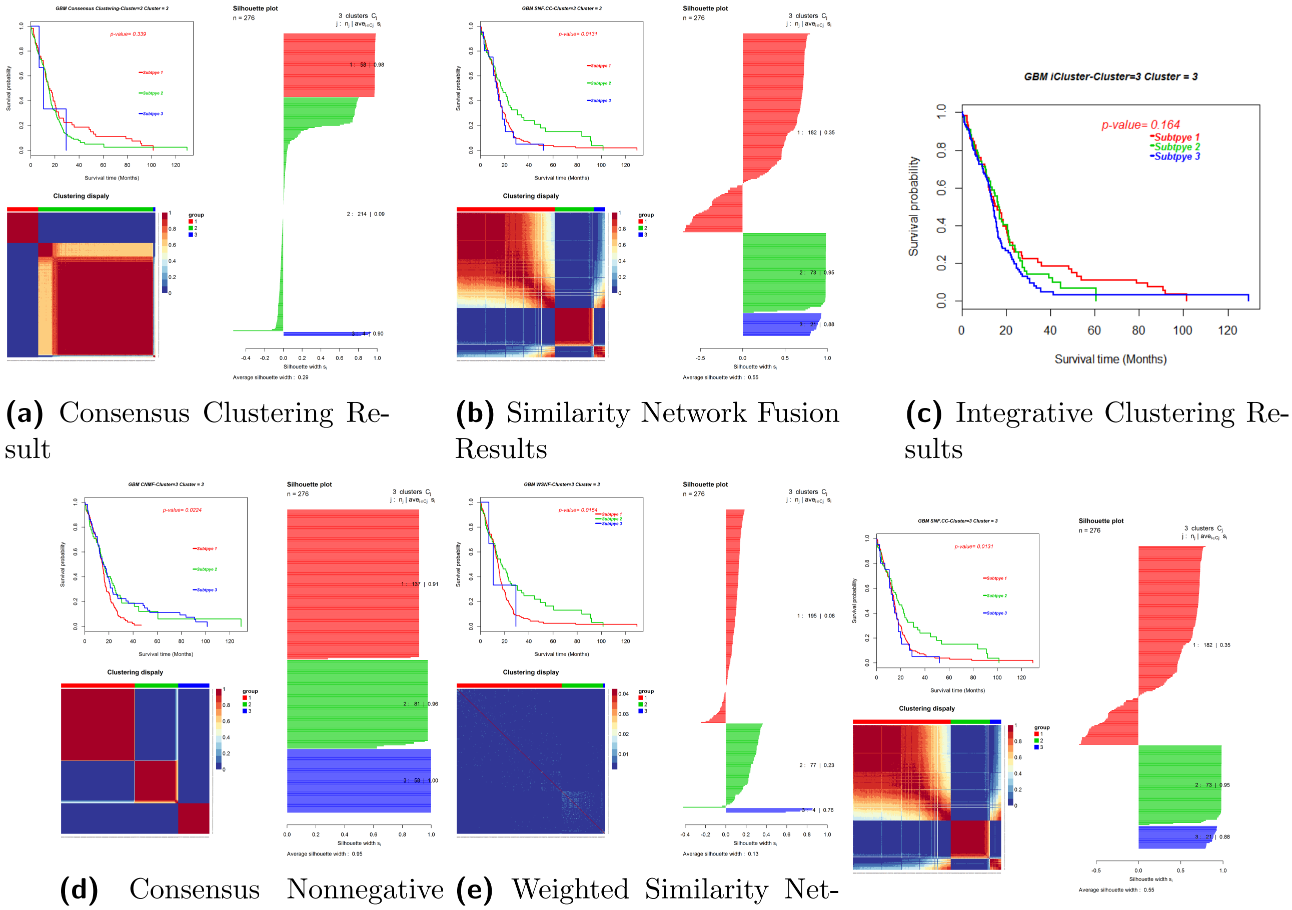


Figure 5: Barplot for the Log-rank test p-values of each cancer subtypes identification method



(a) Consensus Clustering Result

(b) Similarity Network Fusion Results

(c) Integrative Clustering Results

(d) Consensus Nonnegative Matrix Factorization Results

(e) Weighted Similarity Network Fusion Results

(f) SNF+CC Results

Figure 6: Survival curves and Silhouette plots for the identified cancer subtypes using different cancer subtype identification methods

4.1.4 Scenario 4: Investigating how the impact of different genomic data types alters the results with the selected feature selection and cancer subtype identification methods.

In this scenario, the aim was to compare the use of different multi-omics data types. As in the previous scenarios, a glioblastoma multiforme multi-omics dataset was chosen for analysis. Six groups were compared in order to obtain a comprehensive comparison:

1. mRNA-Var-CC
2. mRNA+miRNA-Var-CC
3. mRNA+miRNA-Var-SNF
4. mRNA+DM+miRNA-Var-SNF
5. mRNA+miRNA-COX-SNF
6. mRNA+DM+miRNA-COX-SNF

Consensus Clustering (CC) was the cancer subtype identification method used for single dataset input and Similarity Network Fusion (SNF) the one used for multiple datasets input clustering. Given the diversity across groups regarding data type and methods used in feature selection and cancer identification, it is possible to conclude not only about the impact of different genomic data types, but also about the best feature selection and cancer identification methods. In order to compare the differences between groups, a log-rank test was performed. The test results are presented in Figure 7.

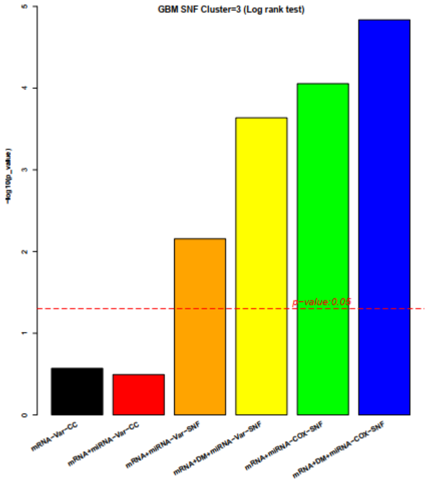


Figure 7: The barplot for the Log-rank test p-values

In the plot, there is a significant difference in p-value between the groups with single dataset input and the groups with multiple datasets. Therefore, based on this results we can infer that the more multi-omics datasets the program receives, the easier it is to identify cancer subtypes. Also, group 6 (multi-omics data with COX model for SNF) was the one that showed higher potential to unveil distinct cancer subtypes with significant varying survival patterns.

Considering that the feature selection by COX and the cancer subtype identification by SNF were the methods previously considered the best among the others in this dataset, we conclude that the results of this scenario are in line with the previous scenarios.

4.2 Testing the program

4.2.1 Evaluation

In what regards to performance, a script similar to the one provided on *CancerSubtypes* site takes approximately 8 min to run, with the BRCA.mRNA dataset. The size and complexity of the dataset being analysed have a significant impact on scalability and performance.

The differential analysis of the expression of genes, doesn't work for some datasets, like the COAD.mRNA and LAML.methylation, from TCGA, due to the inability to form more than one cluster.

Furthermore, the iCluster method didn't converge in all of our testings, what suggests an error on the implementation of the corresponding algorithm.

In terms of server implementation viability, the code can be successfully executed on a server.

Given the examples of scenarios in which to use this package, which provide code to integrate all the steps that can be performed with *CancerSubtypes*, it is relatively simple to adapt those examples, and also the script provided in the site, to the needs of the user.

4.2.2 Test on new data

We used the RTCGA.mRNA library to access mRNA datasets from The Cancer Genome Atlas Project. The dataset selected (BRCA.mRNA) is part of a project focused on Breast Invasive Carcinoma. [ref]

In this example, we decided to compare two methods of cancer subtype identification: CC and SNF. The feature selection method used was the Variance in both cases. The results obtained are presented in Figure 8.

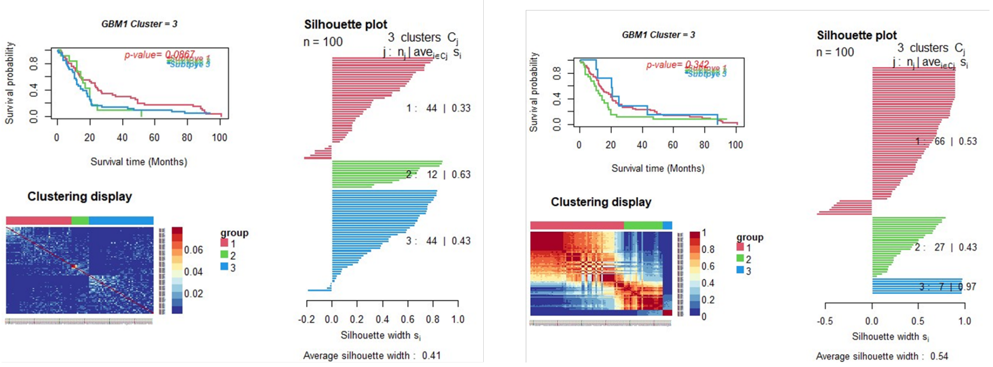


Figure 8: The Survival curves and Silhouette plots for the identified cancer subtypes of BRCA. On the left, the results obtained using SNF and, on the right, using CC.

As previously mentioned, in the survival analysis, a lower p-value indicates a better distinction between subtypes. The p-values obtained were 0.0867 and 0.342 for SNF and CC, respectively. The p-value associated with the SNF method is significantly smaller, indicating its superior ability to differentiate subtypes within this dataset when compared to the CC method. This outcome aligns with the findings observed in other datasets.

5 Conclusion

The R/Bioconductor package CancerSubtypes was developed with the purpose of establishing an accurate identification of several cancer subtypes, based on a multi-omics dataset input. This package includes all the tools needed to perform a reliable and complete analysis, in a standardized framework, from the initial data pre-processing step to the cancer subtype identification methods and validation of the obtained results.

After analysing the original authors' application scenarios, the conclusions reached were that, in terms of feature selection methods included in the package, the Cox Model allowed the best performance of the clustering results. When it comes to the cancer subtypes identification methods, the SNF method appears to be the most suitable, providing the most accurate results among the four other computational techniques analysed. Similarly, it was shown that a higher number of genomic datasets also enhances the quality and validity of the outcomes. The following tests conducted with new mRNA datasets confirmed and emphasized these inferences.

To conclude, the CancerSubtypes package reveals immense potential in computational biology, among several other areas, allowing users to make solid and trustworthy predictions, when analysing and identifying cancer subtypes from multi-omics datasets. This package could act as a starting point to revolutionize and optimize personalized medicine, especially when targeted to cancer diagnosis and treatment.

References

[1] Pavlicová M. Robles-Espinoza C. D. Peña J. G. T. Treviño V. Ayton, S. Multiomics subtyping for clinically prognostic cancer subtypes and personalized therapy: A systematic review and meta-analysis. *Genetics in Medicine*, 24(1):15–25, 2022.

[2] Tamayo P. Golub T. R. Mesirov J. P. Brunet, J.-P. . Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164—4169, 2004.

[3] Wen Y. Xie C. Chen X. He S. Bo X. Zhang-Z. Chen, Y. Mocss: Multi-omics data clustering and cancer subtyping via shared and specific representation learning. *iScience*, 26(8):107378, 2023.

[4] Timothy G Clark, Michael J Bradburn, Sharon B Love, and Douglas G Altman. Survival analysis part i: basic concepts and first analyses. *British journal of cancer*, 89(2):232–238, 2003.

[5] Jin Gu Dongfang Wang. Integrative clustering methods of multi-omics data for molecule-based cancer classifications. *Quantitative Biology*, 4(1):58, 2016.

[6] Yufeng Liu, David Neil Hayes, Andrew Nobel, and JS Marron. Statistical significance of clustering for high-dimension, low–sample size data. *Biostatistics*, 9(1):61–75, 2008.

[7] Tamayo P. Mesirov J. et al. Monti, S. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52(1/2):91–118, 2003.

[8] Ladanyi M. Shen R, Olshen AB. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906—2912, 2009.

[9] Verma S. Kumar S. Jere A. Anamika K. Subramanian, I. Multi-omics data integration, interpretation, and its application. *Bioinformatics and Biology Insights*, 14, 2020.

[10] Sonia Tarazona, Fernando García-Alcalde, Joaquín Dopazo, Alberto Ferrer, and Ana Conesa. Differential expression in rna-seq: a matter of depth. *Genome research*, 21(12):2213–2223, 2011.

[11] Wang C. Tian, S. An ensemble of the icluster method to analyze longitudinal lncrna expression data for psoriasis patients. *Human Genomics*, 15(1), 2021.

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

[13] M. Wilkerson. Consensusclusterplus (tutorial). 2023.

[14] D. S Wishart. Is cancer a genetic disease or a metabolic disease? *EBioMedicine*, 2(6):478–479, 2015.

[15] Le T. D. Liu L. Su N. Wang R. Sun B. Colaprico-A. Bontempi G. Li J. Xu, T. Cancersubtypes: an r/bioconductor package for molecular cancer subtype identification, validation and visualization. *Bioinformatics*, 33(19):3131–3133, 2017.

[16] Le T. D. Liu L. Wang R. Sun B. Li J. Xu, T. Identifying cancer subtypes from mirna-tf-mrna regulatory networks and expression data. *PLOS ONE*, 11(4), 2016.