UNIVERSITAT ROVIRA i VIRGILI

Integration of multi-omics data for cancer subtype identification

Adrià Olomí Farré

FINAL DEGREE PROJECT

Supervised by Dr. Aida Valls Mateu

Bachelor's Degree in Computer Science

In cooperation with

Department of Biostatistics and Bioinformatics

Acibadem University, Türkiye.

Head Dr. Osman Uğur Sezermann

Tarragona, September 2023.

Index

1. Introduction ____ *6*

1.1. Hypothesis and goals ____ *9*

2. Project Development ____ *10*

2.1 Functional requirements ____ *10*

2.2 Design of the software tool ____ *10*

2.2.1. Fundamental Concepts for Work Understanding ____ *13*

2.2.2 Integrative methods used ____ *15*

2.3 Implementation ____ *24*

2.3.1. R programming language ____ *24*

2.3.2. Detailed implementation ____ *24*

2.4. Metrics and evaluation methods ____ *28*

2.4.1. Rand Index and Adjusted Rand Index ____ *28*

2.4.2. Survival Analysis ____ *28*

3. Experimentation and Results ____ *30*

4. Discussion ____ *31*

4.1. Number of datatypes ____ *31*

4.2. Dual and Triple Grouping ____ *32*

4.3. Consensus matrix ____ *33*

4.4. Number of steps in Walktrap algorithm ____ *34*

4.5. Test set ____ *35*

5. Conclusion and future goals ____ *36*

5.1. Workflow assessment and conclusions ____ *36*

5.2. Personal review ____ *36*

6. References ____ *38*

Figures index

Figure 1. *Representation of the omics sciences relationships* ____ *6*
Figure 2. *Detailed schema of the whole workflow that has been developed in this project* _ *12*
Figure 3*. Example of an adjacency matrix and its respective graph representation*____ *13*
Figure 4. *Representation of the integral calculated*____ *13*
Figure 5. *Representation of two (a) consensus matrix histograms* ____ *14*
Figure 6*. Illustrative example of SNF steps*____ *16*
Figure 7. *Schema of SNF algorithm*____ *17*
Figure 8*. Schema of the iCluster+ algorithm* ____ *18*
Figure 9. *Schema of the CCPlus algorithm* ____ *20*
Figure 10. *Schema of the PINS algorithm* ____ *22*
Figure 11*. Walktrap algorithm action* ____ *23*
Figure 12*. Adjacency matrix of the integrative method, and visualization of the graph corresponding to the adjacency matrix*____ *27*
Figure 13. *Survival curves for both women with Astrocytoma and Glioblastoma by diagnosis* ____ *29*

Tables index

Table 1*. Most important cancer omics data repositories*____ *7*
Table 2. *Sub-groups labelling returned by each clustering method* ____ *26*
Table 3. *Adjacency matrices returned by each method* ____ *27*
Table 4*. Subtyping results of PINS, CC, SNF, iClusterPlus, and the integrative method developed for the 6 cancer diseases* ____ *30*
Table 5. *Subtyping results of the integrative method developed for the 6 cancer diseases using dual or triple grouping* ____ *32*
Table 6*. Adjacency and consensus matrix concepts*____ *33*
Table 7*. Study of the outcome of the consensus matrix approach using different values of threshold (from 1.0 to 2.0 by 0.25 intervals)* ____ *33*
Table 8. *Subtyping results of the integrative method developed for the 6 cancer diseases using different number of steps in Walktrap algorithm* ____ *34*
Table 9. *Test set* ____ *35*

Abbreviations

ARI · Adjusted Rand Index

AUC · Area Under the Curve

BRCA · Breast Invasive Carcinoma

BIC · Bayesian Information Criterion

CC · Consensus Clustering

CDF · Cumulative Distribution Function

COAD · Colon Adenocarcinoma

DNA · Deoxyribonucleic Acid

GBM · Glioblastoma Multiforme

KIRC · Kidney Renal Clear Cell Carcinoma

LAML · Acute Myeloid Leukemia

LUSC · Lung Squamous Cell Carcinoma

PINS · Perturbation Clustering for Data Integration and Disease Subtyping

RI · Rand Index

RNA · Ribonucleic Acid

SNF · Similarity Fusion Network

TCGA · The Cancer Genome Atlas

Abstract

Continuous advances in high-throughput technologies allow for measurements of many types of omics data at a highly detailed level. This permits both the scientific and bioinformatics community to develop novel methods to answer different biological questions. Nowadays, one of the main questions to assess is cancer patients' classification. Specifically, classification of patients into existing sub-groups and for new subtypes discovery. State-of-the-art methods in this field are Similarity Fusion Network (SNFtool), iClusterPlus, Perturbation Clustering for Data Integration and Disease Subtyping Plus (PINS+) and Consensus Clustering Plus (CC+). Although they are powerful, they are not always able to subtype data in a correctly manner. Knowing the capabilities of each method and having the goal to improve the data integration process, a workflow that can integrate all the existing subtype methods is developed. The combination of those existing multi-omics data integration packages in a workflow improves the subtyping of cancer patients. Future efforts should be addressed principally into omics databases standardization and, obviously, the improvement of the existing integration methods by searching the best parameters or by finding novel approaches.

1. Introduction

Omics sciences are various disciplines in biology whose names end in the suffix –omics, such as genomics, transcriptomics, proteomics and metabolomics. The main purpose of these 'omics' is to identify, characterize, and quantify all the biological molecules that play a role in the structure, function and dynamics of a cell, tissue, organ or organism (Vailati-Riboni et al., 2017).

Firstly, genomics studies the structure, function, evolution and mapping of the genome. The genome is the set of all the genes that are codified within the Deoxyribonucleic acid (DNA). The genes contain all the genetic information and for the protein production. For its part, the transcriptome is the set of ribonucleic acid (RNA) that can be obtained from the genome, a field that is studied by transcriptomics. One type of RNA - messenger RNA (mRNA) - brings the genetic sequence of a gene and is read by an organelle, called ribosome, to synthesize proteins. All the set of proteins is called proteome. Proteomics is, then, the science that studies those proteins and their biochemical properties and functional roles, and how their quantities, modifications, and structures change during growth and in response to internal and external stimuli. Finally, the last level of the omics science is metabolomics. This field studies small molecules, called metabolites. Collectively, these small molecules and their interactions within a biological system are known as the metabolome. The whole cascade process is exemplified in *Figure 1*.

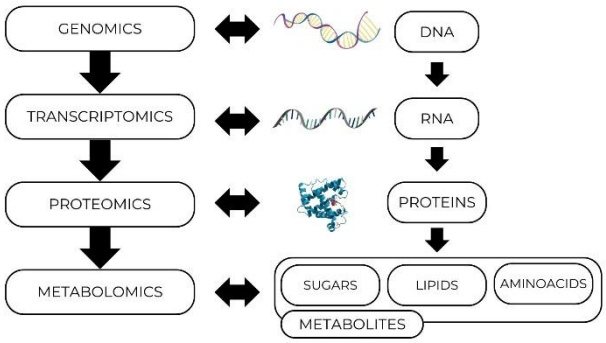


Figure 1. Representation of the omics sciences relationships.

All together, these wide set of science fields can bring responses and lead to insights that are crucial for improve biological knowledge of all organisms. Moreover, they are relevant to understand the mechanism of diseases, as well as the discovery of biomarkers, known to be substances that can be related with a disease development.

One of the most studied diseases by the global scientific community is cancer. The word cancer - malignant tumour or neoplasm - is a generic term for a large group of diseases that can affect any part of the body. The main feature of cancer is the rapid creation of abnormal cells that grow beyond their usual boundaries, and which can then invade nearby parts, such as close tissues, and which can also spread to other organs, in a process named as metastasis. Widespread metastases are the primary cause of death from cancer (Cancer, n.d;Ferlay et al., 2021).

Cancer is a death leading cause worldwide, almost reaching the amount of 10 million deaths in 2020. In terms of new detected cases, the most common cancer diagnoses were breast (2.26

million cases), lung (2.21 million cases) and colon-rectum (1.93 million cases). In mortality terms, the most common cancer causes of death were lung (1.80 million deaths), colon-rectum (916 000 deaths) and liver (830 000 deaths). However, the most common cancers vary between countries.

The Cancer Genome Atlas (TCGA; https://cancergenome.nih.gov/) is one of the most important omics caner data repositories (see *Table 1*). It contains collections of multi-omics datasets for 33 types of cancer for 20 000 individual tumour samples. Among them, there are DNA, RNA, protein and epigenetic data together with histological and clinical data, trespassing the size of 2.5 petabytes of data. Moreover, it saves both molecular and genetic profiles from primary tumour and their respective subtypes. (*The Cancer Genome Atlas Program (TCGA) - NCI*, n.d.). These all characteristics lead TCGA to be one of the most used sources of multi-omics data when dealing with cancer data.

Table 1. Most important cancer omics data repositories. Abbreviations: CNV, copy number variation; miRNA, microRNA; RPPA, reverse phase protein array; SNP, single-nucleotide polymorphism; SNV, single-nucleotide variant. Taken from (Subramanian et al., 2020).

DATA REPOSITORY	WEB LINK	DISEASE	TYPES OF MULTI-OMICS DATA AVAILABLE
The Cancer Genome Atlas (TCGA)	https://cancergenome.nih.gov/	Cancer	RNA-Seq, DNA-Seq, miRNA-Seq, SNV, CNV, DNA methylation, and RPPA
Clinical Proteomic Tumor Analysis Consortium (CPTAC)	https://cptac-data-portal.georgetown.edu/cptacPublic/	Cancer	Proteomics data corresponding to TCGA cohorts
International Cancer Genomics Consortium (ICGC)	https://icgc.org/	Cancer	Whole genome sequencing, genomic variations data (somatic and germline mutation)
Cancer Cell Line Encyclopedia (CCLE)	https://portals.broadinstitute.org/ccle	Cancer cell line	Gene expression, copy number, and sequencing data; pharmacological profiles of 24 anticancer drugs
Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)	http://molonc.bccrc.ca/aparicio-lab/research/metabric/	Breast cancer	Clinical traits, gene expression, SNP, and CNV
TARGET	https://ocg.cancer.gov/programs/target	Pediatric cancers	Gene expression, miRNA expression, copy number, and sequencing data
Omics Discovery Index	https://www.omicsdi.org	Consolidated data sets from 11 repositories in a uniform framework	Genomics, transcriptomics, proteomics, and metabolomics

Cancer is a heterogeneous disease. Hence, is not unreasonable to think that different types of cancer can also have several subtypes with different phenotypic and molecular profiles (Kuijjer et al., 2018). The patient classification is then relevant for the early diagnosis of cancer and for choosing an adequate treatment. An early diagnosis of cancer can increase the survival probability of those patients. In that sense, the identification of a clinically relevant subtype is crucial for selecting and administering the most effective treatment, as different cancer subtypes may respond differently to specific treatments (Gao et al., 2019).

This thesis will work with an integrative view to omics for cancer subtype determination. The main purpose of an integrative analysis is to give a comprehensive vision for the subgroup identification. This is not considering only one datatype e.g., mRNA, but other levels like miRNA, DNA methylation and more.

Talking about omics data integration, there are a lot of approaches that can be used for that goal. One method consists of analysing each datatype independently. Then, a group of experts analyse the results and try to integrate all the results. However, this approach may lead to

irregular results that can be difficult to interpret (Muzny et al., 2012). Another strategy is to concatenate all the inputs to a single matrix and then apply clustering methods, such as correlation distance and partitioning around medoids, to classify the patients. However, the integration of data without considering their scale, their source and the methods that have been used usually leads to wrong classifications (Kim et al., 2015).

Finally, there are currently few studies that can integrate muti-omics data interpretation together with clinical data. Actually, this approach could have a lot of interests. For instance, there is the possibility to stablish relations between some clusters of samples and the survival probabilities of those samples, compared to other clusters. Moreover, majority of those studies are not still robust. However, new progressions are taking place nowadays to relate both aspects.

Even though multi-omics data integration leads to a comprehensive knowledge of both biological processes and diseases, it comes with a large set of challenges. One big challenge is the enormous heterogeneity that exists among all the individuals, generating a big contrast in the data values between them. Moreover, these data sets are very large and hard to study. Hence, an intensive computation is often needed. Additionally, data is sometimes coming from different platforms. Every platform has its specific data format.

In that sense, a pre-processing step must be performed before the beginning of the integration. Commonly, the pre-processing step include data filtering, normalization, batch effects removal and a final verification. Is important to set the pre-processing parameters to reduce the data size significantly (if possible) but also avoiding relevant data features removal. However, choosing the optimal settings to reach this balance is a challenging task.

Depending on its goal, the methods to integrate different multi-omics data types, can be distinguished as following:

- *Disease subtyping* · They try to classify the cases into different subtypes. This can help to discover the etiology of the disease and identify suitable interventions for patients belonging to different subtypes. Furthermore, they try to find new subtypes of the disease.
- *Disease insights* · This method allows the discovery of key factors of different diseases. It permits the study of the complex biology of each disease.
- *Biomarker prediction* · They can help to find the variables (a group of factors) that cause a disease. They can also predict the mortality risk and the features associated with it.

In the present work, the focus is set in the disease subtyping. Then, the efforts have been directed to the patients subclassification. This could conduct into different situations as (1) classification of patients into existing sub-groups and (2) new subtypes discovery.

The programming language R is an open-source tool that has a widely set of packages that are focused on biological purposes. It is a tool easy to use for the scientific community and has a big set of both bioinformatical and plotting tools that help his use. All these characteristics have converted R in one of the most used packages to face biological questions, including multi-omics data integration.

In that sense, some methods have tried to perform this integration. Specifically, the R packages called Similarity Fusion Network (SNFtool), iClusterPlus, Perturbation Clustering for Data Integration and Disease Subtyping Plus (PINS+) and Consensus Clustering Plus (CC+). Although these approaches have been improved in the recent years, they are still not working for all types of cancer. Consequently, some patients subtyping results are wrong and must be afforded considering more different parameters.

1.1. Hypothesis and goals

Multi-omics data integration methods together with clinical data have taken the front seat for discovering useful insights in human health. Even though some tools are starting to perform well for specific cases, a confident method for subtyping patients from different cancer types is still missing. Additionally, there is no evidence of a complete workflow starting from the data download until the results visualization and interpretation.

Hence, the hypothesis is: the combination of the existing multi-omics data integration packages in R improve the subtyping of cancer patients.

In order to satisfy the hypothesis and the main goal, the work is divided into the following specific goals:

- Develop a pre-processing step that can reduce the gap within the size reduction of the data and the loss of relevant information.
- Integrate the existing R packages for multi-omics data integration.
- Find the best methods to assess the workflow data outcome.
- Study the main parameters implied in the whole process and try to find the most optimal values for each case.

In order to check the developed workflow, 6 different cancer datasets are downloaded from the TCGA public and free repository. The datasets include 3 datatypes files, including mRNA expression, miRNA expression and DNA-methylation beta-values. The datasets refer to different types of cancer: (1) Colon Adenocarcinoma (COAD), (2) Glioblastoma Multiforme (GBM), (3) Kidney Renal Clear Cell Carcinoma (KIRC), (4) Breast Invasive Carcinoma (BRCA), (5) Acute Myeloid Leukemia (LAML) and (6) Lung Squamous Cell Carcinoma (LUSC).

2. Project Development

2.1 Functional requirements

To solve the main problem of integrating multi-omics data for cancer subtyping detection, the list of the main requirements that must be achieved in this work is the following:

- 1st requirement. Sample files preparation:
  - Sample file download · Be able to download files from cancer repositories depending on some search terms and filters.
  - Sample file standardization · Usually data is coming from different sources; hence, they come with different formats. However, all the files must have the same extension and the same input format in order to be processed by the workflow.
- 2nd requirement. Data integration:
  - Choose the best k (number of clusters) for each situation. Is hard to choose a k value without knowing the samples constitution. To solve that, a set of metrics must be used to evaluate as deeply as possible the situation and, hence, choose a k value according to the terms.
  - Each method is facing the problem by different ways. They are not changing only method, but the pre-processing part including the normalization and features filtering by relevance. A workflow gathering all pre-processing steps model may be interesting to check how each method is affording the problem.
  - Evaluation metrics and data assessment. There are a wide range of metrics to evaluate the data outcome. However, some of them strongly depends on clinical data existence, that is, a register alive and dead patients together with the number of days to death. Find the best metrics in each situation could be interesting to face correctly the assessment.

2.2 Design of the software tool

The workflow designed to solve this problem consists of two basic modules: (A) sample files preparation and (B) integration, as displayed in *Figure 2*:

(A) Sample file preparation module

This module is assuming the files have been correctly downloaded from TCGA and have been distributed to their respective directories (‘Control and ‘Disease’). Frequently, when performing a clinical experiment, both healthy (control) and diagnosed patients are included in the study. Control directories contain files related to the samples of the Control group; these are, the patients that are healthy. Disease directories contain files related to the samples of the Disease group; these are, the patients diagnosed of cancer. Each file refers to one datatype analysis of one sample. Hence, each sample has three different files: one for the mRNA expression, another for the miRNA expression and the last one for the DNA-methylation.

The first module, as the name is suggesting, is the responsible to prepare all the files that are required to perform the clustering algorithm.

Another requirement to assure is the fact that the project is dealing with the same samples all the time. This means that if some file – related to one specific sample – is missing for one datatype, then it must be removed from the analysis. For instance, there is a sample 'A' with both mRNA and miRNA expression values; but if the DNA-methylation file is missing, this sample will not be included in the analysis.

In the mRNA and miRNA files, there is information related to the number of reads per mRNA or miRNA fragment. The number of reads is the number of fragments that we are observing that fragment in a specific sample. For example, if we are observing 10 000 times the miRNA-AAA fragment, the number of reads will be 10 000.

However, the integration methods cannot work with the number of reads. Hence, a transformation of the vales must be performed. This transformation consists of converting all these read values to fold-change values. This parameter is relating the number of reads within the control group and the disease group. The result of the fold change calculation tells us how many times a feature is more abundant (or less abundant) in the experimental condition compared to the control condition. If the fold change is equal to 1, it means there is no difference in abundance between the two conditions. If the fold change is greater than 1, it indicates an increase in abundance in the experimental condition compared to the control. If the fold change is less than 1, it indicates a decrease in abundance in the experimental condition. To perform this conversion, functions from *DESeq2* R package are applied to the sample files of mRNA and miRNA datatype. These new files contain a unique fold-change value per feature per sample.

For its part, DNA methylation is a chemical modification in which methyl groups (-CH3) are added to the cytosine (C) bases in CpG regions (dinucleotides composed of a cytosine followed by a guanine) of DNA. These modifications can impact gene expression and play a crucial role in gene regulation. Beta values are commonly used in DNA methylation studies to represent the degree of methylation at a specific CpG site on a scale ranging from 0 to 1, where a beta value (β) of 0 indicates that the site is completely unmethylated (no methyl groups), and a beta value (β) of 1 indicates that the site is fully methylated (all methyl groups are present). That is, DNA-methylation files contain values (called beta-values) that are assigned to features (called CpG sites). Each sample has the same number of features. Those beta-values are compatible as input of the algorithm. Hence, in the DNA-methylation files, the values do not need to be converted to other formats.

However, these methylation files are usually bigger than the mRNA and miRNA ones. This is frequently due to the extra information that contain. Because the range of features (CpG sites) is big, most of times all these features are not covered with values, and they contain zeroes. To solve that, a filtering step consisting of irrelevant features removal is needed. This step is performed once the beta-values of all the features coming from all the samples have been gathered. Is then when there is the possibility to identify which features have majority of zero values, and they can be removed consequently.

Once finished the first module *(A) Sample file preparation*, the module *(B) Integration* is where all the multi-omics data integration using clustering methods is taking place. But, before

deeping into the algorithms of this part, a previous explanation of the tools set used in the project must be exhibited. Find this basic explanation of the methods and the metrics used in the following section.

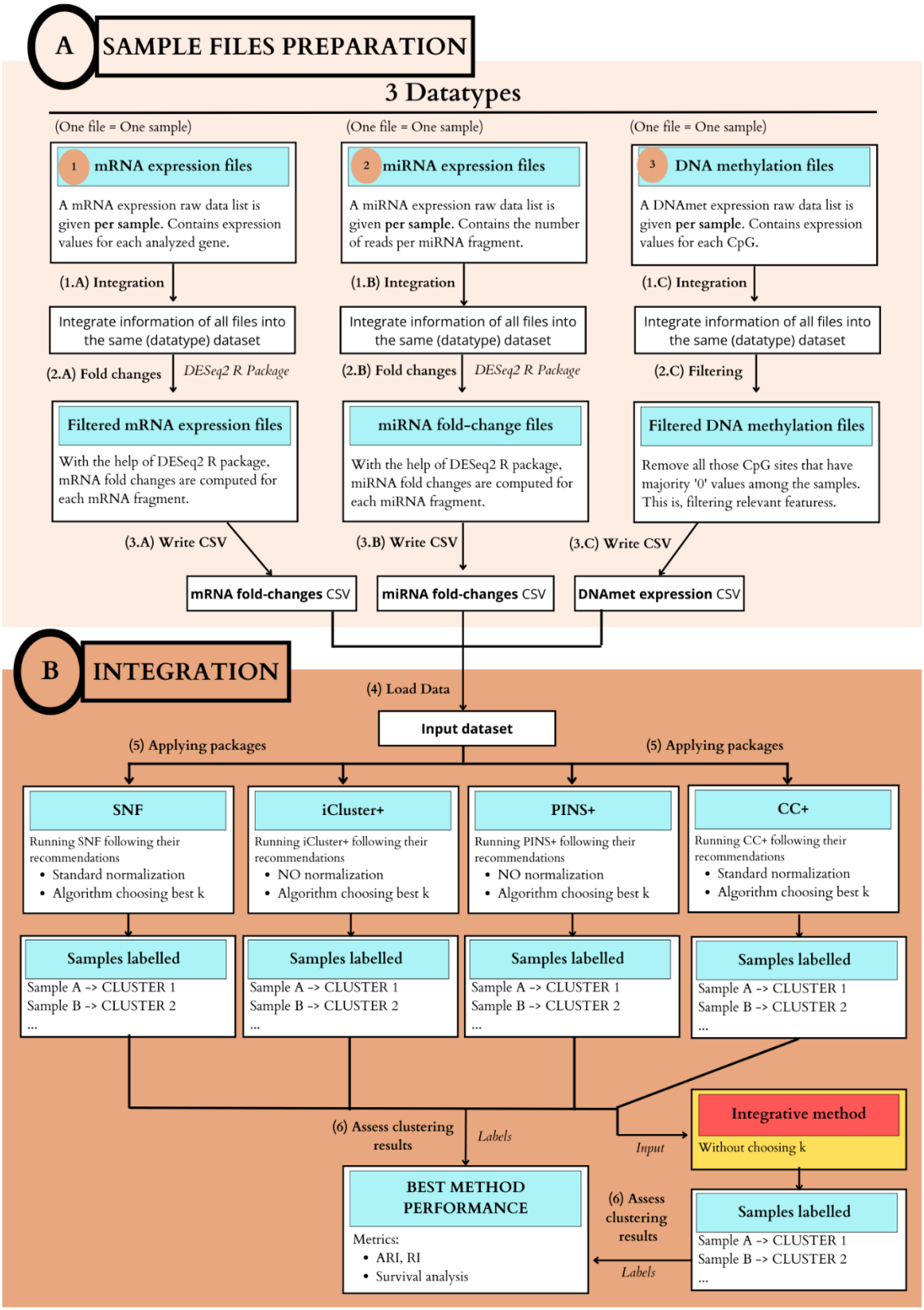


Figure 2. Detailed schema of the whole workflow that has been developed in this project.

2.2.1. Fundamental Concepts for Work Understanding

In the present manuscript, some advanced terms related to graph theory and clustering methods will be introduced. Find the basic definitions below.

2.2.1.1. Adjacency matrix

An adjacency (or connectivity) matrix is a data structure used in graph theory to represent a finite graph. It provides a concise way to describe the relationships between nodes (or vertices) in a graph. In an adjacency matrix, rows and columns correspond to graph nodes, and the entries in the matrix indicate whether there is an edge connecting two nodes.

In *Figure 3*, an example of an adjacency matrix id displayed together with its respective graph representation.

	A0CQ	*A0CU*	*A0CY*	*A0CZ*	*A0D0*
A0CQ	0	1	0	1	0
A0CU	1	0	0	1	0
A0CY	0	0	0	0	1
A0CZ	1	1	0	0	0
A0D0	0	0	1	0	0

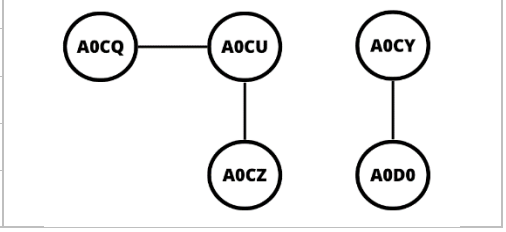


Figure 3. Example of an adjacency matrix and its respective graph representation.

2.2.1.2. Area under the curve (AUC)

The area under a curve between two points is found out by doing a definite integral between the two points. To find the area under the curve y = f(x) between x = a & x = b, integrate y = f(x) between the limits of a and b. This area can be calculated using integration with given limits. The area calculated is exemplified in *Figure 4*.

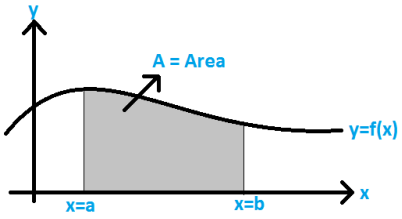


Figure 4. Representation of the integral calculated. Grey zone is the zone defined as Area under the Curve.

2.2.1.3. Cumulative Distribution Function (CDF)

The cumulative distribution function (CDF) calculates the cumulative probability for a given x-value. Use the CDF to determine the likelihood that a random observation taken from the population will be less than or equal to a particular value. The cumulative distribution function, F(t) ranges from 0 to 1.

The goal must be to have a CDF of 1 at t=0, that is, F(0) = 1; and AUC = 1. That would indicate that we have chosen the correct number of clusters.

2.2.1.4. Histogram of consensus matrix

A consensus matrix, unlike adjacency matrix, is a matrix containing continuous real values between 0 and 1. '0' values mean no interaction, '1' values mean interaction and the values between them are chances of interaction, where values close to 1 are more likely to indicate interaction and values close to 0 are more likely to indicate no interaction.

In a histogram of a consensus matrix, there is the representation of the values density in the matrix. In the 'x' edge there are the range of values appearing in the matrix, that is, interval values between 0 and 1. The density is represented in the edge 'y', that is, the frequency of appearances of each value in the matrix. For example, if we have ten '1' values and four '0' values the bar of x=1 will be higher than the bar of x=0.

When plotting a histogram of a consensus matrix entries, a perfect consensus would translate into two bins centered at 0 and 1. On the other hand, a histogram representing a bad consensus or noise pollution, would display more than two bins; actually, that is, it will display multiple bins between x=0 and x=1. There is an exemplification of these situations in *Figure 5*.

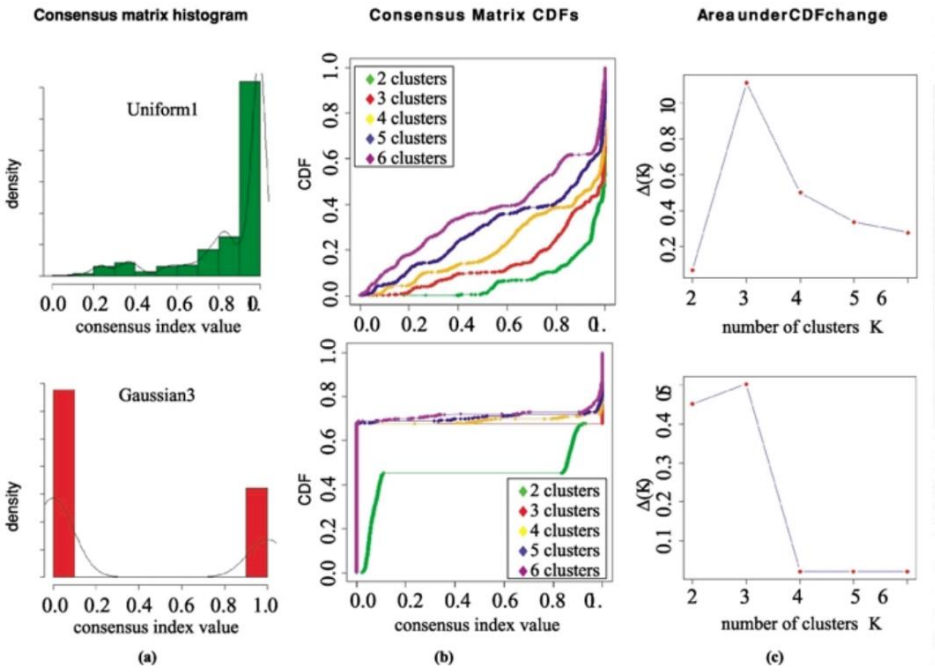


Figure 5. Representation of two (a) consensus matrix histograms, (b) two(Monti et al., 2003) consensus matrix CDFs and two (c) AUC plots. In the illustration, two examples are represented: Uniform1 and Gaussian3. Uniform1 dataset is generated from a uniform 600-dimensional hypercube. It is a dataset generated to evaluate the behaviour of the clustering methodology when applied to data known not to contain distinct sub-populations. On the other hand, Gaussian3 is a dataset that represents the union of three Gaussian distributions in a 600-dimensional space. Hence, is designed with 3 implicit sub-populations (Monti et al., 2003).

2.2.1.5 Bayesian Information Criterion

The Bayesian Information Criterion (BIC) is a statistical measure used in data analysis and statistical inference. BIC is primarily used in the context of regression models but can also be applied in other contexts. Its main goal is to select the best model among a set of alternative models fitted to a dataset. BIC is based on maximizing the likelihood of the model and penalizing more complex models to prevent overfitting.

BIC seeks a balance between model fit to the data (through likelihood) and model complexity (through the number of parameters). It penalizes models with a higher number of parameters, which helps avoid overfitting by favoring simpler models. In practice, the model with the lowest BIC value is chosen as it indicates a better trade-off between fit and complexity.

2.2.2 Integrative methods used

This section explains the methods used for the integration of the multi-omics data. Remark that the following methods were designed with disease subtyping purposes. Advantages and disadvantages of each of these methods are also discussed.

2.2.2.1 Similarity Network Fusion (SNF)

Similarity network fusion is a network-based approach to integrate multi-omics data sets using a network fusion method. First, SNF creates an individual network for each data type and then fuses these into a single similarity network using a network fusion approach. (Wang et al., 2014).

The algorithm of SNF starts with similarity matrices of each datatype. Hence, the number of matrices is equal to the number of datatypes to integrate. Each matrix can be understood as a similarity network where nodes are samples, and edges are weighted with pairwise similarities among samples (*Figure 6.a,b,c*). After the similarity matrices are initialized, SNF uses a non-linear method that iteratively updates every network with the purpose of making it more similar to the others constantly, at every iteration (*Figure 6.d*). After some iterations, a single similarity network is obtained, and data have been clustered (*Figure 6.e*).

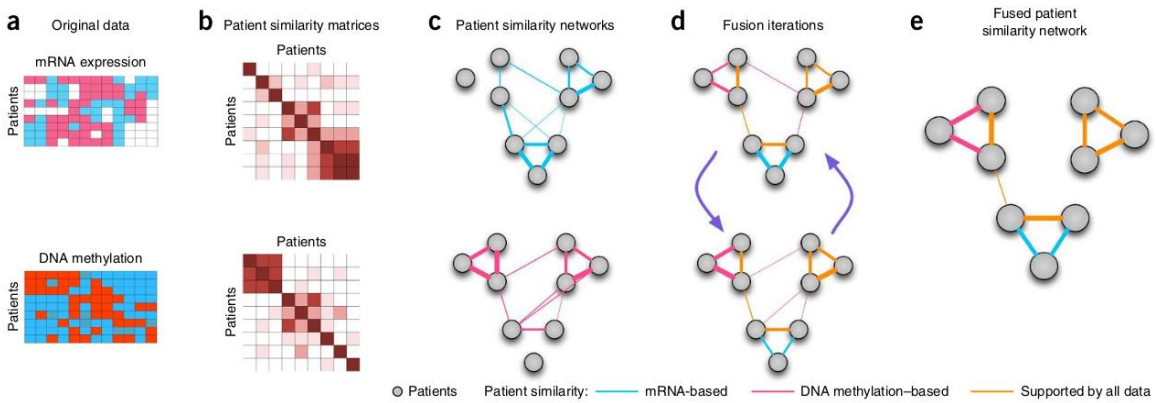


Figure 6. Illustrative example of SNF steps. (a) Example representation of mRNA expression and DNA methylation data sets for the same cohort of patients. (b) Patient-by-patient similarity matrices for each data type. (c) Patient-by-patient similarity networks, equivalent to the patient-by-patient data. Patients are represented by nodes and patients' pairwise similarities are represented by edges. (d) Network fusion by SNF iteratively updates each of the networks with information from the other networks, making them more similar with each step. (e) The iterative network fusion results in convergence to the final fused network. Edge color indicates which data type has contributed to the given similarity (Wang et al., 2014).

A schema of the algorithm can be found in *Figure 7*. Specifically, the process consists in the following:

(0) Suppose we have n samples (e.g, patients) and m measurements (e.g., mRNA gene expression). A patient similarity network is represented as a graph G = (V, E). The vertices V correspond to the patients $\{x_1, x_2, \square, x_n\}$ and the edges E are weighted by how similar the patients are.

(1) Edge weights are represented by an n × n similarity matrix W with $W_{(i,j)}$ indicating the similarity between patients x_i and $x_j$and are computed with a formula including euclidean distance and average distance between nodes and their neighbours.

(2) After this step, normalization of W matrices is performed. Normalized matrices are noted as P.

(3) Let N_i represent a set of x_i's neighbours including x_i in G. Given a graph, G, they use K nearest neighbours (KNN) to measure local affinity. With those affinity values, an affinity matrix is built for each dataset, noted as S.

(4) Then, SNF iteratively updates S similarity matrix using P normalized matrix. Every time an iteration is finished, a normalization is performed. This process is performed until new S matrices do not show differences within recent updates.

(5) Integration of the resulting similarity matrices of each datatype is performed

(6) The learned status matrix can be used, then, to subtype each sample. That is, labelling all the samples to the different sub-groups.

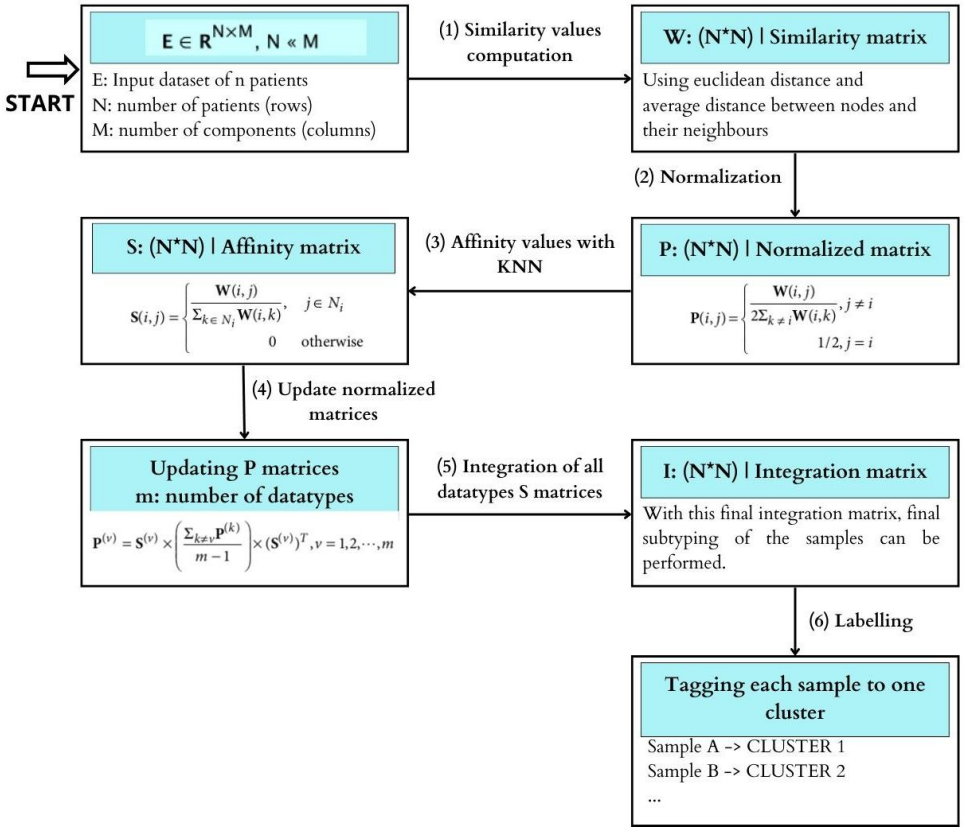


Figure 7. Schema of SNF algorithm.

The advantage of this method is that the weak connections (noise) disappear with iterations, whereas the strong connections are propagated till convergence. On the other hand, the main drawback is that SNF treats all available data types equally, which may lead to diluting signals and make a false 'fusion' due to different platforms or reliability levels of data types. Moreover, the unstable nature of kernel-based clustering makes the algorithm sensitive to small changes in molecular measurements or in its parameter settings.

SNF is available on CRAN repository. More information can be found at *https://cran.r-project.org/package=SNFtool*.

2.2.2.2 iCluster+

ICluster+ method can perform pattern discovery that integrates diverse data types: binary (somatic mutation), categorical (copy number gain, normal, loss), and continuous (gene expression) values.

A schema of the algorithm can be found in *Figure 8*. The detailed algorithm is written as follows:

(0) Let N be the number of patients and M the number of measurements/features for each patient. The input is a dataset (matrix).
(1) Given k (number of latent variables), they estimate the lasso penalty parameters that minimize a Bayesian Information Criterion (BIC) using a uniform design.
(2) To select the best k, the deviance ratio is computed. That is the ratio of the log-likelihood(fitted) - log-likelihood (null model) divided by the log-likelihood (full model)- log-likelihood (null model). The deviance ratio can be interpreted as the "% explained" variation.
(3) Best k is the one where the curve of "% explained" variation levels off.
(4) Subtype samples into different sub-groups using the k previously computed.

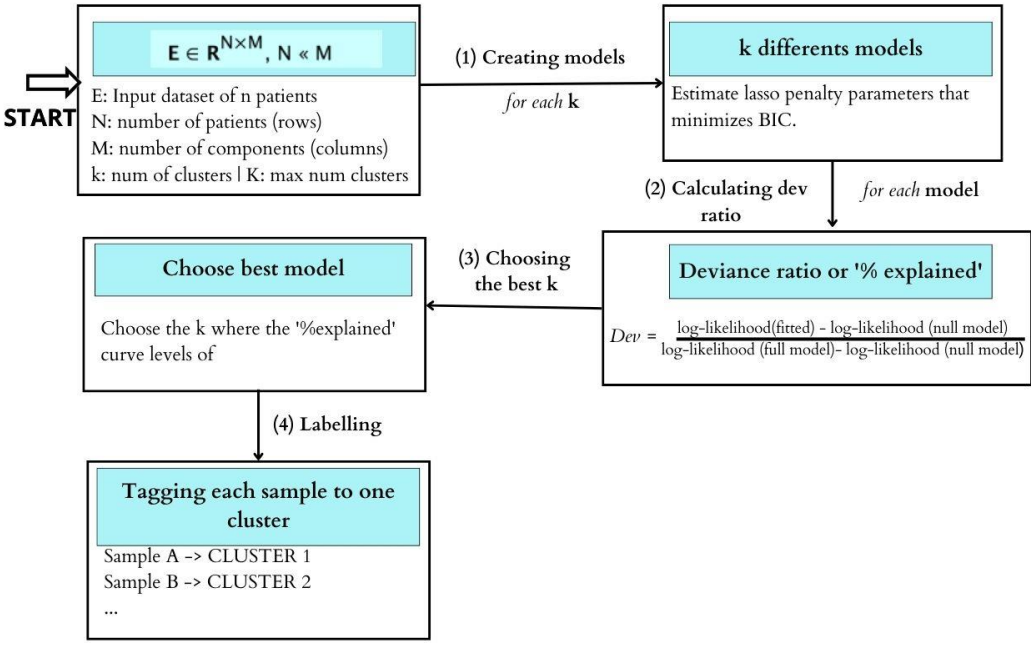


Figure 8. Schema of the iCluster+ algorithm.

Though powerful, these approaches are limited by their strong assumptions about the data and by the gene selection step used to reduce the computational complexity. It is very sensitive to its parameters.

iCluster is available on Bioconductor repository. More information can be found at *https://bioconductor.org/packages/release/bioc/html/iClusterPlus.html*

2.2.2.3 Consensus Clustering Plus (CC+)

CC+ provides quantitative and visual 'stability' evidence derived from repeated subsampling and clustering. It reports a consensus of these repetitions, which is robust relative to sampling variability (Wilkerson & Hayes, 2010). A schema of the algorithm can be found in *Figure 9*. The detailed algorithm is written as follows:

(0) Let N be the number of patients and M the number of measurements/features for each patient. The input is a dataset (matrix), noted as R.
(1) Let $k \in$ [2..K] where K is the maximum number of clusters to test. Now, let $h \in$ [1..H] where H is the number of resampling iterations. Then, for each k subsample a proportion of items and a proportion of features of R. Note S as these subsample matrices.
(2) Cluster subsampled matrices. Let H be these clustered matrices. User can choose within three algorithms: agglomerative hierarchical clustering, k-means or a custom algorithm. Remind that, in this point, H connectivity – clustered - matrices for each k value have been generated.
(3) Calculate a consensus matrix noted as C, result of properly normalized sum of the connectivity matrices of all the perturbed datasets of that k value. In C, if two samples have been mostly grouped together, they will have a pair-wise value close to 1. Otherwise, if they have not been grouped together, this value will be close to 0.
(4) Calculate consensus matrix histogram, cumulative distribution function (CDF) and the area under the curve (AUC) of each clustering.
(5) The optimal number of clusters k is chosen where the area under the curve levels off and $\Delta(\hat{k})$ approaches zero. Furthermore, this assumption must be supported with a CDF ≈ 1 and with heterogeneous histogram (almost binarized plot with major presence of '0' and '1'). That is, the clustering exhibiting the least disruption when noise was added to original data.
(6) Data clustering based on the best k. Labelling samples to sub-groups.

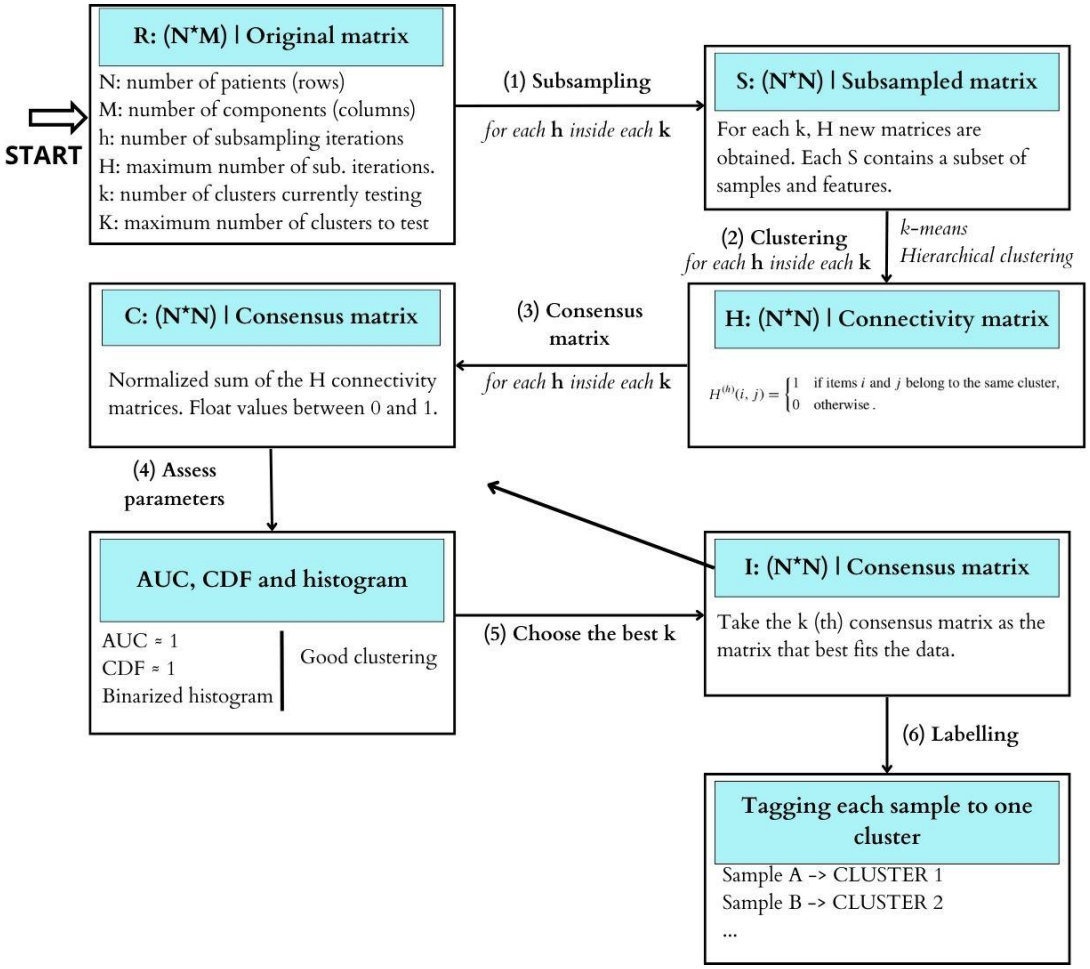


Figure 9. Schema of the CCPlus algorithm.

CC is not designed to integrate multiple data types. In that case, instead of performing a full integration, each data type will be analysed separately. After that, all data type results will be concatenated to perform the "integration".

CC is available on Bioconductor repository. More information can be found at *https://bioconductor.org/packages/release/bioc/html/ConsensusClusterPlus.html*.

2.2.2.4 Perturbation clustering for data INtegration and disease Subtyping Plus(PINS+)

To identify subtypes, PINS identifies how often the patients are arranged together in 1) a single cluster; 2) when the data are perturbed; 3) when using different types of omics data, and 4) when a different clustering technique is used. Patients who are strongly connected in all the scenarios are clustered together into a subtype. A schema of the algorithm can be found in *Figure 10*. The detailed algorithm is written as follows:

(0) Let N be the number of patients and M the number of measurements/features for each patient. The input is a dataset (matrix) $E \in R^{NxM}$.
(1) PINS makes a partition of the patients using all possible number of clusters k ∈ [2..K] After that, (K-1) partitionings are obtained, one for each value of k.
(2) For each partitioning, it builds the pair-wise connectivity matrix. These matrices can only have '0' or '1' values. If two patients are connected by a '1', they are clustered together. Otherwise, a '0' means they are not clustered together.
(3) Then, it adds gaussian noise to the original datatype matrices E (from step 0), in a process they called "perturbation". The perturbed matrices are noted like H_k, one for each k value. To avoid not perturbing the data sufficiently, but also to try to not increment the differences between subtypes, the variance of the perturbation noise added is equal to the median variance of the data.
(4) Cluster the perturbed matrices $H_k$with k-means algorithm varying values of k ∈ [2..K]. That is, for each partitioning, they obtain k new matrices by applying k-means. This is in total, k^2 matrices. Let Q be those clustered matrices. For each H_k, they obtain $Q_k^{(1)}$, $Q_k^{(2)}$, . . . , $Q_k^{(K)}$, where each k ∈ [2..K]
(5) Build the $G_k^{k'}$connectivity matrix for each $Q_k^{k'}$perturbed and clustered matrix obtained in the last step.
(6) For each k, calculate the A_k perturbed connectivity matrix by averaging the $G_k^{k'}$connectivity matrices obtained in the last step.
(7) Calculate the discrepancy between original and perturbed data using a difference matrixD_k, as follows: $D_k = |C_k - A_k|$
(8) Calculate cumulative distribution function (CDF) of each clustering.
(9) Calculate the area under the curve (AUC). If the difference matrix contains almost '0' values, it means the perturbation did not change clustering results. Hence, CDF ≈ 1 and AUC ≈ 1.
(10) The optimal k number of clusters is where the AUC is maximized. That is, the clustering exhibiting the least disruption when noise was added to original data.

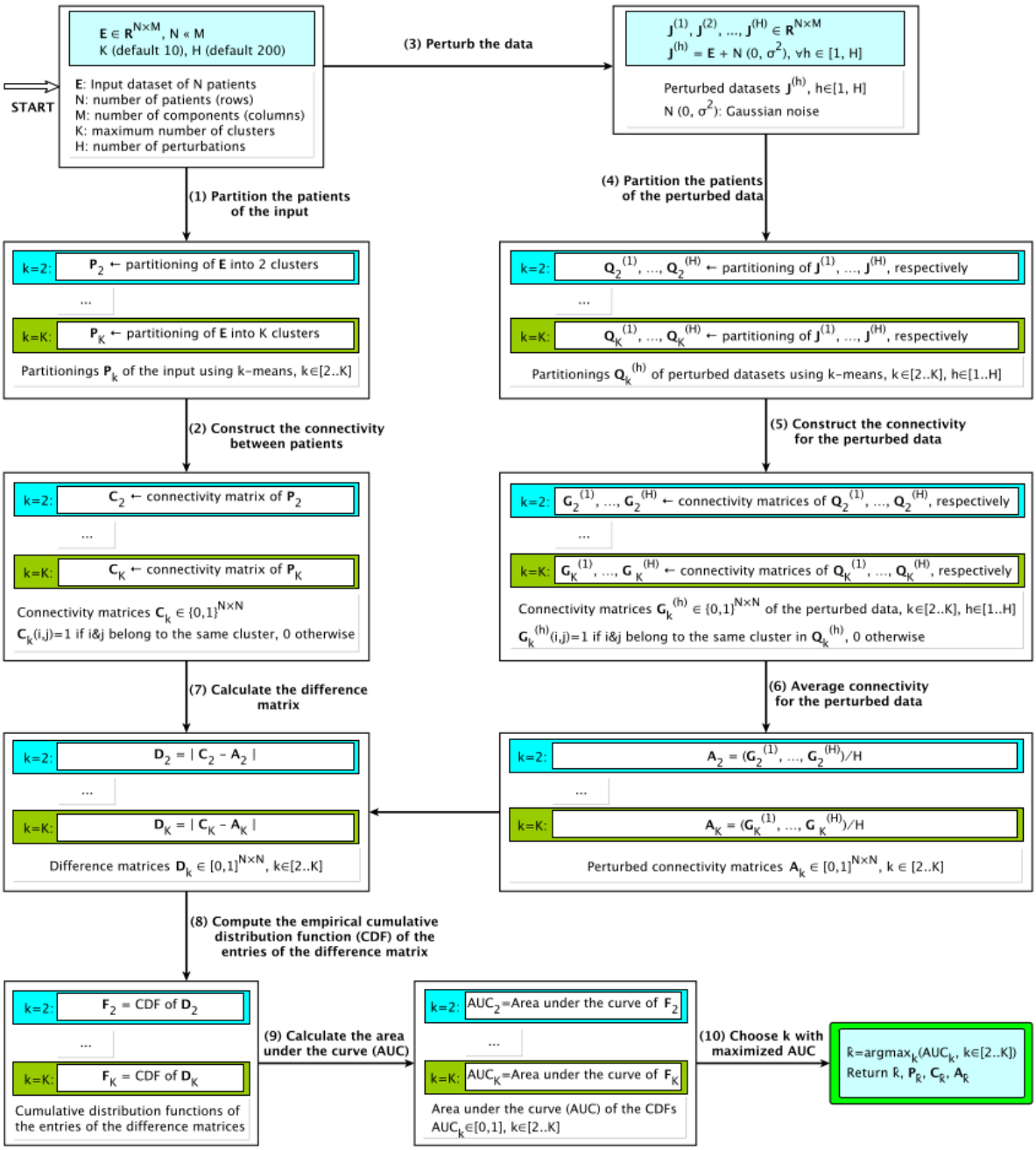


Figure 10. Schema of the PINS algorithm (Nguyen et al., 2017).

PINS is able to extract subtypes with significant survival differences even after data integration, unlike all other existing approaches. On the other hand, the main limitation of PINS is that it treats all data types equally in determining subtypes. This may not always be appropriate. For instance, for GBM, the results show that methylation plays a major role in determining distinct subtypes. Additionally, in terms of time complexity, PINS needs a significantly longer time than CC and SNF to perform an analysis on large datasets. One of the reasons is that they run k-means multiple times to make sure that the results are stable and reproducible.

PINS is available on CRAN repository. More information can be found at *https://cran.r-project.org/package=pins*.

2.2.2.5 Walktrap method

The Walktrap method is a community detection algorithm for graphs that is based on the concept of random walks to identify community structures in a graph (see *Figure 11*). This method is available in the 'igraph' package in R and is used to find natural groupings of nodes in a network.

Here is an explanation of the Walktrap method step by step:

1. *Random Walk* · The algorithm begins by performing random walks on the graph. A random walk is a process in which you start at a node and randomly choose one of its neighbors as the next step, repeating this process at each subsequent step. The goal of the random walks is to simulate how the graph would be explored starting from an initial point.
2. *Similarity Between Random Walks* · During the random walks, similarities between walks originating from different nodes are recorded. Similarity is measured in terms of how much two random walks have in common in their paths. If two random walks visit similar nodes at similar steps, they are considered similar.
3. *Hierarchical Clustering* · Based on the similarities between random walks, the Walktrap algorithm uses hierarchical clustering techniques (e.g., agglomerative linkage) to group nodes that tend to be close in terms of similarity into larger clusters.
4. *Community Detection* · The algorithm stops at some level of the hierarchy, usually when a specific resolution value is reached or when a clear community structure is detected. Nodes grouped together at this point are considered members of the same community.
5. *Community Detection Results* · The result of the algorithm is an assignment of nodes to communities. Each node belongs to a single community, and the goal is generally to maximize similarity between random walks within the same community and minimize similarity between random walks in different communities.

The Walktrap algorithm is useful for detecting communities in graphs where random walks tend to stay within well-defined communities due to the network's structure. It's important to note that the algorithm can be tuned using parameters such as the number of steps in random walks to adapt to different types of graphs and community detection resolutions.

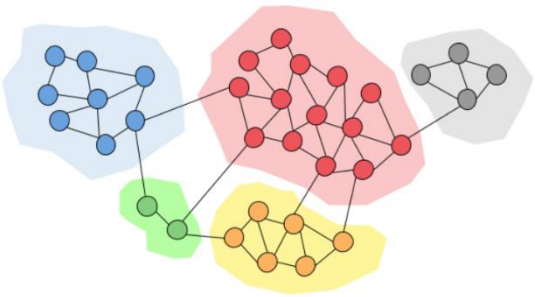


Figure 11. Walktrap algorithm action.

2.3 Implementation

In the following section, the most relevant aspects of the implementation are described as well as the justifications of each decision.

2.3.1. R programming language

R is the chosen language to perform the integrative analysis in this context. R is one of the most used programming languages in integrative analysis due to various reasons:

- Huge number of libraries · R has a big set of libraries that can deal with biological and omics data.
- Advanced data visualization · It has powerful tools with data visualization capabilities, like *ggplot2*. It allows to create personalized plots that can help in the results exposure.
- Statistical tools · The set of advanced statistical tools in R can help in the process of finding patterns, relationships and significance differences among data.
- Data integration · Different libraries in R make the process of data loading into the environment easier. Moreover, most of them can load data coming from different sources like "txt", "csv" (comma-separated values) and "json".
- Interoperability · R can be integrated with more languages like Python (*reticulate* package), C++ (*Rcpp* package), Java (*rJava* package) or SQL (*RMySQL* package). This allows the user to use the power of each language at any time.
- Documentation and support · There are a lot of tutorials, courses and documentation related to R packages that can help to the programming learning. Furthermore, R has a big user community that have created packages and that discuss about issues in online forums. That can help when you are locked in a development step.

All our methods introduced previously – SNF, iCluster+, PINS+, CC+ - are R packages that are freely available on either CRAN or Bioconductor repository.

2.3.2. Detailed implementation

In the following section, find a detailed implementation of the module (B) Integration (see *Figure 2*).

The input dataset of this module is a set of either comma-separated-values files, or R objects. Each file or object contains the values of a specific datatype corresponding to each sample. In this case, the chosen datatypes have been mRNA expression, miRNA expression and DNA methylation values. Hence, we have three different files or objects as inputs of the methods.

This module can be divided into 4 different steps: (1) Pre-processing, (2) Selection of the k value for each method, (3) Samples labelling for each method and finally (4) the integrative method gathering all the individual results. An explanation of each step is described in the next sections.

2.3.2.1 Pre-processing

A normalization of the values is performed depending on the method. According to the manuals and manuscripts of the methods.

For SNF, standard normalization is performed. The goal of this step is to center all the values to make comparison between them. That is reached making the values have a zero mean and a standard deviation of one.

For iClusterPlus, the Median Absolute Deviation (MAD) is computed for all the features of each datatype. Then, it keeps only the features with highest MAD; that is, the most relevant features. A part of that, no normalization is performed.

For PINS+, the pre-processing step is skipped.

For CC+, the pre-processing step consists of center all the features values of each sample around the median of features values. Furthermore, a selection of the x most relevant features is performed. After dealing with all the subtypes, standard normalization is performed before the integration.

2.3.2.2 Selecting k

For SNF, the number of clusters is calculated using the function *estimateNumberOfClustersGivenGraph*. This method is returning different cluster number values in order of preference. In the workflow, the first choice is always selected because, in principle, it means the best value choice.

For iClusterPlus, a deviance ratio is first computed in order to choose the number of cluster that best fits the model, as explained previously in Methodology. The *k number of clusters* value is selected where this ratio levels off.

For PINS+, there is not the need to select a *k* number of cluster value. The number of clusters is automatically selected by *SubtypingOmicsData* function. This function is also labelling the samples into different subgroups.

For CC+, a k number of clusters value for each datatype is selected according to the AUC plots. The k that best fits each datatype where the area under the curve levels off and $\Delta(k)$ approaches zero. After selecting k values for each datatype, a k value for the integration of the datatypes is selected following the same strategy: choose k according where the AUC levels off and $\Delta(k)$ approaches zero,

2.3.2.3. Labelling

Each method is labelling then the samples, depending on if they belong to the same cluster or not. If two samples belong to the same cluster, they will be labelled with the same cluster number.

Once all the algorithms have returned the cluster predictions (see *Table 2*, *Table 3*), the integrative method of all those results can begin.

2.3.2.4. Integrative method

(a) 1st step

The goal of the first step is to build a consensus graph depending on the clustering results of each method. Our integrative method does not need to calculate the k number of clusters with which the data can be divided. It only receives the labels of the samples returned by each method. With that information, it checks which samples are grouped together the most by building adjacency matrices. The assumption is: if the samples are grouped together in most cases, then they belong to the same cluster. That is, if two samples have been grouped together more 3 or more times (taking all the 4 algorithms labellings) they can be considered as they belong to the same cluster, as shown in *Figure 12*. With all the samples classified into different clusters, an initial graph is built.

(b) 2nd step

In the second step, Walktrap algorithm is applied to the graph obtained in the previous phase. This algorithm can provide crucial insights into the natural grouping of nodes with similarities in their connections. By classifying nodes based on their membership in these communities, it becomes possible to capture latent patterns and relationships that may not be apparent in traditional classification. The goal is to remove weak iterations that are present in the initial graph as well as to intensify the strong iterations. This may lead to improved classification accuracy by considering the influence of the network structure on node labelling and, thus, enhancing the model's ability to capture intrinsic cohesion within interconnected data.

Table 2. Sub-groups labelling returned by each clustering method. Data coming from 5 samples of BRCA dataset.

Sample	**SNF**	**ICluster+**	**PINSPlus**	**CC**
TCGA-A2-A0CQ	1	5	2	2
TCGA-A2-A0CU	1	2	2	2
TCGA-A2-A0CY	2	5	1	1
TCGA-A2-A0CZ	1	2	2	1
TCGA-A2-A0D0	3	2	1	1

Table 3. Adjacency matrices returned by each method. '1' means connection between samples inside the same cluster; then, they belong to the same cluster. '0' means no connection between those samples, that is they belong to different clusters. Data coming from 5 samples of BRCA project.

SNF	*A0CQ*	*A0CU*	*A0CY*	*A0CZ*	*A0D0*
A0CQ	0	1	0	1	0
A0CU	1	0	0	1	0
A0CY	0	0	0	0	0
A0CZ	1	1	0	0	0
A0D0	0	0	0	0	0

IC+	*A0CQ*	*A0CU*	*A0CY*	*A0CZ*	*A0D0*
A0CQ	0	0	1	0	0
A0CU	0	0	0	1	1
A0CY	1	0	0	0	0
A0CZ	0	1	0	0	1
A0D0	0	1	0	1	0

PINS	*A0CQ*	*A0CU*	*A0CY*	*A0CZ*	*A0D0*
A0CQ	0	1	0	1	0
A0CU	1	0	0	1	0
A0CY	0	0	0	0	1
A0CZ	1	1	0	0	0
A0D0	0	0	1	0	0

CC	*A0CQ*	*A0CU*	*A0CY*	*A0CZ*	*A0D0*
A0CQ	0	1	0	0	0
A0CU	1	0	0	0	0
A0CY	0	0	0	1	1
A0CZ	0	0	1	0	1
A0D0	0	0	1	1	0

A

	A0CQ	*A0CU*	*A0CY*	*A0CZ*	*A0D0*
A0CQ	0	1	0	0/1	0
A0CU	1	0	0	1	0
A0CY	0	0	0	0	0/1
A0CZ	0/1	1	0	0	0/1
A0D0	0	0	0/1	0/1	0

B

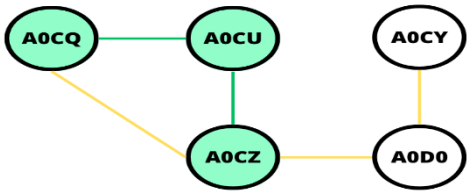


Figure 12. (A) Adjacency matrix of the integrative method. Green values (1) are connections that are supported, at least, by two methods. Yellow values (0/1) are connections that are supported by only two methods. That is, only half of the methods support that interaction, meaning it is a weak connection. (B) Visualization of the graph corresponding to the adjacency matrix. The nodes (represented in circles) represent the samples and the edges (represented in lines) represent the connection between them. Two nodes connected with an edge belong to the same cluster. Data coming from 5 samples of BRCA project.

2.4. Metrics and evaluation methods

The effectiveness of these clustering techniques needs to be rigorously assessed and validated to ensure their utility in real-world applications. Here, a wide array of metrics and methodologies are used specifically designed to quantify the quality and performance of clustering solutions. These metrics not only help analysts determine the validity of the clusters formed but also assist in the comparison of different clustering algorithms.

2.4.1. Rand Index and Adjusted Rand Index

With ground truth data, there is the possibility to use Rand Index (RI) (Rand, 1971) or the Adjusted Rand Index (ARI) (Hubert & Arabie, 1985). Since the true disease subtypes are known in the datasets used for test, the Rand Index (RI) (Rand, 1971) and Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) can be used to assess the performance of the resulted subtypes.

Specifically, RI measures the agreement between a given clustering and the ground truth. Deeping into the parameter, RI = (a + b)/ ( ) where *a* is the number of pairs that belong to the same true subtype and are clustered together, *b* is the number of pairs that belong to different true subtypes but they are not grouped together, and () is the number of possible pairs that can be formed from the N samples. Then, RI is the fraction of pairs that are grouped in the same way – they can be together or not - in the two partitions compared. For example, a RI = 0.7 means 70% of pairs are grouped in the same way.

The Adjusted Rand Index (ARI) is the corrected version of the Rand Index. The ARI takes values from -1 to 1, where 1 indicates perfect agreement between two clusters, 0 indicates a random subtyping and -1 indicates that the two clusters are completely different. Hence, ARI values close to 1 are the ones that can be expected from a good clustering performance.

As commented, this metric can only be used while ground truth data exists. On the present scenario, with unsupervised data, there is the must to rely in another assessment resources, like the survival analysis.

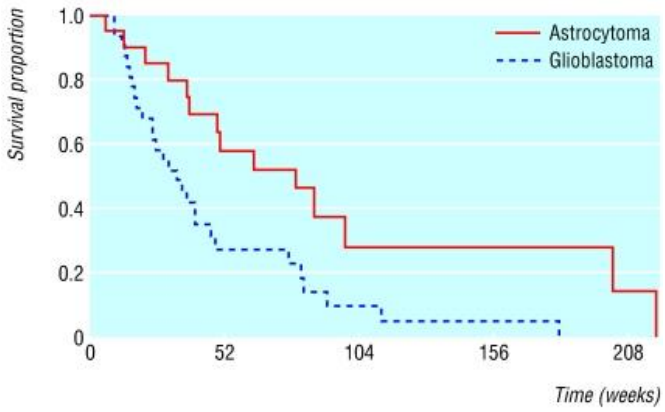
2.4.2. Survival Analysis

Survival Analysis is a tool used to assess the time until an event occurs. This “event” could be any number of alternatives. In the present study case, the event that wants to be studied is the death. In other words, is follow-up of the time from the diagnosis of a disease to death.

In the Cox Regression Model, the dependent variable is the risk. The risk is defined as the probability of dying (or the specific event in question) given that patients have survived up to a given point in time, or the risk for death at that specific moment. (Bewick et al., 2004)This can also be thought of as the instantaneous risk of experiencing the event of interest at a certain time point. (Schober & Vetter, 2021).

The log-rank test in Cox Regression Model is used to test whether the difference between survival times between two groups is statistically different or not. The test returns a p-value. Usually, we determine that two groups are statistically different when Cox p-value < 0.05. The lower the P value is, the less likely it is that such differential survival was observed by chance, i.e., the more significantly different the survival profiles is between subtypes.

In order to exemplify this explanation, an example is attached in *Figure* 13 to clarify and understand how the survival analysis is working (Bland & Altman, 2004).



Astrocytoma	6, 13, 21, 30, 31, 37, 38, 47, 49, 50, 63, 79, 80, 82, 82, 86
Glioblastoma	10, 10, 12, 13, 14, 15, 16, 17, 18, 20, 24, 24, 25, 28, 30, 33, 34, 35, 37, 40, 40, 40, 46, 48, 70, 76, 81, 82, 91, 112, 181

Figure 13. Survival curves for both women with Astrocytoma and Glioblastoma by diagnosis. A table with the survival weeks of the patients belonging to each group is also attached (Bland & Altman, 2004).

The first death was in week 6 - astrocytoma group - when one patient died. At the beginning of that week, there were 51 subjects alive, so the risk of death was 1/51. There were 20 patients in group 1 – supposing the null hypothesis was true - the expected number of deaths in group 1 was $20 \times 1/51 = 0.39$. At the same time, the expected number of deaths in glioblastoma group was $31 \times 1/51 = 0.61$. In week 10 more events took place, when there were two deaths. There were now 19 and 31 patients alive in the two groups, one having died in week 6, so the probability of death in week 10 was 2/50. The expected numbers of deaths were $19 \times 2/50 = 0.76$ and $31 \times 2/50 = 1.24$ respectively.

Taking the calculations for each time of death, expected deaths were 22.48 in group 1 and 19.52 in group 2, and the observed numbers of deaths were 14 and 28. We can now use a χ^2 test of the null hypothesis. The test statistic is the sum of $(O - E)^2/E$ for each group, where O and E are the totals of the observed and expected events. Here $(14 - 22.48)^2 / 22.48 + (28 - 19.52)^2 / 19.52 = 6.88$. The degrees of freedom are the number of groups minus one, i.e. $2 - 1 = 1$. From a table of the χ^2 distribution we get $P < 0.01$, so that the difference between the groups is statistically significant.

The feeling is that the most significant assessment is provided by the survival analysis of the subgroups discovered. Actually, there is the aim to discover the subtypes that have the potential to make a difference in the clinical practice and, from that perspective, an approach that can distinguish between patients with the more and less aggressive disease subtypes is more interesting.

3. Experimentation and Results

In this chapter, the experimentation and validation of the approach made in this project is presented.

Table 4 is displaying the firsts results obtained once the workflow has been executed. The results display the number of clusters chosen by each method for each condition. The visualization is divided by projects and by the number of datatypes that are intervening in each case, using all the sample of the dataset (#patients). Survival analysis cox p-value is chosen as a metric because is considered as the best to assess the results, as explained before. Situations where cox-p values < 0.05 display clustering statistically significant and can be accepted as potentially possible classifications.

Table 4. Subtyping results of PINS, CC, SNF, iClusterPlus, and the integrative method developed for the 6 cancer diseases. For each disease, the first row displays the results using mRNA, DNA methylation data, and miRNA while the other three rows display the results using two types of data. Since iClusterPlus is unable to subtype miRNA data for KIRC, LAML, LUSC, BRCA, and COAD, the results for any combination with miRNA is shown as NA. Asterisk (*) values were obtained using only 3 methods (SNF, PINS+ and iClusterPlus; in those cases, the samples were grouped if, at least, two methods classified them together. Cox-p values < 0.05 are statistically significant and are coloured with green colour. The rest of situations are not coloured.

TCGA DATASET			**SNF**		**ICLUSTER`+**		**PINS+**		**CC+**		**INTEGR.**	
NAME	#Patients	Datatype	k	Cox p	k	Cox p	k	Cox p	k	Cox p	k	Cox p
BRCA	172	mRNA, methyl	2	0.3793	10	0.5024	7	0.4664	7	0.4728	12	0.4600
		mRNA, miRNA	2	0.5793	NA	NA	8	0.7317	8	0.3756	3	0.9600*
		methyl, miRNA	2	0.9690	NA	NA	8	0.0193	7	0.9226	5	0.1100*
		All	2	0.3983	10	0.5024	3	0.3215	7	0.5807	5	0.0019
GBM	273	mRNA, methyl	5	0.0030	10	0.0003	5	0.0004	7	0.0183	7	0.0054
		mRNA, miRNA	2	0.5586	7	0.2945	4	0.0322	7	0.0615	6	0.1400
		methyl, miRNA	4	0.0296	5	<0.0001	8	0.2804	6	0.0288	10	0.0980
		All	4	0.0208	5	0.0616	2	0.4388	7	0.0396	5	<0.0001
KIRC	124	mRNA, methyl	2	0.8772	6	0.0795	4	0.0022	6	0.2108	8	0.6300
		mRNA, miRNA	2	0.8772	NA	NA	4	0.0212	7	0.0159	3	0.0025*
		methyl, miRNA	2	0.3568	NA	NA	4	0.0576	9	0.6333	8	0.0009*
		All	2	0.6913	6	0.0795	3	<0.0001	6	0.2051	9	0.1400
LAML	164	mRNA, methyl	5	0.0003	5	0.4951	7	0.0475	6	0.1107	12	0.5100
		mRNA, miRNA	3	0.0304	NA	NA	5	0.0063	5	0.0135	5	0.8200*
		methyl, miRNA	4	0.0047	NA	NA	3	0.1443	7	0.0179	5	0.7000*
		All	4	0.0017	5	0.4951	3	0.1099	8	0.0568	8	0.0210
LUSC	110	mRNA, methyl	2	0.7021	4	0.3887	9	0.2129	5	0.5493	10	0.0003
		mRNA, miRNA	2	0.0415	NA	NA	3	0.1246	6	0.4347	3	0.0220*
		methyl, miRNA	2	0.9418	NA	NA	4	0.0068	8	0.4253	4	0.0580*
		All	3	0.0870	4	0.3887	4	0.0083	6	0.7941	8	0.0110
COAD	146	mRNA, methyl	2	0.7837	10	0.2361	8	0.0864	5	0.2254	11	0.1000
		mRNA, miRNA	3	0.2131	NA	NA	3	0.6607	5	0.7471	4	0.0038*
		methyl, miRNA	2	0.1078	NA	NA	3	0.6607	8	0.3484	2	0.5700*
		All	2	0.9032	10	0.2361	3	0.6607	5	0.2254	10	0.0100

The initial results obtained are highly promising and, in the majority of cases, they outperform previous approaches. The cox-p values are below 0.05 – for all datatypes integration - in all the cases, but for the KIRC dataset. This is an outcome not seen before in other approaches and seems to be a big improve while clustering these six cancer datasets.

In the present case, the integrative method developed is finding more clusters than the other approach. This fact is suggesting that the previous methods were not splitting the data enough. This could mean that those algorithms were grouping samples together when, in reality, they were part of different groups.

In other words, the outcome suggests that the workflow is removing weak iterations between samples while making stronger 'true' iterations.

4. Discussion

The workflow developed is still primitive and seems to be very sensible to some parameters. It may be relevant to study the main parameters that affect to specific parts of the workflow, or even, the ones that affect to the whole workflow. Finding the parameters that change drastically the outcome is interesting to find limitation steps inside the process as well as stablish which parameters are offering best performances in each case.

The titles of the following sub-sections display the parameters that are modified in each specific case.

4.1. Number of datatypes

Sometimes, one datatype is not enough differentiating and is not apporting useful information when clustering. Furthermore, it can perturb the results that the other datatypes may offer, by adding noise. In order to check this fact, a study of this phenomenon is performed.

As displayed in *Table 4*, the workflow was run using either two or three datatypes. Moreover, all the combinations among mRNA, miRNA and DNA-methylation datatypes were used to perform it, but for iClusterPlus.

The results suggest that the clustering algorithm using only two datatypes do not over-perform the process using three datatypes. Observing the results, it may be logical to conclude that the bigger number of datatypes, the more comprehensive view of the data. Hence, clustering algorithms will become better when integrating more number of datatypes.

4.2. Dual and Triple Grouping

In previous sections it is told that if two samples have been grouped together more 3 (triple grouping) or more times (taking all the 4 algorithms labellings) they can be considered as they belong to the same cluster.

However, and once seen last results, it could be interesting to make a comparison among perform the integrative method changing the previously described assumption. Hence, an study is made considering the following assumption: if, at least, two samples are grouped together by the algorithms (dual grouping), they can be considered as belonging to the same cluster.

Quadruple grouping makes no sense, because it is almost impossible - in most cases - that all the methods can return the same cluster composition. This will translate into a lot of clusters (probably <20) that will conduct to good cox p-value metrics. However, this good cox p-value will be caused by samples that are alone in their clusters, that is, clusters of only one node. When individual clusters are appearing, cox-p value are improving consequently. For example, if we have a set of 50 samples and the algorithm is clustering the samples into 50 different clusters, the cox-p value will be the best. However, this is not a real representation of the dataset. Hence, the main goal is to have the least individual clusters together with the minimum number of clusters possible minimizing the cox-p value.

Deeping into this comparison, the parameters of the algorithm are changed to check if the outcome is varying. The results are displayed in *Table 5*:

Table 5. Subtyping results of the integrative method developed for the 6 cancer diseases using dual or triple grouping. Cox-p values < 0.05 are statistically significant and are coloured with green colour. The rest of situations are not coloured.

TCGA DATASET			**DUAL GROUPING**		**TRIPLE GROUPING**	
NAME	#Patients	Datatype	k	Cox p	k	Cox p
BRCA	172	All	3	0.9900	5	0.0019
GBM	273	All	4	0.0010	5	<0.0001
KIRC	124	All	6	0.0440	9	0.1400
LAML	164	All	3	0.6600	8	0.0210
LUSC	110	All	3	0.0089	8	0.0110
COAD	146	All	5	0.5500	10	0.0100

As *Table 5* is showing, triple grouping is over-performing dual grouping approach in most of the cases. However, there is an exception for KIRC disease, where dual grouping is performing better than triple.

According to the premise, it makes sense that the triple grouping works better than the double one. Triple grouping means a most consensus approach and stronger relationships between samples grouped together. That is because at least 3 methods have labelled them as they belong to the same cluster. On the other hand, double grouping means weaker interactions between samples and will translate, in most cases, into a worse performance.

4.3. Consensus matrix

Another option is to use the consensus matrices, that is, using the matrices that are relating samples with continuous values between 0 and 1, instead of using discrete values (0 and 1) like normally. This fact may conduct to a higher detail degree of the interactions between the samples and can give a deeper vision of the groups hidden in the dataset. *Table 6* clarifies this approach.

Hence, the consensus matrix of each method must be obtained. After all the consensus matrices have been obtained, an addition operation of all the matrices is performed to obtain an integrative consensus matrix. That matrix will have continuous values between 0 and 4. Once this matrix is obtained, it is converted to an adjacency matrix using a specific threshold. Values below the threshold will turn into 0. On the other hand, values above the threshold will turn into 1.

Table 6. Adjacency and consensus matrix concepts. While in adjacency matrix, there are discrete values, in the consensus there are continuous values. Exemplification of the conversion from integrative consensus matrix to adjacency matrix using a threshold=2.5. Data coming from GBM project.

DISCRETE ADJACENCY	*0010*	*0011*	*0014*	*0021*	*0024*
0010	0	1	0	0	1
0011	1	0	0	1	1
0014	0	0	0	0	0
0021	0	1	0	0	1
0024	1	1	0	1	0

THR=2.5

CONTINUOUS CONSENSUS	*0010*	*0011*	*0014*	*0021*	*0024*
0010	0.00	2.67	1.67	2.33	2.67
0011	2.67	0.00	1.33	2.67	3.01
0014	1.67	1.33	0.00	1.00	1.32
0021	2.33	2.67	1.00	0.00	2.68
0024	2.67	3.01	1.32	2.68	0.00

However, it is only possible to retrieve the consensus matrix of three of the methods (SNF, PINS+ and CC+). For the other method, iCluster+, there is no way to return the consensus matrix, but the adjacency matrix. Hence, the only solution is to perform the addition operation between three consensus matrices (continuous values) and one adjacency matrix (discrete values). A study of the importance of the threshold value is made in *Table 7*. The threshold value rank starts at 1.00 and ends at 2.00.

Table 7. Study of the outcome of the consensus matrix approach using different values of threshold (from 1.0 to 2.0 by 0.25 intervals). The outcome is assessed with the number of clusters obtained and its respective cox p-value. Cox-p values < 0.05 are statistically significant and are coloured with green colour. The rest of situations are not coloured. All the datasets are run.

TCGA DATASET			**THR=1.0**		**THR=1.25**		**#THR=1.50**		**#THR=1.75**		**#THR=2.00**		**DISCRETE VALUES MATRIX**	
NAME	#Patients	Datatype	k	Cox p	k	Cox p	k	Cox p	k	Cox p	k	Cox p	k	Cox p
BRCA	172	All	4	0.300	6	0.1500	5	0.2400	7	0.4600	9	0.6600	5	0.0019
GBM	273	All	4	0.0011	4	0.0290	5	0.0840	7	0.0150	8	0.1900	5	<0.0001
KIRC	124	All	6	<0.0001	6	0.0470	5	0.1100	5	0.1100	6	0.1700	9	0.1400
LAML	164	All	5	0.5700	5	0.5100	5	0.5100	5	0.5300	10	0.7300	8	0.0210
LUSC	110	All	4	0.0081	4	0.1700	4	0.1700	4	0.1700	6	0.4000	8	0.0110
COAD	146	All	5	0.0740	5	0.0310	12	0.2800	11	0.5900	14	0.6300	10	0.0100

Consider the first approach the one using adjacency matrix since the beginning; find results in the column 'Discrete Values Matrix' in *Table 7*. In general lines, the consensus matrix approach is not performing better than the first approach. However, in the KIRC disease, the results obtained by this new approach are surprisingly good. With a cox p-value < 0.0001, it can classify correctly the data for the first time. The first approach was not able to classify KIRC data in a good manner.

4.4. Number of steps in Walktrap algorithm

The number of random steps is one of the main parameters that affects the Walktrap algorithm for community detection in graphs. A higher number of steps tends to result in finer or lower-resolution community detection, meaning the graph will be divided into smaller and more detailed communities. Conversely, a lower number of steps tends to group more vertices into larger communities, leading to lower resolution.

By adjusting this parameter, different results can be obtained. Sometimes, it's necessary to fine-tune this parameter to achieve the desired level of community detail. Find a study of this parameter applied to the manuscript data in *Table 8*:

Table 8. Subtyping results of the integrative method developed for the 6 cancer diseases using different number of steps in Walktrap algorithm. Cox-p values < 0.05 are statistically significant and are coloured with green colour. The rest of situations are not coloured.

TCGA DATASET			**#STEPS=3**		**#STEPS=4**		**#STEPS=5**	
NAME	#Patients	Datatype	k	Cox p	k	Cox p	k	Cox p
BRCA	172	All	3	0.0021	5	0.0019	5	0.0038
GBM	273	All	5	<0.0001	5	<0.0001	5	<0.0001
KIRC	124	All	9	0.1500	9	0.1400	6	0.0730
LAML	164	All	6	0.5600	8	0.0210	8	0.0310
LUSC	110	All	7	0.0110	8	0.0110	7	0.0083
COAD	146	All	9	0.0093	10	0.0100	11	0.0054

Table 8 suggests that maintaining the number of steps between 3 and 5 will translate in similar outcomes in most cases. But there is one exception, for LAML disease. In that case, a number of steps lower than 3 is classifying wrongly the data, leading to a high cox-p value.

4.5. Test set

In this last section of the discussion, all the tests are gathered in a unique table. In *Table 9*, all the approaches previously commented are showed together with their outcome. We assess the results depending on the number of cox p-values below than 0.05 for each situation. Data were gathered from previous sections and Tables.

Table 9. Test set. Green colour remark best value for each parameter.

Number of datatypes	**Two**		**Three**
	3 out of 6		5 out of 6
Grouping	**Dual**		**Triple**
	3 out of 6		5 out of 6
Type of matrix	**Consensus**		**Adjacency**
	3 out of 6		5 out of 6
#steps Walktrap	**Three**	**Four**	**Five**
	4 out of 6	5 out of 6	5 out of 6

According to *Table 9*, the best parameters are the following:

- The best number of datatypes is three, instead of two or one.
- The best grouping method is the triple. That is, classifying two samples together into the same cluster if they have been clustered together for, at least, three methods.
- Adjacency matrix (discrete values) gives better values than using consensus matrix (continuous float values).
- Finally, setting four or five random steps in the Walktrap conduct to the same outcome.

5. Conclusion and future goals

Once finished the implementation and deeply analyzed the outcome, find a brief review of the most relevant features and achievements acquired during this project in the next sections.

5.1. Workflow assessment and conclusions

The hypothesis of the project was "the combination of the existing multi-omics data integration packages in R improve the subtyping of cancer patients". At this point, it can be concluded that the hypothesis was correctly verified. The workflow combines a set of methods and algorithms that improve the classification of 6 cancer datasets into different subgroups.

According to the requirements of the project, most of the goals were accomplished. Firstly, a pre-processing step was developed to reduce the data size of DNA-methylation files, not for the other datatypes. The download module to retrieve data from TCGA was not implemented due to the lack of time. This is an important missing point to deal with, in future implementations. However, sample files standardization was achieved using an automated script. Secondly, the existing R packages – for data clustering - were integrated. This includes finding the best k for each method according to different metrics (AUC, CDF and BIC) allowing the algorithms to label the samples accordingly. Thirdly, the main parameters that affect the whole process were tested, finding the way how they disrupt the results. Additionally, the best values for each parameter were also found. Survival analysis was used as a metric to assess the data outcome, considering it the best metric to verify the results.

Multi-omics data integration is currently one of the fields that is experiencing more progress in the last years. Future efforts should be addressed principally into omics databases standardization and, obviously, the improvement of the existing integration methods by searching the best parameters or by finding novel approaches.

5.2. Personal review

When I started my Erasmus+ Traineeship in Acibadem University (Istanbul, Türkiye) in the first semester of 2022-2023. My mentor, Dr. Osman Uğur Sezermann, explained me different options to carry a project. Among all the options and taking into consideration my degree's background and interests, I decided to take a chance into multi-omics data integration field.

At that point, the main challenges were basically three. Firstly, it was the first time that I was abroad for a long time. Moreover, I did not have my family and friends with me. Additionally, the language could be a barrier because it was the first time that I was doing my routine full in English, and even more, in Turkish. Secondly, I had never performed a multi-omics data integration before. Furthermore, I only stayed in contact with one omics science – metabolomics- in another traineeship. But in this project, we were focusing on another omics sciences - genomics and transcriptomics – that are completely different to metabolomics.

Thirdly and finally, I knew I had a really short period of time (only 4 months) to learn about that kind of analysis and to propose a method that could improve the existing ones.

Given the circumstances, it was a really challenging task to finish the project before 4 months. I could do most of the work during my traineeship in Istanbul, but it could not be finished, so I continued working on it when I came back to Catalunya. During the second semester I also got a job and had finish my subjects of university. That situation decelerated the project pace until the end of my subjects in university.

However, I am glad of all the work that I have been done during all this journey and I hope to get more time to improve my knowledge about multi-omics data integration.

6. References

Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 12: Survival analysis. *Critical Care*, *8*(5), 389. https://doi.org/10.1186/CC2955

Bland, J. M., & Altman, D. G. (2004). Statistics Notes: The logrank test. *BMJ : British Medical Journal*, *328*(7447), 1073. https://doi.org/10.1136/BMJ.328.7447.1073

Cancer. (n.d.). Retrieved August 12, 2023, from https://www.who.int/news-room/fact-sheets/detail/cancer

Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., & Bray, F. (2021). Cancer statistics for the year 2020: An overview. *International Journal of Cancer*, *149*(4), 778–789. https://doi.org/10.1002/IJC.33588

Gao, F., Wang, W., Tan, M., Zhu, L., Zhang, Y., Fessler, E., Vermeulen, L., & Wang, X. (2019). DeepCC: a novel deep learning-based framework for cancer molecular subtype classification. *Oncogenesis 2019 8:9*, *8*(9), 1–12. https://doi.org/10.1038/s41389-019-0157-8

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification 1985 2:1*, *2*(1), 193–218. https://doi.org/10.1007/BF01908075

Kim, S. H., Herazo-Maya, J. D., Kang, D. D., Juan-Guardela, B. M., Tedrow, J., Martinez, F. J., Sciurba, F. C., Tseng, G. C., & Kaminski, N. (2015). Integrative phenotyping framework (iPF): integrative clustering of multiple omics data identifies novel lung disease subphenotypes. *BMC Genomics*, *16*(1). https://doi.org/10.1186/S12864-015-2170-4

Kuijjer, M. L., Paulson, J. N., Salzman, P., Ding, W., & Quackenbush, J. (2018). Cancer subtype identification using somatic mutation data. *British Journal of Cancer*, *118*(11), 1492. https://doi.org/10.1038/S41416-018-0109-7

Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, *52*(1–2), 91–118. https://doi.org/10.1023/A:1023949509487/METRICS

Muzny, D. M., Bainbridge, M. N., Chang, K., Dinh, H. H., Drummond, J. A., Fowler, G., Kovar, C. L., Lewis, L. R., Morgan, M. B., Newsham, I. F., Reid, J. G., Santibanez, J., Shinbrot, E., Trevino, L. R., Wu, Y. Q., Wang, M., Gunaratne, P., Donehower, L. A., Creighton, C. J., … Thomson., E. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature 2012 487:7407*, *487*(7407), 330–337. https://doi.org/10.1038/nature11252

Nguyen, T., Tagett, R., Diaz, D., & Draghici, S. (2017). A novel approach for data integration and disease subtyping. *Genome Research*, *27*(12), 2025–2039. https://doi.org/10.1101/GR.215129.116

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846–850. https://doi.org/10.1080/01621459.1971.10482356

Schober, P., & Vetter, T. R. (2021). Kaplan-Meier Curves, Log-Rank Tests, and Cox Regression for Time-to-Event Data. *Anesthesia and Analgesia*, *132*(4), 969–970. https://doi.org/10.1213/ANE.0000000000005358

Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics and Biology Insights*, *14*. https://doi.org/10.1177/1177932219899051

The Cancer Genome Atlas Program (TCGA) - NCI. (n.d.). Retrieved August 4, 2023, from https://www.cancer.gov/ccg/research/genome-sequencing/tcga

Vailati-Riboni, M., Palombo, V., & Loor, J. J. (2017). What are omics sciences? *Periparturient Diseases of Dairy Cows: A Systems Biology Approach*, 1–7. https://doi.org/10.1007/978-3-319-43033-1_1/COVER

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., & Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods 2014 11:3*, *11*(3), 333–337. https://doi.org/10.1038/nmeth.2810

Wilkerson, M. D., & Hayes, D. N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*, *26*(12), 1572. https://doi.org/10.1093/BIOINFORMATICS/BTQ170