Integrative Analysis of Omics Data with Biological Knowledge in Translational Medicine

Ferran Briansó

Facultat de Biologia
Departament de Genètica, Microbiologia i Estadística

Universitat de Barcelona

A thesis submitted for the degree of

Doctor of Philosophy

XXXX XX 2024

For XXXXX XXXXXX

Acknowledgements

. . . . . .

Ferran Briansó
Mataró, BCN
XX XXXXXX 2024

Abstract (wip)

Over the past decade, advancements in omics technologies have facilitated the high-throughput monitoring of molecular and organism processes. These techniques have been widely applied to identify biological agents and to characterize biochemical systems, often focusing on the discovery of therapeutic targets and biomarkers related with specific diseases (Cisek et al., 2016),(K. Wang et al., 2014),(F. Wang et al., 2014). While many single-omic approaches target comprehensive analysis of genes (genomics), mRNA (transcriptomics), proteins (proteomics), and metabolites (metabolomics) among other, there is still field to improve omics data analyses through integrative methods (Wanichthanarak et al., 2015),(Gomez-Cabrero et al., 2014). In this sense, the integrative point of view defined in the paragraph above, applied to multi-omics data, is a promising approach to achieve better biomarker development in biomedical research projects, and this is the core idea of this work.

As the field of omics has evolved from analyzing a unique type of data to multiple types, it has been natural to extend the previous use of multivariate techniques to this new situation. With this aim classical and new multivariate techniques have been applied to the analysis of multi-omics datasets. Many of these techniques are dimension reduction methods that aim at finding main sources of variability in the data while maximizing some information characteristic such as the variance of each dataset, the correlation between groups of variables or other. Examples of such techniques are well consolidated methods such as Principal Component Analysis (PCA), Singular Value Decomposition (SVD), Correspondence Analysis (CA), and Partial Least Squares (PLS). Besides these more "novel" approaches have been used such as: Principal Components Regression, Coinertia and Multiple Coinertia Analysis, Generalized SVD, Sparse PLS, Multiple Factor Analysis (MFA), or combined versions of them (M. Wheelock & E. Wheelock, 2013),(Lê Cao et al., 2009),(Culhane et al., 2003). Publications such as (Meng et al., 2016), (Cavill et al., 2016), (Wu et al., 2019), (Subramanian et al., 2020), (Krassowski et al., 2020), and (Cantini et al., 2021), are good reviews of the state of the art of using multivariate and joint reduction methods for Integrative Multi-Omics Analysis.

Dimension reduction methods, especially those that are able to deal with situations that are typical from the omics context (with many more variables than samples, or possibly sparse matrices with many missing values), have been of

great help in visualizing datasets or even for performing variable selection to find biomarkers for a given situation (Wu et al., 2019). There is however one point where they underperform other approaches, that is, the difficulty in interpreting results from a biological point of view. This is relatively reasonable, because the most of these methods work by creating new variables that are some type of linear combination from the original ones. While this is useful, for example, for removing redundancy, this does not provide any clues on what these new dimensions may mean from a biological point of view.

This problem has been known since the beginning of using multivariate methods with omics data, but only a few approaches have been taken to deal with this. The first attempts to introduce biological information in the analyses consisted of using the most well-known database of biological functions, the Gene Ontology (GO) (Ashburner et al., 2000). Fellenberg (Busold et al., 2005) introduces a way to integrate Gene Ontology information with Correspondence Analysis to facilitate the interpretation of microarray data. De Tayrac et al. (Tayrac et al., 2009) applies multiple factor analysis to the integrative analysis of microarray and DNA copy number data. They apply GO Terms on data visualizations by treating these terms as supplemental information. In recent years the representation of biological knowledge has shifted from Gene Ontology to using Gene Sets (Huang et al., 2009). Meng and Culhane (Meng et al., 2016) have introduced the Integrative Clustering with Gene Set Analysis where gene set expression analysis is performed based on multiple omics data; and Tyekucheva et al. (Tyekucheva et al., 2011), go one step further and use the results of Gene Set Expression Analysis (GSEA) to integrate different omics data.

The previous approaches show several things: Although the idea that integrating quantitative data with biological knowledge may increase interpretability, the number of successful attempts to do this is still small. In this thesis, the use of either classical GO Terms or more flexible annotations (Gene Sets or custom annotations), will be combined with different approaches, and combinations of them if needed, to guide integrative analysis and to improve its biological interpretability from the point of view of the biomedical researchers.

Contents

Appendices

List of Figures

List of Tables

List of Abbreviations

1-D, 2-D . . . One- or two-dimensional, referring **in this thesis** to spatial dimensions in an image.

Otter . . . . . One of the finest of water mammals.

Hedgehog . . . Quite a nice prickly friend.

1 Introduction

Contents

1.1 Content of the introductory text (wip)

The general concept of Data Integration can be defined as the combination of data from different sources to provide users with a unified view of the data (Lenzerini, 2002). However, the practical meaning of the term Integration may vary from, for instance, the computational combination of data to the combination of studies

performed independently, the simultaneous analysis of multiple variables on multiple datasets, or any possible approach for homogeneously querying heterogeneous data sources. Therefore, in many cases, integrative analysis may be preferable to a simple combination of data from distinct sources. Integrative analysis allows not only the combination of heterogeneous data but also the combined use of these data to obtain the most relevant information and, what is better, to be able to extract some information that could not be unveiled by the separate analysis of each of the original data types.

Over the past decade, advancements in omics technologies have facilitated high-throughput monitoring of molecular and organism processes. These techniques have been widely applied to identify biological agents and characterize biochemical systems, often focusing on the discovery of therapeutic targets and biomarkers related to specific diseases (Cisek et al., 2016),(K. Wang et al., 2014), and (F. Wang et al., 2014). Although many single-omic approaches target comprehensive analysis of genes (genomics), mRNA (transcriptomics), proteins (proteomics), and metabolites (metabolomics), among others, there is still a need to improve omics data analyses through integrative methods (Wanichthanarak et al., 2015),(Gomez-Cabrero et al., 2014). In this sense, the integrative point of view defined in the paragraph above, applied to multi-omics data, is a promising approach to achieving better biomarker development in biomedical research projects, and this is the core idea of this work.

As the field of omics has evolved from analyzing unique types of data to multiple types, it is natural to extend the previous use of multivariate techniques to this new situation. With this aim, classical and new multivariate techniques have been applied for the analysis of multi-omics datasets. Many of these techniques are dimension reduction methods that aim to find the main sources of variability in the data while maximizing some information characteristics, such as the variance of each dataset and the correlation between groups of variables. Examples of such techniques are well-consolidated methods such as Principal Component Analysis (PCA), Singular Value Decomposition (SVD), Correspondence Analysis (CA), and

Partial Least Squares (PLS). Besides these more "novel" approaches have been used such as: Principal Components Regression, Coinertia and Multiple Coinertia Analysis, Generalized SVD, Sparse PLS, Multiple Factor Analysis (MFA), or combined versions of them (M. Wheelock & E. Wheelock, 2013),(Lê Cao et al., 2009),(Culhane et al., 2003). Publications such as (Meng et al., 2016), (Cavill et al., 2016), (Wu et al., 2019), (Subramanian et al., 2020), (Krassowski et al., 2020), and (Cantini et al., 2021), are good reviews of the state of the art of using multivariate and joint reduction methods for Integrative Multi-Omics Analysis.

Dimension reduction methods, especially those that are able to deal with situations that are typical of the omics context (with many more variables than samples, or possibly sparse matrices with many missing values), have been of great help in visualizing datasets or even in performing variable selection to find biomarkers for a given situation (Wu et al., 2019). However, there is one point where they underperform other approaches: the difficulty in interpreting results from a biological point of view. This is relatively reasonable because most of these methods work by creating new variables that are a type of linear combination from the original ones. While this is useful, for example, for removing redundancy, it does not provide any clues on what these new dimensions may mean from a biological point of view.

This problem has been known since the introduction of multivariate methods with omics data, but only a few approaches have been taken to deal with this problem. The first attempts to introduce biological information in the analyses consisted of using the most well-known database of biological functions, the Gene Ontology (GO) (Ashburner et al., 2000). Fellenberg (Busold et al., 2005) introduces a way to integrate Gene Ontology information with Correspondence Analysis to facilitate the interpretation of microarray data. De Tayrac et al. (Tayrac et al., 2009) applies multiple factor analysis to the integrative analysis of microarray and DNA copy number data. They applied GO Terms on data visualizations by treating these terms as supplemental information. In recent years, the representation of biological knowledge has shifted from Gene Ontology to Gene Sets (Huang et al.,

2009). Meng and Culhane (Meng et al., 2016) have introduced the Integrative Clustering with Gene Set Analysis where gene set expression analysis is performed based on multiple omics data; and Tyekucheva et al. (Tyekucheva et al., 2011), go one step further and use the results of Gene Set Expression Analysis (GSEA) to integrate different omics data.

Altogether, the previous approaches show several things: Although the idea that integrating quantitative data with biological knowledge may increase interpretability, the number of successful attempts to do this is still small. In this thesis, the use of either classical GO Terms or more flexible annotations (Gene Sets or custom annotations) will be combined with different approaches, and combinations of them, if needed, to guide integrative analysis and improve its biological interpretability from the point of view of biomedical researchers.

1.2 Background/State of the Art

Falta desenvolupar punts

1.2.1 Omics data analyses

3 problemes esencials (veure projecte recerca Alex):

- **Omics data may be partly incomplete**, especially in multiomics studies, where not all types of data are usually available for all individuals.
- **The results of these analyses are difficult to interpret**. If we agree that the ultimate goal of many analyzes is a better understanding of the underlying biological processes, for example, in a disease study context, it should be possible to establish a clear relationship between the outcome of an analysis and what this means biologically. And this is not always so.
- **These kind of data analytics are difficult to standardize**, as it is not easy to make complex pipelines of multi-omics analyses, which integrate multiple processes with multiple sources, easy to reproduce or communicate.

Més el problema de la p»n (Dimensionallity Reduction Techniques; The p»n situation)

1.2.2 The problem of having partly incomplete data

Having partly incomplete data is a common challenge in biomedical multi-omics data analyses, where not all omics layers or samples have complete measurements for all variables of interest. This problem, known as missing data, can hinder the integrative analysis and interpretation of multi-omics datasets.

Missing Data Types: Missing data can occur in various forms in multi-omics datasets. For example, some omics layers may have missing values for certain variables (e.g., genes, proteins, metabolites), or specific samples may be missing data for certain omics layers. This can result from technical limitations, experimental design, or inherent biological variability.(Flores et al., 2023)

Impact on Analysis: Incomplete data can introduce biases and distort the results of multi-omics analyses. It can affect downstream statistical analyses, clustering, network inference, and machine learning algorithms, leading to inaccurate or unreliable findings. Addressing missing data appropriately is crucial for obtaining valid and meaningful results. Reference:

Missing Data Mechanisms: Understanding the underlying mechanisms of missing data is essential for selecting appropriate imputation methods. Missing data can occur due to different mechanisms, such as missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). These mechanisms influence the choice of imputation techniques and the assumptions made during data analysis.(Little & Rubin, 2002)

Imputation Methods: Imputation techniques are employed to estimate missing values in multi-omics datasets. Various imputation methods, including mean imputation, regression imputation, multiple imputation, and machine learning-based approaches, have been proposed to handle missing data in different omics layers. Each method has its assumptions, strengths, and limitations, and the choice of imputation strategy should be carefully considered. Reference: Buuren, S. V.,

& Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45(3), 1-67.

Uncertainty and Sensitivity Analysis: Dealing with missing data introduces uncertainty in the imputed values and subsequent analyses. Sensitivity analyses, such as multiple imputation and bootstrapping, can help assess the robustness of the results to missing data assumptions and imputation methods. Reference: Sterne, J. A., et al. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. BMJ, 338, b2393.

Addressing the issue of incomplete data in multi-omics analyses is crucial to avoid biased or misleading results. By utilizing appropriate imputation methods and understanding the missing data mechanisms, researchers can mitigate the impact of missing data and enhance the accuracy and reliability of their analyses.

1.2.3 Results interpretation in the context of integrative multi-omics data analyses

Interpretation of results in integrative multi-omics data analyses is a critical challenge due to the complexity and high dimensionality of the data, as well as the need to integrate information from multiple omics layers. Here, I will explain the problem of result interpretation in this context and provide relevant bibliographic references.

Data Integration Challenges: Integrating multi-omics data involves combining information from different molecular layers such as genomics, transcriptomics, proteomics, and metabolomics. Each omics layer provides a unique perspective on biological processes, and integrating these layers can reveal comprehensive insights. However, interpreting the integrated results becomes challenging due to the heterogeneity and scale differences among the omics data. Reference: Wang, X., & Zhang, B. (2018). Integrating multiple 'omics' data for biomarker discovery and clinical assessment. Molecular & Cellular Proteomics, 17(6), 991-1003.

Dimensionality and Complexity: Multi-omics data analyses often result in high-dimensional datasets with numerous features, making it difficult to interpret the results directly. The challenge lies in identifying the most relevant features

or patterns and extracting meaningful biological insights from the vast amount of data. Reference: Nguyen, T. M., et al. (2019). Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. Advances in Genetics, 103, 143-175.

Contextual Interpretation: Interpreting multi-omics results requires considering the biological context, such as pathways, networks, and regulatory interactions. Understanding how different omics layers interact and influence each other within biological systems is crucial for accurate interpretation. Reference: Mei, H., et al. (2017). The road beyond omics: Integration of multi-omics data for the inference of regulatory networks and precision medicine. Computational and Structural Biotechnology Journal, 15, 359-366.

Validation and Biological Significance: Integrative multi-omics analyses often generate numerous associations, correlations, or biomarkers. However, validating and determining the biological significance of these findings is a key challenge. Experimental validation, functional enrichment analysis, and comparison with existing knowledge are essential for confirming the biological relevance of the results. Reference: Sun, H., et al. (2020). Strategies for interpreting multi-omics studies in schizophrenia and other neuropsychiatric disorders. Journal of Psychiatric Research, 129, 121-133.

Visualization and Interactive Tools: Visualizing and exploring multi-omics data can aid in result interpretation. Interactive visualization tools that integrate different omics layers, provide network views, and enable user-driven exploration can facilitate the interpretation process. Reference: Swatloski, T., & et al. (2020). Multi-Omics Data Integration, Interpretation, and Its Application. Genes, 11(10), 1162.

In summary, the problem of result interpretation in integrative multi-omics data analyses stems from the challenges of data integration, high dimensionality, contextual understanding, validation, and visual exploration. Addressing these challenges requires a combination of statistical methods, biological knowledge, and interactive tools to extract meaningful insights from the integrated data.

1.2.4 Approaches for the biological and clinical interpretation

The biological and clinical interpretation of multi-omics data analysis results is crucial for gaining insights into the underlying molecular mechanisms, identifying biomarkers, and understanding disease processes.

1. **Pathway and Functional Enrichment Analysis**: Pathway and functional enrichment analysis aim to identify overrepresented biological pathways, gene sets, or functional categories that are significantly associated with the differentially expressed genes or other omics features. These analyses help in understanding the biological processes, molecular functions, and cellular components that are affected in a particular condition or disease. Citation: Khatri, P., et al. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. PLoS Computational Biology, 8(2), e1002375.

2. **Network Analysis**: Network analysis involves the construction and analysis of biological networks, such as gene regulatory networks or protein-protein interaction networks, using multi-omics data. Network-based approaches help in identifying key hub genes, modules, or subnetworks that play important roles in disease progression or phenotype. Citation: Barabási, A. L., et al. (2011). Network medicine: A network-based approach to human disease. Nature Reviews Genetics, 12(1), 56-68.

3. **Machine Learning and Predictive Modeling**: Machine learning algorithms, such as random forests, support vector machines, or deep learning models, can be applied to multi-omics data to develop predictive models for disease diagnosis, prognosis, or treatment response. These models can uncover potential biomarkers or patterns in multi-omics data and provide insights into disease classification and personalized medicine. Citation: Alizadeh, A. A., et al. (2000). Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. New England Journal of Medicine, 344(14), 1031-1037.

4. **Integration of Multi-Omics Data**: Integrative analysis methods aim to combine and analyze different omics datasets, such as transcriptomics, proteomics, and epigenomics, to identify molecular interactions and relationships across different layers of biological information. These methods enable a more comprehensive understanding of the molecular mechanisms underlying complex diseases or biological processes. Citation: Liu, Y., et al. (2014). A survey of integrative analysis methods for multi-omics data. Statistical Methods in Medical Research, 27(11), 3061-3077.

5. **Data Visualization**: Data visualization techniques, such as heatmaps, scatter plots, or network visualizations, play a crucial role in the interpretation of multi-omics data analysis results. Visualizations help in identifying patterns, clusters, and relationships between variables, enabling researchers to generate hypotheses and communicate findings effectively. Citation: Gehlenborg, N., et al. (2010). Visualization of omics data for systems biology. Nature Methods, 7(3), S56-S68.

These methods, among others, contribute to the biological and clinical interpretation of multi-omics data analysis results, providing insights into disease mechanisms, biomarker discovery, and potential therapeutic targets.

1.2.5 Data processing and standarization

Data processing and standardization are critical steps in biomedical multi-omics data analyses to ensure data quality, comparability, and compatibility across different omics layers and studies. In this context, I will explain the problem of data processing and standardization and provide relevant bibliographic references.

Data Preprocessing: Raw multi-omics data often require preprocessing steps to handle technical variations, correct systematic biases, and remove noise. This may involve background correction, normalization, batch effect removal, and quality control measures to ensure data quality and comparability. Reference: Tarazona,

S., et al. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Research, 43(21), e140.

Integration Challenges: Integrating multi-omics data involves combining information from different omics layers, which may have distinct measurement scales, dynamic ranges, and data distributions. Harmonizing the data across omics layers is necessary to enable meaningful comparisons and integrative analyses. Reference: Meng, C., et al. (2014). Dimension reduction techniques for the integrative analysis of multi-omics data. Briefings in Bioinformatics, 17(4), 628-641.

Missing Data Handling: In multi-omics datasets, missing data can be present due to technical limitations or experimental designs. Proper handling of missing data, such as imputation or exclusion strategies, is crucial to avoid biases and ensure accurate analyses. Reference: Zhou, Y., et al. (2021). Missing data imputation in single-cell RNA sequencing and its implications in integrative multi-omics analysis. Briefings in Bioinformatics, 22(5), bbaa212.

Standardization and Metadata: Standardization of data formats, annotation, and metadata is vital for data sharing, reproducibility, and cross-study comparisons. The use of common data standards and ontologies facilitates data integration and harmonization efforts. Reference: Sansone, S. A., et al. (2012). Toward interoperable bioscience data. Nature Genetics, 44(2), 121-126.

Quality Control: Implementing quality control measures is essential to identify and remove low-quality or unreliable data points. Quality control procedures can include outlier detection, sample exclusion criteria, and identifying technical artifacts or batch effects. Reference: Leek, J. T., et al. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics, 11(10), 733-739.

Effective data processing and standardization in multi-omics analyses are crucial for accurate and meaningful interpretations. These steps ensure data quality, comparability, and compatibility, enabling integrative analyses and cross-study comparisons.

1.2.6 Tools for the development of bioinformatics pipelines in biomedical multi-omics data integration

1.2.7 Motivation for Integrative analysis

The fable of the blind men and the elephant (https://en.wikipedia.org/wiki/Blind_men_and_an_elephant) is a metaphorical story that can be applied to various contexts, including the motivation behind using distinct omics data types in biomedical integrative data analyses. In this fable, several blind men touch different parts of an elephant and form their own interpretations based on the limited information they gather from their individual experiences. See Figure 1.1. In the parable, several blind men touch different parts of an elephant, but each one perceives only a small aspect of the whole animal. As a result, they form vastly different and often conflicting impressions of what an elephant is. Each blind man, based on his limited sense of touch, describes the elephant differently. One might touch the tail and think the elephant is like a rope, while another feeling the leg believes it's like a tree trunk. Yet another touching the ear might think it's like a fan. None of them, however, comprehends the entirety of the elephant. See Figure 1.2.

The parable is often interpreted to convey the idea that individuals may have partial, subjective truths based on their limited experiences and perspectives. It's a metaphor for the limitations of perception and the importance of considering multiple viewpoints to arrive at a more complete understanding of a complex reality. Similarly, in biomedical research, different omics data types provide distinct perspectives on biological processes, and no single omics layer can fully capture the complexity of the underlying system. Each omics layer, such as genomics, transcriptomics, proteomics, and metabolomics, provides specific insights into different molecular components and interactions. By integrating these diverse data types, we aim to create a more comprehensive and accurate understanding of the biological system, similar to how the blind men can form a more complete understanding of the elephant by sharing and integrating their individual observations.

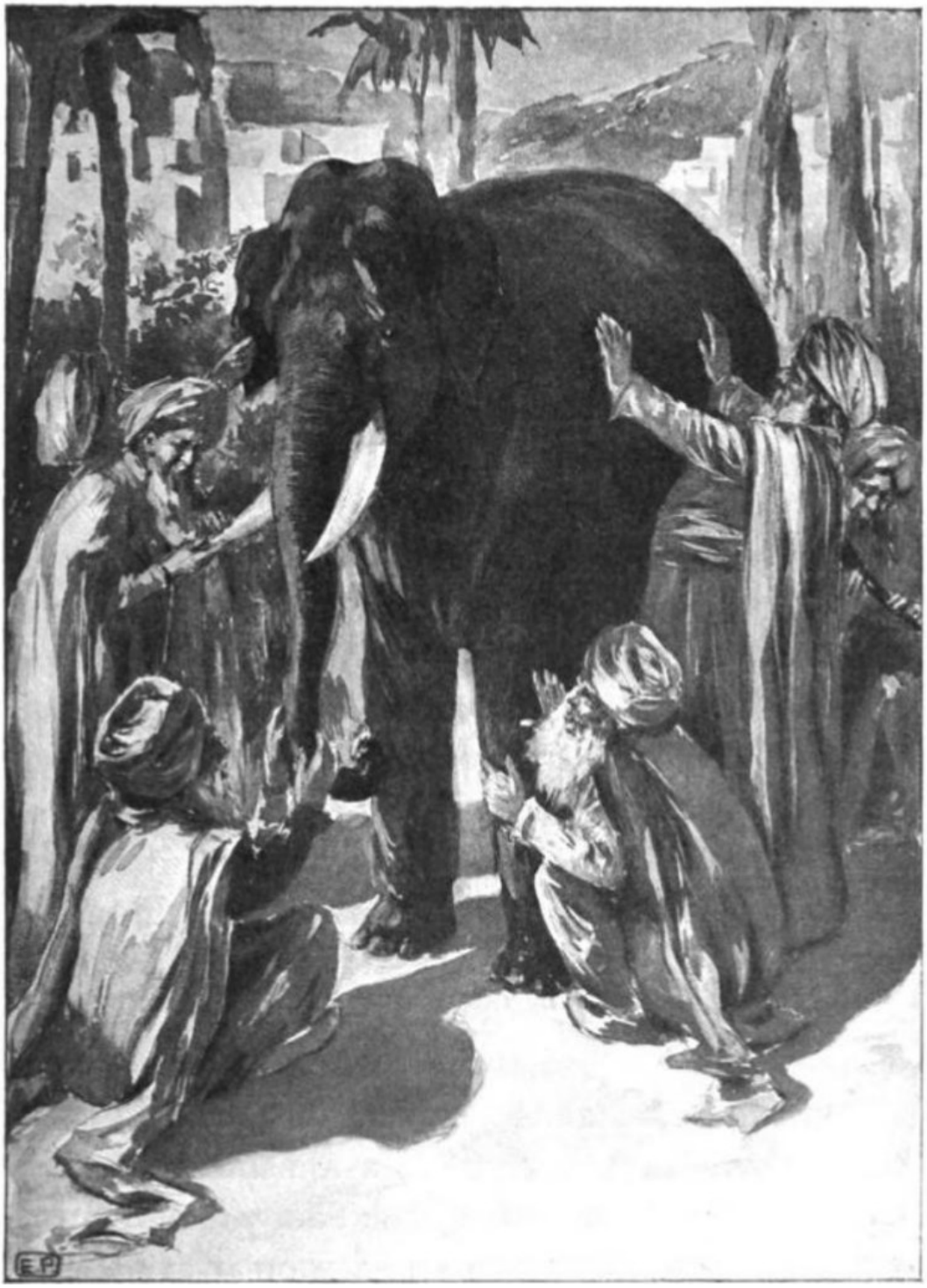

Figure 1.1: The blind men and the elephant. By Illustrator unknown - From The Heath readers by grades, D.C. Heath and Company (Boston), p. 69., Public Domain, https://commons.wikimedia.org/w/index.php?curid=4581263

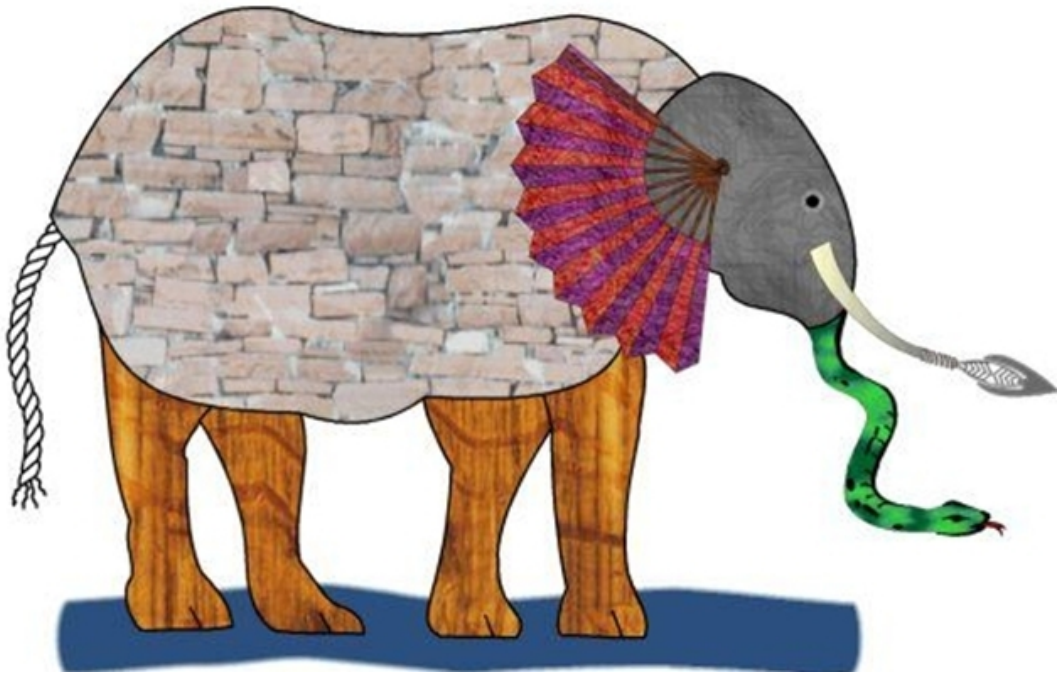

Figure 1.2: Reconstruction of the elephant as the blind men perceive it. Image source: http://doug-johnson.squarespace.com/blue-skunk-blog/2012/12/8/the-blind-men-and-the-elephant.html;jsessionid=30DE43866E1B453471B75CB39688E2CB.v5-web003

Each omics data type reveals a specific aspect of biological information. For example, genomics focuses on the DNA sequence, providing insights into genetic variations and potential disease-causing mutations. Transcriptomics examines gene expression levels, helping us understand which genes are active in a given condition. Proteomics investigates the expression and abundance of proteins, shedding light on protein-protein interactions and signaling pathways. Metabolomics analyzes small molecules, providing insights into metabolic pathways and cellular processes.

By integrating these different omics layers, we can overcome the limitations of each individual data type and gain a more holistic understanding of biological phenomena. Integrative multi-omics data analyses enable us to uncover complex relationships, identify key biological pathways, discover biomarkers, and generate more accurate predictions for diseases and therapeutic interventions.

Just as the blind men needed to collaborate and share their individual perceptions to form a complete understanding of the elephant, biomedical researchers can leverage the strengths of different omics data types and integrate their findings to

reveal a more comprehensive picture of biological systems. Integrative approaches allow us to move beyond isolated observations and capture the intricate interplay among genes, proteins, metabolites, and other molecular entities.

In conclusion, the fable of the blind men and the elephant serves as an analogy for the motivation behind using distinct omics data types in biomedical integrative data analyses. Just as the blind men's individual perceptions were limited, focusing on a single omics data type can lead to an incomplete understanding of complex biological processes. Integration of diverse omics data types enables us to overcome these limitations and gain a more comprehensive understanding of the intricacies of living systems.

Interpretability is a weak point of most multi omics approaches

Methods focus much more on feature selection discovery and interaction highlighting measurement than on clinical or biological interpretability.

1.2.8 Application of existing approaches for multi-omics data integration

MCIA, RGCCA, MFA... Cavill, 2016; Culhane 2003...

1.2.9 Revisió de metodes de creacio pipelines

2 Objectives

The main objectives of this work are the following:

1. To make an empirical comparison of some of the currently available dimension reduction techniques applied for the integration of omics data, focused on their ability to include biological annotations,

2. To develop methods and workflows able to apply these techniques, focusing on the matching of distinct omics datasets relying on biological knowledge,

3. To apply these methods to specific translational biomedical research cases, such as an integrative analysis of transcriptomics and proteomics data to study ischemic stroke, as well as to public datasets, which can be easily shared and are not as restricted by sample sizes as other projects.

4. To implement the knowledge acquired with this work into the appropriate bioinformatics tools, e.g. R packages or web-based tools, that will be used in future biomedical research projects for providing a better interpretation of this kind of studies.

All these objectives are in agreement with the tasks defined within a project partially supported by Grant MTM2015-64465-C2-1-R (MINECO/FEDER) from the Ministerio de Economía y Competitividad (Spain), to which the PhD Thesis proposed here is related.

Ein Mann, der recht zu wirken denkt,
Muß auf das beste Werkzeug halten
The man who seeks to be approved,
must stick to the best tools for it
— Goethe's *Faust. Eine Tragödie* (1808).

3 Methodology

Contents

3.1 Working phases

Working phases, with the corresponding steps, followed in order to achieve the above objectives: SHA DE SEPARAR ENTRE AQUI I ELS RESULTATS

1. Application of integrative multi-omics methods to (I) the analysis of specific data sets provided by research units from our former affiliation center, VHIR, and other research institutions that we collaborate with (Rodríguez-Hernández et al., 2016), (Rodriguez-Fernandez et al., 2018), (Simats et al., 2020) and (II) to the integrative analysis of larger data sets from public

data bases, such as Breast Cancer samples from the TCGA project [TCGA Research Network: http://cancergenome.nih.gov/], [TCGA-BRCA Project: https://portal.gdc.cancer.gov/projects/TCGA-BRCA].

2. Development of methods, either in terms of new algorithms or in terms of combinative workflows, which will be able to improve, and facilitate, the analysis and biological interpretation of those data sets to be integrated.

3. Implementation of the methods developed for this study in the appropriate bioinformatics tools, such as an R package or a web-based application, to facilitate their use in the context of biomedical research projects.

Here follows a brief description of these main five activities, the methods in which they are initially based, the objectives that they are related to, and the corresponding results:

1. Application of some state-of-the-art methods for integrative multi-omics data analysis to the study of human brain tissue samples, collected by the Neurovascular Diseases Laboratory at Vall d'Hebron Research Institute. This part is already finished, and led to publications in 2018 and 2021 (Simats et al., 2020), (Ramiro et al., 2021). Researchers obtained different omics data from necropsies, which had been processed to obtain mRNA, microRNA and protein expression values. Each dataset had been first analyzed independently using standard bioinformatics protocols [R Development Core Team. 2008]. These analyses allowed selecting subsets of relevant features, for each type of data, to be used in the integrative analysis. Among all available options, we decided to use two distinct and complementary approaches: (I) Multiple Co-inertia Analysis implemented in Bioconductor packages made4 (Culhane et al., 2005) and mogsa (Singh et al., 2016), and (II) Regularized Canonical Correlation Analysis with Sparse Partial Least Squares regression (sPLS), provided by mixomics R package (Rohart et al., 2017). This work had been presented at some meetings (Briansó et al., 2016a), (Briansó et al.,

2016b), (García-Berrocoso et al., 2016), (García-Berrocoso et al., 2017) and in an already published extended abstract's series book (Briansó et al., 2017). This step had been obviously useful for the achievement of the objective number 3 explained in the previous section, which aims on the study of the regulome's response to ischemic stroke, but also useful for detecting the advantages and drawbacks of the methods applied, thus setting the basis for the work regarding to objective number 2.

2. Reproduction of the same analyses steps performed in point 1) above with publicly available databases, such as distinct omics data from 150 samples from the TCGA-BRCA collection. This data set contains the expression or abundance of mRNA, miRNA and proteomics for 150 breast cancer samples previously prefiltered, as explained in Rohart et al. [29], and allows identifying a good multi-omics signature to discriminate between Basal, Her2 and Luminal A breast cancer subtypes. This work is already finished, and complies with objectives 3 and 2.

3. Use of all the data sets analyzed up to this point to make a comparison of results between the main implemented methods, and eventually some others, which is the aim of objective 1. This is based on quantitative and qualitative comparison and visualization methods, such as those explained by Thallinger [24] and Martin [25], going from simple Venn diagrams to more complex, network analysis, software such as some specific R packages [20] or Cytoscape [26]. The focus here is to use graphical visualization elements to compare the results of the analyses with and without the addition of biological information.

4. Development of new methods and/or workflows in order to improve and/or combine the benefits from the selected approaches, with focus in those allowing the addition of biological significance to the integration process. Here follows an overview of the methods developed to expand the original datasets (X, Y) with annotations (Ax, Ay) to obtain new blocks of data (Nx, Ny,and

Nxy). And the workflow has been implemented adapting the integrative pipelines applied so far to the R targets package [33], a pipeline toolkit that improves reproducibility, skipping unnecessary steps already up to date and showing tangible evidence that the results match the underlying code and data. The development of this targets workflow is intended to comply with the objective number 2 of this working plan.

5. Implementation of the methods resulting from 4) as a new R package to be submitted to Bioconductor repository [27], and, finally, to complete objective 4 of this thesis plan, as a web application [28] to be used in further steps of the current biomedical research projects in which our collaborators are implied, as well as in future studies.

3.2 Methodology

In the context of multi-omics data integration, our proposal relies on the idea that incorporating biological annotations into data sets before integrative analysis enriches outcomes and enhances their biological interpretability. So, augmenting quantitative omics data with contextual biological knowledge deepens our understanding of complex biological phenomena. To do that, we begin with meticulous data quality assessment and standardization, laying the foundation for reliable analyses. We then infuse biological knowledge using standard biological annotations, creating "Expanded Datasets" that provide context for comprehensive analysis. Advanced dimension reduction techniques can be applied then in illuminating hidden patterns and relationships between data sources or blocks and, finally, the semi-automation capabilities of the Targets R package let us to build an easy-to-use implementation of the process.

3.2.1 Data Quality Assessment and Format Review

Prior to commencing the integrative analysis, a rigorous data quality assessment and format review were conducted to ensure the reliability and compatibility of the

input datasets. This critical step aimed to identify and rectify any discrepancies, inconsistencies, or errors that might affect the subsequent analyses. The following procedures were employed:

- Data Source Selection: Datasets from distinct omics technologies, including genomics, transcriptomics, proteomics, and metabolomics, were obtained from reliable sources and repositories. It is essential to note that data sources were carefully selected to ensure consistency and adherence to standardized formats.

- Data Preprocessing: Raw omics data underwent preprocessing to address issues such as missing values, outliers, and data normalization. This preprocessing step was essential to enhance data quality and comparability.

- Data Format Standardization: Datasets were reviewed for consistency in data formats, including file types, column naming conventions, and units of measurement. Non-standardized data were transformed to a common format to facilitate downstream integration.

- Quality Control: Quality control checks were performed to assess the reliability of data sources. This included evaluating data reproducibility, assessing batch effects, and conducting statistical tests to identify data points or samples requiring further investigation.

3.2.2 Pre-processing for Integration of Biological Knowledge: Generating the "Expanded Datasets"

The integration of biological knowledge into the omics datasets was achieved through a pre-processing step aimed at expanding the data matrices with annotations accessed from specialized R libraries, which provided direct access to curated biological databases such as Gene Ontology (GO) and pathway information (e.g., KEGG). This process resulted in what we term "Expanded Data Sets", which

include the original biological features (e.g. gene expressions) as well as new variables coming from the annotation of biological terms. The following steps detail the pre-processing procedure:

- Selected biological annotations: Specialized R libraries, dedicated to biological knowledge integration, were employed to access and retrieve up-to-date annotations from databases such as GO and KEGG.
- Data-Annotation Mapping: Each omics dataset was mapped to the retrieved biological annotations based on identifiers (e.g., gene or protein names) using the capabilities of the R libraries. This step facilitated the linking of omics data with biological knowledge.
- Annotation Integration: The annotated information accessed through the specialized R libraries was integrated into the original omics datasets, resulting in expanded data matrices that combined the original quantitative omics measurements with new quantified features associated with the given biological annotations.

PENDENT DE DETALLAR:

- Significació biològica, com faig les anotacions
- incloure aquí Biological Interpretation
- Expansió de les matrius (numeritzar anotacions, creació de noves vars a partir de les anotacions)

Start the process already having a couple [punt de millora: admetre 3 o + inputs] of data sets from distinct 'omics sources, mapped to gene ids (if GO annotation has to be performed), containing the results from a selection of differentially expressed genes or most relevant proteins analysis, or similar.

[explicar aquí els requeriments de format dels data sets d'entrada!!]

For each input data set, if annotations are not already provided, two distinct basic annotation methods can be performed:

```
  [1] "RTN2"      "NDRG2"     "CCDC113"   "FAM63A"    "ACADS"     "GMDS"      "HLA.H"     "SEMA4A"    "ETS2"      "LIMD2"     "NME3"
 [12] "ZEB1"      "CDCP1"     "GIYD2"     "RTKN2"     "MANSC1"    "TAGLN"     "IFIT3"     "ARL4C"     "HTRA1"     "KIF13B"    "CPPED1"
 [23] "SKAP2"     "ASPM"      "KDM4B"     "TBXAS1"    "MT1X"      "MED13L"    "SNORA8"    "RGS1"      "CBX6"      "WWC2"      "TNFRSF12A"
 [34] "ZNF552"    "MAPRE2"    "SEMA5A"    "STAT5A"    "FLI1"      "COL15A1"   "C7orf55"   "ASF1B"     "FUT8"      "LASS4"     "SQLE"
 [45] "GPC4"      "AKAP12"    "AGL"       "ADAMTS4"   "EPHB3"     "MAP3K1"    "PRNP"      "PROM2"     "SLCO3A1"   "SNHG1"     "PRKCDBP"
 [56] "MXI1"      "CSF1R"     "TANC2"     "SLC19A2"   "RHOU"      "C4orf34"   "LRIG1"     "DOCK8"     "BOC"       "C11orf52"  "S100A16"
 [67] "NRARP"     "TTC23"     "TBC1D4"    "DEPDC6"    "ILDR1"     "SDC1"      "STC2"      "DTWD2"     "TCF4"      "ITPR2"     "DPYD"
 [78] "NME1"      "EGLN3"     "CD302"     "AHR"       "LAPTM4B"   "OCLN"      "HIST1H2BK" "HDAC11"    "C18orf1"   "C6orf192"  "AMPD3"
 [89] "COL6A1"    "RAB3IL1"   "APBB1IP"   "PSIP1"     "EIF2AK2"   "CSRP2"     "EIF4EBP3"  "LYN"       "WDR76"     "SAMD9L"    "ASPH"
[100] "RBL1"      "SLC43A3"   "HN1"       "TTC39A"    "MTL5"      "NES"       "APOD"      "RIN3"      "ALCAM"     "C1orf38"   "PLCD3"
[111] "BSPRY"     "NTN4"      "IL1R1"     "EMP3"      "ZKSCAN1"   "FMNL2"     "OGFRL1"    "IRF5"      "IGSF3"     "DBP"       "CNN2"
[122] "CAMK2D"    "SIGIRR"    "AKAP9"     "ICA1"      "FGD5"      "DSG2"      "E2F1"      "QSOX1"     "TOB1"      "CSF3R"     "SHROOM3"
[133] "CCDC80"    "FRMD6"     "CXCL12"    "CCNA2"     "TIGD5"     "ALDH6A1"   "POSTN"     "FZD4"      "NCAPG2"    "SDC4"      "SNED1"
[144] "PLEKHA4"   "KCNAB2"   "SH3KBP1"   "IGSF9"     "DNLZ"      "S1PR3"     "PTPRE"     "FLJ23867"  "PLSCR1"    "LMO4"      "IFITM2"
[155] "LRRC25"    "TST"       "NCF4"      "NCOA7"     "IL4R"      "CCDC64B"   "SGPP1"     "RUNX3"     "SLC5A6"    "IFIH1"     "PREX1"
[166] "PLAUR"     "CDK18"     "SLC43A2"   "GK"        "ICAM2"     "YPEL2"     "CBR1"      "MEX3A"     "ZNRF3"     "PTPRM"     "C1orf162"
[177] "GAS6"      "C1QB"      "PVRL4"     "CTSK"      "MRVI1"     "LEF1"      "PLCD4"     "ZNF37B"    "MEGF9"     "GINS2"     "FAM13A"
[188] "CPT1A"     "SNX10"     "TRIM45"    "ELP2"      "ALOX5"     "AMN1"      "CERCAM"    "SEMA3C"    "KRT8"      "TP53INP2"  "JAM3"
[199] "ZNF680"    "PBX1"
```

Figure 3.1: List of gene symbols used as example

(i) a basic GO mapping, returning annotations to those GO entities for which we find more than a certain number of features (gene ids coming from our data set) annotated to them,

[mostrar formula] [mostrar exemple]

(ii) a Gene Enrichment Analysis (based on Hypergeometric tests against all GO categories, with FDR correction[ref clusterProfiler]) is performed in order to retrieve the most relevant annotations to that set of genes/features.

[mostrar exemple de llista de gens]

[afegir aquí la opció d'afegir les anotacions com a individus suplementaris enlloc de variables]

Figure 3.2 is an example.

Alternatively, manual annotations can be provided (eg. GO terms, canonical pathways, or even annotation to custom entities) as an optional input file.

[mostrar el format requerit].

Other annotation methods can be implemented, as functions to be used by the main pipeline, if more complex methods for biological information addition are required.

[Mostrar el format final de les anotacions, com a matrius dels data sets amb anotacions binàries 1/0 com a columnes extra]

Once the annotations are already computed, mapping each feature of the input data set to the corresponding biological entity, they can be used to generate new

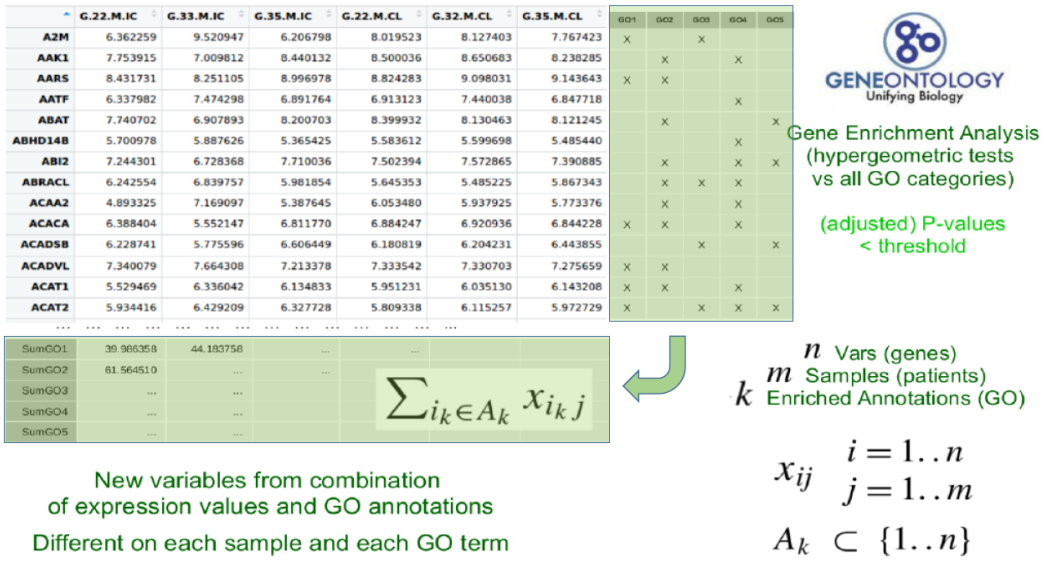

Figure 3.2: Addition of GO terms

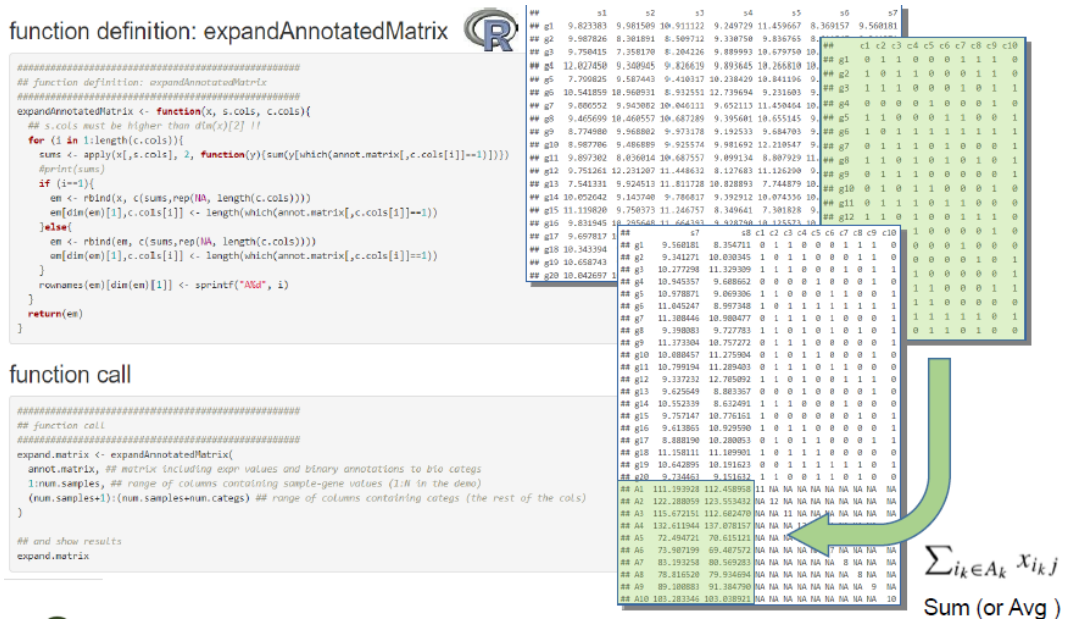

Figure 3.3: Addition of news feats

features (as new rows), computing the average value [punt de millora: funció de ponderació] of the expression/intensity values from all original features being mapped to the annotated biological entities.

Once we have the annotated matrices (Figure 3.5, highlighted in blue) we proceed to generate the Expanded matrices (in green) by casting these annotations as numerical values, that is, calculating the average of the numerical expressions

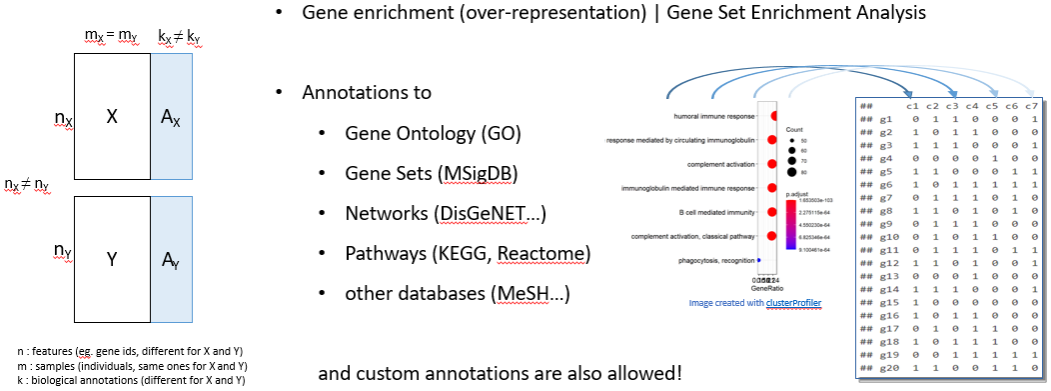

Figure 3.4: Gene enrichment diagram

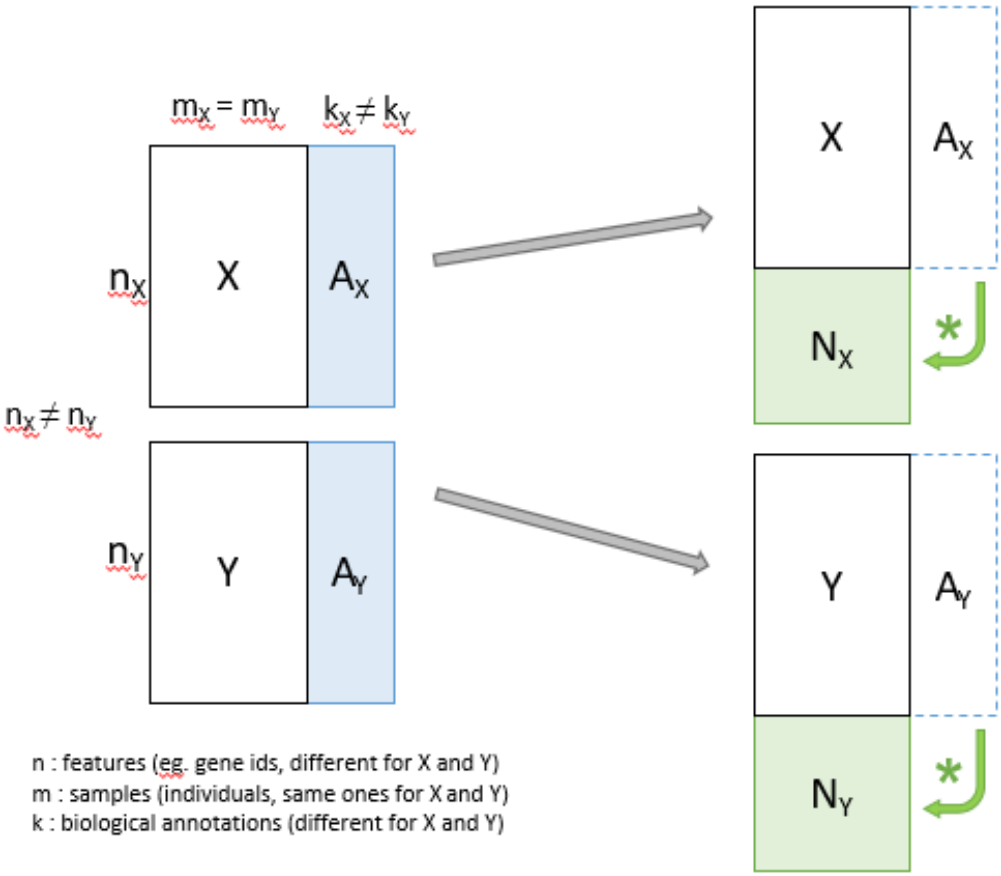

Figure 3.5: Matrix expansion diagram

$$N_G = \phi(G, A_G), \quad G \in X, Y, \ldots,$$

$$N_X = \frac{1}{r_X}(X \times A'_X),$$

$$N_Y = \frac{1}{r_Y}(Y \times A'_Y),$$

Figure 3.6: Addition of new feats (2)

of each individual for the variables annotated to each category. This is done with the matrix product of the initial numerical values (expression, proteins...) with the transposed matrices of their annotations, and then with the inverse matrix of a diagonal matrix of the count of how many annotations each category or entity annotated has had.

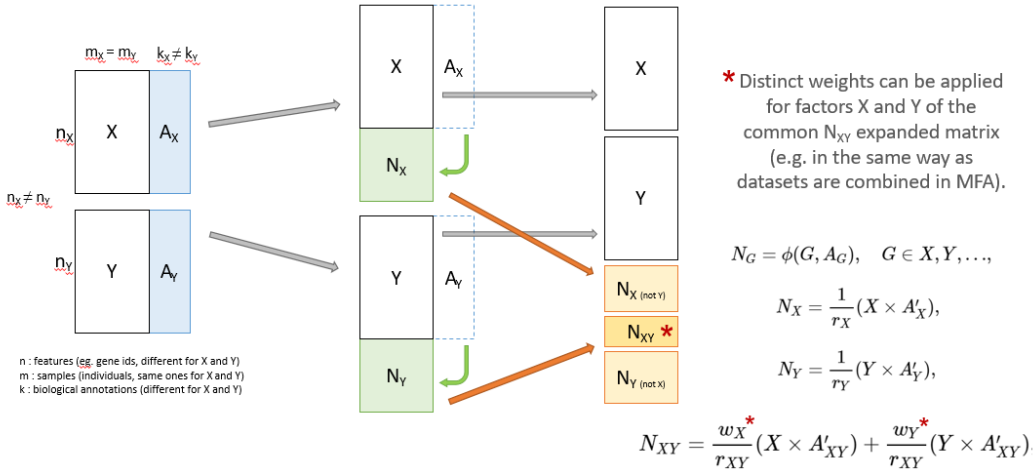

Figure 3.7: Matrix expansion diagram (2)

3.2.3 Integrative Analysis with Joint Dimension Reduction Techniques

To uncover meaningful insights from the expanded data sets and extract relevant information from the integrated omics and biological knowledge, contrasted joint dimension reduction techniques were employed. These techniques enable the simultaneous analysis of multiple data types and facilitate the identification of key patterns and relationships. The following methods were applied:

- Multiple Factor Analysis (MFA): MFA, adapted for multi-omics data, was utilized to identify sources of variability in the integrated dataset while considering both quantitative omics data and biological annotations. MFA aims to maximize relevant information within the data while accounting for the hierarchical structure of the biological knowledge.

- Multiple Co-Inertia Analysis (MCIA): MCIA, a technique that aligns the covariance structures of multiple datasets, was employed to explore relationships between omics measurements and biological annotations. MCIA seeks to identify common patterns and associations between these data sources.

- Regularized Generalized Canonical Correlation Analysis (RGCCA): RGCCA was used to identify latent variables that capture joint information from

omics data and biological annotations. RGCCA extends canonical correlation analysis to handle multi-view data integration and helps reveal correlated features across datasets.

PUNTS A INCLOURE:

- Reducció de dimensió. Anàlisi factorial en detall (MFA), + MCIA + RGCCA
- incloure aquí % variabilitat explicat segons la estructura de la intersecció de les 2 taules (article Lovino 2022)
- avantatge del MFA és que podem definir blocs de variables!
- no mirem unicament si guanyem variabilitat, sino tambe si millorem interpretabilitat biologica

3.2.4 Semi-Automation using the Targets R Package

The semi-automation of the integrative analysis process was facilitated by leveraging the Targets R package, which provides an efficient and user-friendly framework for building and managing complex analysis pipelines. In the development of the Targets pipeline, careful management of functions and parameters was essential to ensure a systematic and reproducible workflow. The following principles were applied:

- Function Modularity: Functions within the Targets pipeline were designed to be modular, focusing on specific tasks or analyses. This modularity enhanced code readability and maintainability.
- Parameterization: Parameters for each function and analysis step were carefully defined, allowing for flexibility and adaptability in the pipeline. This parameterization enabled the adjustment of analysis settings without modifying the underlying code.

- Dependency Management: Dependencies between different analysis steps were explicitly defined within the pipeline. This ensured that each step was executed in the correct order, and dependencies were automatically managed by the Targets package.

- Error Handling: Error handling procedures were implemented to capture and address potential issues during pipeline execution. This included the ability to handle errors, retries, and reporting of errors for troubleshooting. (NO APLICAT ARA PER ARA!)

PENDENT A AMPLIAR:

- Introduccio al paquet Targets en general i de les seves caracteristiques...

The R 'targets' package is a powerful tool for building and managing data science and data analysis pipelines. It is primarily designed for workflow automation, dependency management, and parallel processing in R projects. This package is useful for the following purposes:

1. Define and Manage Workflows: You can create a directed acyclic graph (DAG) that represents the workflow of your data analysis or machine learning project. Each node in the graph corresponds to a target, which can be a data file, an R script, or any other computational task.

2. Manage Dependencies: 'targets' allows you to specify dependencies between targets, ensuring that tasks are executed in the correct order. If a target depends on another target, it won't be executed until its dependencies are up-to-date.

3. Parallel Processing: One of the strengths of 'targets' is its ability to parallelize tasks. It can automatically determine which targets can be executed concurrently, improving the efficiency of your workflows, especially when working with large datasets or computationally intensive tasks.

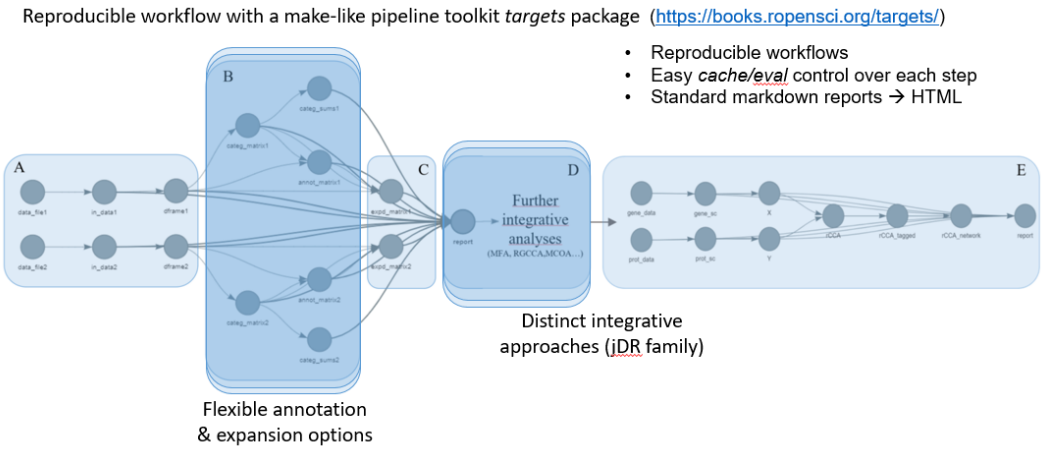

Figure 3.8: Workflow overview

4. Incremental Builds: When you make changes to your code or data, 'targets' can identify the minimal set of targets that need to be recomputed, saving time and computational resources. This is particularly useful for iterative development and experimentation.

5. Reports and Logging: 'targets' provides tools for generating reports and logging the progress of your workflow, making it easier to track and document your work.

6. Integration: It can be seamlessly integrated with other R packages and tools, such as 'drake' for more advanced data workflow management.

So, the 'targets' package is especially valuable for projects where data processing is a significant component, and you need a structured way to manage the various steps of your analysis or modeling pipeline. It helps ensure that your analyses are reproducible, efficient, and well-documented.

- Sistema que hem aplicat per crear el pipeline amb Targets...

Targets workflow diagram (Figure 3.8) showing the steps corresponding with the complete process: The pipeline starts from (A) a couple of 'omics-derived input data sets (e.g. pre-processed gene expression and protein abundance matrices).

These are converted to R data frames with features in rows and samples in columns. Then, a data frame containing related annotations (B) is created, or loaded, for each given input matrix, and used to expand these original data, in order to end up with a pair of data frames (C) containing the original values plus the average expression/abundance values of the features related to each annotation as new features in additional rows. After that, distinct Dimension Reduction Methods are applied to perform the integrative analysis (D), and finally, an R markdown report (E) is rendered to show steps and main results of the full process.

4

Results

Contents

Text de presentacio dels resultats...

Fer que 4.1 sigui l'actual 4.2 (tota la info d'aplicar el mètode)

ESTRUCTURA DELS RESULTATS:

4.1 Resultats del nou mètode

4.2 Resultats d'implementació en paquet R amb targets

4.3 Exemples i aplicacions

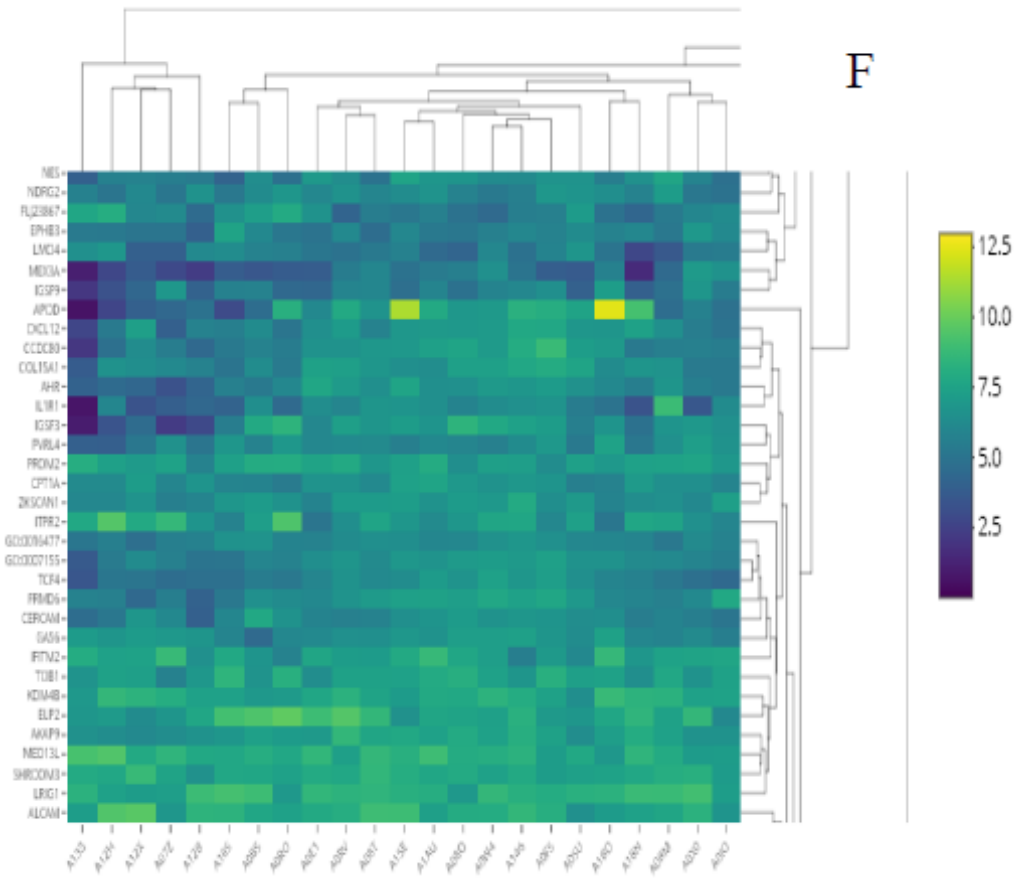

Figure 4.1: Heapmap of an expanded matrix

4.1 Results from the analysis of human brain tissue samples

4.2 Results from the expansion of omics data with biological annotations

Figure 4.1 is an snapshot (F) of one of the heat maps created to show the expanded matrices obtained in (Figures 3.5 i 3.6 prèvies, de Methods).

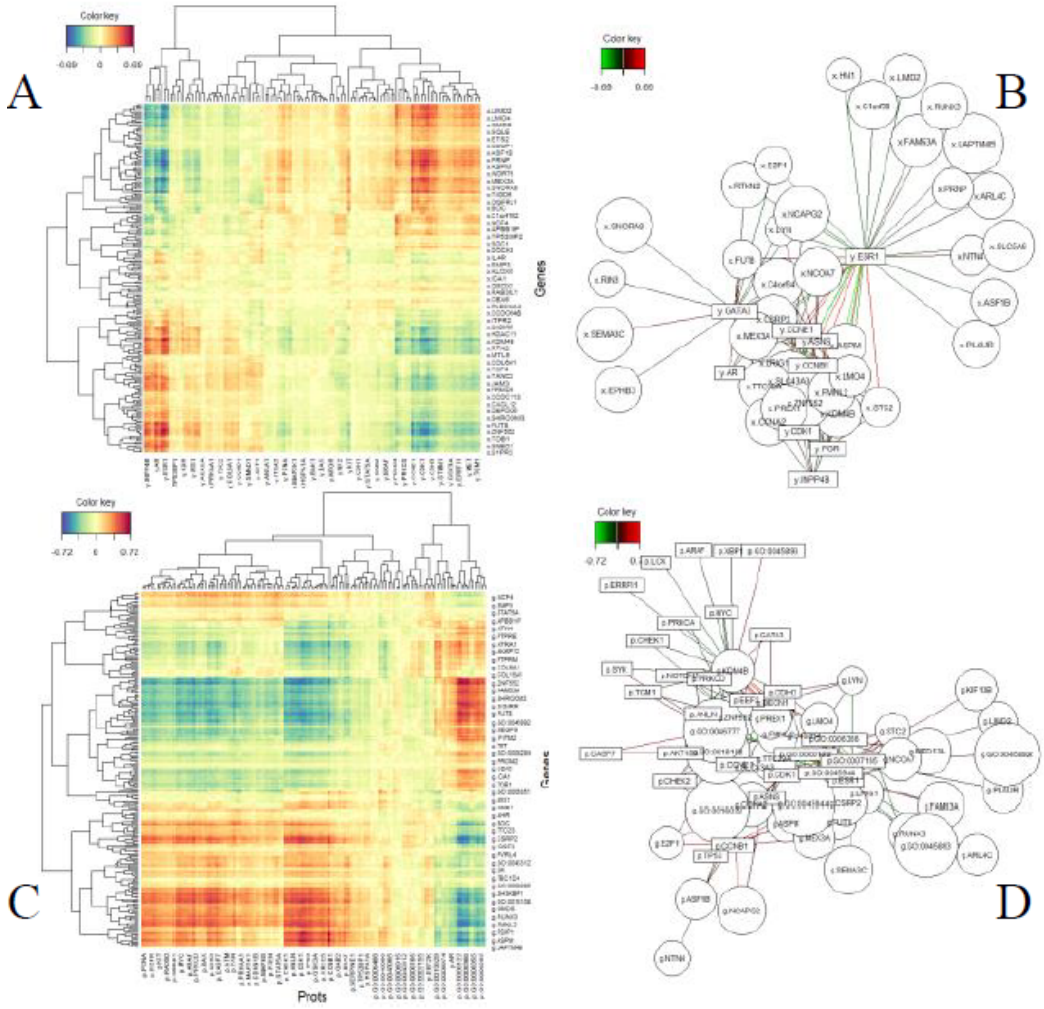

Figure 4.2: BRCA results overview

4.3 Results from the analysis of 150 TCGA-BRCA samples

Figure 4.2 contains some of the graphical results of the analysis of the 150 samples from TCGA-BRCA: Heat maps (A, C) and association networks (B, D) resulting from the integration by Regularized Canonical Correlations Analysis with mixomics R package. Performed with the original data sets (A, B) or using data expanded with biological annotations to Gene Ontology (C, D), so adding some GO terms to the features from each source, where the outputs contain higher level of information (higher density in both type of plots).

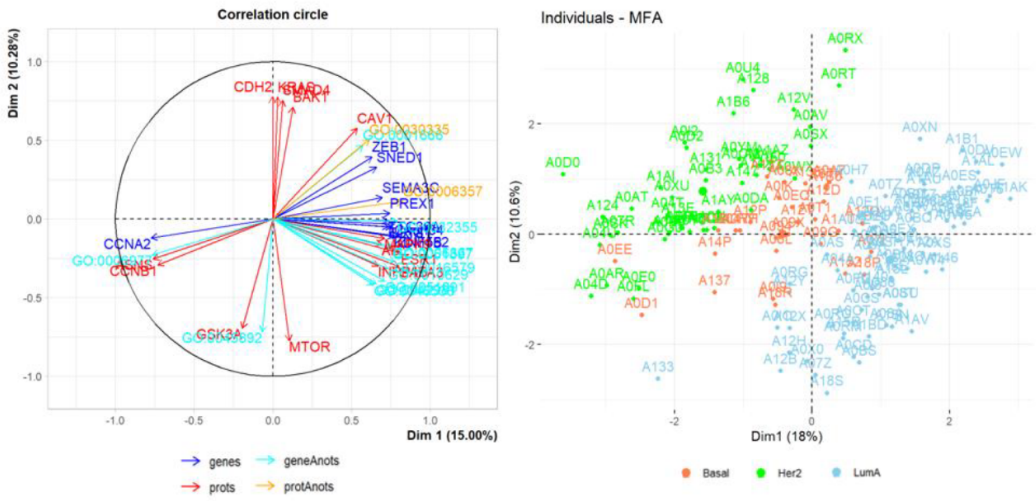

Figure 4.3: BRCA results with MFA

4.4 Results from the application of MFA on TCGA-BRCA data with, and without, expanded data

Figure 4.3 includes a Correlation Circle (left), with most relevant genes, proteins and added GO annotations. Distribution of samples (right) along the first two plotted dimensions. Both results coming from the application of Multiple Factor Analysis (FactoMineR and factoextra R packages) performed on the same 150 samples (Basal, Her2 and LuminalA conditions) from TCGA-BRCA.

4.5 Resultats de la creacio del paquet amb Targets...

5 Discussion

Potser no cal posar la TOC aqui£?

Resum de l'article. Apuntant a les conclusions. Comentant problemes i limitacions (emprar combinacions lineals de variables per crear-ne de noves).

Possibles extensions [punts de millora] Comentar i descriure cadascun d'ells:

- Poder fer servir 3 o més conjunts de dades
- Poder ponderar els pesos de les anotacions, segons tipus, data set d'origen, etc.
- Permetre treballar amb dades faltants o, fins i tot, blocs de dades faltants.
- Millorar les opcions del paquet: mètodes d'anotació bio, mètodes d'integració, tipus de gràfics resultants...

There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved.

— Darwin's *On the Origin of Species* (1859).

6 Conclusions

If we don't want Conclusion to have a chapter number next to it, we can add the `{-}` attribute.

Conclusion 1

The need for a better biological interpretation of multi-omics integrative methods let us to consider the inclusion of biological information during (not after) the analysis process

Conclusion 2

We propose a method focused on the expansion of the starting omics datasets, by adding new annotation-derived features to those matrices, before applying the integrative analysis

Conclusion 3

This approach allows the inclusion of relevant information from the main biological annotation tools, as well as any custom annotation, combined with the use our preferred Dimension Reduction techniques

Conclusion 4

We have implemented a pipeline for reproducible and easy-to-use execution, that facilitates the control of each step, the visualization of results and their reporting to PDF/HTML formats.

Appendices

A

The First Appendix

This first appendix includes an R chunk that was hidden in the document (using `echo = FALSE`) to help with readibility:

In 02-rmd-basics-code.Rmd

And here's another one from the same chapter, i.e. Chapter ??:

B

The Second Appendix, for Fun

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Briansó, F., García-Berrocoso, T., Montaner, J., & Sánchez-Pla, A. (2016a). Integrative analysis of transcriptomics and proteomics data for the characterization of brain tissue after ischemic stroke. *XXVIIIth International Biometric Conference IBC2016*.

Briansó, F., García-Berrocoso, T., Montaner, J., & Sánchez-Pla, A. (2016b). Multivariate methods for the integrative analysis of transcriptomics and proteomic data in a study on ischemic stroke. *The 15th European Conference on Computational Biology ECCB*.

Briansó, F., García-Berrocoso, T., Montaner, J., & Sánchez-Pla, A. (2017). Integrative Analysis of Transcriptomics and Proteomics Data for the Characterization of Brain Tissue After Ischemic Stroke. In E. A. Ainsbury, M. L. Calle, E. Cardis, J. Einbeck, G. Gómez, & P. Puig (Eds.), *Extended Abstracts Fall 2015* (pp. 21–27). Springer International Publishing. https://doi.org/10.1007/978-3-319-55639-0_4

Busold, C. H., Winter, S., Hauser, N., Bauer, A., Dippon, J., Hoheisel, J. D., & Fellenberg, K. (2005). Integration of GO annotations in Correspondence Analysis: Facilitating the interpretation of microarray data. *Bioinformatics*, *21*(10), 2424–2429. https://doi.org/10.1093/bioinformatics/bti367

Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E., & Baudot, A. (2021). Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature Communications*, *12*(1), 124. https://doi.org/10.1038/s41467-020-20430-7

Cavill, R., Jennen, D., Kleinjans, J., & Briedé, J. J. (2016). Transcriptomic and metabolomic data integration. *Briefings in Bioinformatics*, *17*(5), 891–901.

https://doi.org/10.1093/bib/bbv090

Cisek, K., Krochmal, M., Klein, J., & Mischak, H. (2016). The application of multi-omics and systems biology to identify therapeutic targets in chronic kidney disease. *Nephrology Dialysis Transplantation*, *31*(12), 2003–2011. https://doi.org/10.1093/ndt/gfv364

Culhane, A. C., Perrière, G., & Higgins, D. G. (2003). Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics*, *4*(1), 59. https://doi.org/10.1186/1471-2105-4-59

Culhane, A. C., Thioulouse, J., Perrière, G., & Higgins, D. G. (2005). MADE4: An R package for multivariate analysis of gene expression data. *Bioinformatics*, *21*(11), 2789–2790. https://doi.org/10.1093/bioinformatics/bti394

Flores, J. E., Claborne, D. M., Weller, Z. D., Webb-Robertson, B.-J. M., Waters, K. M., & Bramer, L. M. (2023). Missing data in multi-omics integration: Recent advances through artificial intelligence. *Frontiers in Artificial Intelligence*, *6*, 1098308. https://doi.org/10.3389/frai.2023.1098308

García-Berrocoso, T., Goicoechea, L., Simats, A., Briansó, F., Gonzalo, R., Martínez-Saez, E., Moliné, T., Sánchez-Pla, A., & Montaner, J. (2016). Exploring brain gene expression changes following ischemic stroke through microarrays. *X Simposi de Neurobiologia de La Societat Catalana de Biologia*.

García-Berrocoso, T., Simats, A., Briansó, F., Llombart, V., Hainard, A., Sánchez-Pla, A., Sanchez, J., & Montaner, J. (2017). Integrative analysis of transcriptomics and proteomics data for the molecular characterization of human brain after ischemic stroke. *28th Symposium on Cerebral Blood Flow, Metabolism and Function*.

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A., & Tegnér, J. (2014). Data integration in the era of omics: Current and future challenges. *BMC Systems Biology*, *8*(2), I1. https://doi.org/10.1186/1752-0509-8-S2-I1

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, *37*(1), 1–13. https://doi.org/10.1093/nar/gkn923

Krassowski, M., Das, V., Sahu, S. K., & Misra, B. B. (2020). State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. *Frontiers in Genetics*, *11*. https://www.frontiersin.org/articles/10.3389/fgene.2020.610798

Lê Cao, K.-A., Martin, P. G., Robert-Granié, C., & Besse, P. (2009). Sparse canonical methods for biological data integration: Application to a cross-platform study. *BMC Bioinformatics*, *10*(1), 34. https://doi.org/10.1186/1471-2105-10-34

Lenzerini, M. (2002). Data integration: A theoretical perspective. *Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 233–246. https://doi.org/10.1145/543613.543644

Little, R. J. A., & Rubin, D. B. (2002). Missing Data in Experiments. In *Statistical Analysis with Missing Data* (pp. 24–40). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119013563.ch2

M. Wheelock, Å., & E. Wheelock, C. (2013). Trials and tribulations of 'omics data analysis: Assessing quality of SIMCA-based multivariate models using examples from pulmonary medicine. *Molecular BioSystems*, *9*(11), 2589–2596. https://doi.org/10.1039/C3MB70194H

Meng, C., Zeleznik, O. A., Thallinger, G. G., Kuster, B., Gholami, A. M., & Culhane, A. C. (2016). Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics*, *17*(4), 628–641. https://doi.org/10.1093/bib/bbv108

Ramiro, L., García-Berrocoso, T., Briansó, F., Goicoechea, L., Simats, A., Llombart, V., Gonzalo, R., Hainard, A., Martínez-Saez, E., Canals, F., Sanchez, J.-C., Sánchez-Pla, A., & Montaner, J. (2021). Integrative Multi-omics Analysis to Characterize Human Brain Ischemia. *Molecular Neurobiology*, *58*(8), 4107–4121. https://doi.org/10.1007/s12035-021-02401-1

Rodriguez-Fernandez, S., Pujol-Autonell, I., Brianso, F., Perna-Barrull, D., Cano-Sarabia, M., Garcia-Jimeno, S., Villalba, A., Sanchez, A., Aguilera, E., Vazquez, F., Verdaguer, J., Maspoch, D., & Vives-Pi, M. (2018). Phosphatidylserine-Liposomes Promote Tolerogenic Features on Dendritic Cells in Human Type 1 Diabetes by Apoptotic Mimicry. *Frontiers in Immunology*, *9*, 253. https://doi.org/10.3389/fimmu.2018.00253

Rodríguez-Hernández, C. J., Mateo-Lozano, S., García, M., Casalà, C., Briansó, F., Castrejón, N., Rodríguez, E., Suñol, M., Carcaboso, A. M., Lavarino, C., Mora, J., & Torres, C. de. (2016). Cinacalcet inhibits neuroblastoma tumor growth and upregulates cancer-testis antigens. *Oncotarget*, *7*(13), 16112–16129. https://doi.org/10.18632/oncotarget.7448

Rohart, F., Gautier, B., Singh, A., & Cao, K.-A. L. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Computational Biology*, *13*(11), e1005752. https://doi.org/10.1371/jo

urnal.pcbi.1005752

Simats, A., Ramiro, L., García-Berrocoso, T., Briansó, F., Gonzalo, R., Martín, L., Sabé, A., Gill, N., Penalba, A., Colomé, N., Sánchez, A., Canals, F., Bustamante, A., Rosell, A., & Montaner, J. (2020). A Mouse Brain-based Multi-omics Integrative Approach Reveals Potential Blood Biomarkers for Ischemic Stroke. *Molecular & Cellular Proteomics: MCP*, *19*(12), 1921–1936. https://doi.org/10.1074/mcp.RA120.002283

Singh, A., Gautier, B., Shannon, C. P., Vacher, M., Rohart, F., Tebbutt, S. J., & Cao, K.-A. L. (2016). *DIABLO – an integrative, multi-omics, multivariate method for multi-group classification*. bioRxiv. https://doi.org/10.1101/067611

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

Tayrac, M. de, Lê, S., Aubry, M., Mosser, J., & Husson, F. (2009). Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach. *BMC Genomics*, *10*(1), 32. https://doi.org/10.1186/1471-2164-10-32

Tyekucheva, S., Marchionni, L., Karchin, R., & Parmigiani, G. (2011). Integrating diverse genomic data using gene sets. *Genome Biology*, *12*(10), R105. https://doi.org/10.1186/gb-2011-12-10-r105

Wang, F., Chen, C., & Wang, D. (2014). Circulating microRNAs in cardiovascular diseases: From biomarkers to therapeutic targets. *Frontiers of Medicine*, *8*(4), 404–418. https://doi.org/10.1007/s11684-014-0379-2

Wang, K., Huang, C., & Nice, E. (2014). Proteomics, genomics and transcriptomics: Their emerging roles in the discovery and validation of colorectal cancer biomarkers. *Expert Review of Proteomics*, *11*. https://doi.org/10.1586/14789450.2014.894466

Wanichthanarak, K., Fahrmann, J. F., & Grapov, D. (2015). Genomic, Proteomic, and Metabolomic Data Integration Strategies. *Biomarker Insights*, *10*(Suppl 4), 1–6. https://doi.org/10.4137/BMI.S29511

Wu, C., Zhou, F., Ren, J., Li, X., Jiang, Y., & Ma, S. (2019). A Selective Review of Multi-Level Omics Data Integration Using Variable Selection. *High-Throughput*, *8*(1), 4. https://doi.org/10.3390/ht8010004