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

. . . . . .

Thanks to Ulrik Lyngs for providing the Oxford University Markdown template that I used for writing this thesis (Lyngs, 2019)

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

List of Figures

List of Tables

List of Abbreviations and Specific Terminology

Expanded Datasets . Referring **in this thesis** to the data matrices containing the original expression or quantification omics data as well as those numerical variables coming from the annotation of biological terms.

GO . . . . . Gene Ontology [https://geneontology.org/].

KEGG . . . . . Kyoto Encyclopedia of Genes and Genomes [https://www.genome.jp/kegg/]

MCIA . . . . . Multiple Co-Inertia Analysis

MFA . . . . . Multiple Factor Analysis

microRNA . . . . . Micro RiboNucleic Acid

mRNA . . . . . Messenger RiboNucleic Acid

TCGA . . . . . The Cancer Genome Atlas project [http://cancergenome.nih.gov/].

TCGA-BRCA . . . . . The Cancer Genome Atlas - BReast CAncer project [https://portal.gdc.cancer.gov/projects/TCGA-BRCA].

rCCA . . . . . Regularized Canonical Correlation Analysis

RGCCA . . . . . Regularized Generalized Canonical Correlation Analysis

sPLS . . . . . sparse Partial Least Squares regression

1 Introduction

Contents

1.1 Content of the introductory text

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 multi-

ple 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

FFF:WORK IN PROGRESS

1.2.1 Omics data analyses

Omics data encompasses comprehensive information about a biological system, encompassing its entirety. The term "omics" originates from the Greek word "oma," meaning "a collection" or "a mass." Omics data is generated through high-throughput analysis technologies that enable the measurement of gene expression, protein composition, DNA structure, metabolism, and more.

Among the primary omics data types one can highlight the following:

- Genomics: The study of the genome, which comprises the complete set of genes within an organism.
- Transcriptomics: The investigation of gene expression, focusing on the amount of messenger RNA (mRNA) produced from each gene.
- Proteomics: The examination of proteins, the molecules that carry out the majority of biological functions.

- Metabolomics: The study of metabolism, the ensemble of chemical reactions occurring within an organism.
- Epigenomics: The exploration of changes in gene expression that are not attributed to alterations in DNA sequence.

FFF:INSERIR AQUÍ IMATGE AMB ESQUEMA DE LA OMICS CASCADE

These layers provide complementary biological information and collectively offer a comprehensive view of biological systems.

Omics data integration stands as a complex endeavor requiring advanced statistical and computational methods. It is employed for a range of biomedical applications, such as identifying novel genes and proteins linked to diseases, developing new drugs, and enhancing diagnostic accuracy.

FFF: 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.

FFF: MES EL TEMA DE LA p»n (Dimensionallity Reduction Techniques; The p»n situation, ja en part superat)

1.2.2 The problem of data incompleteness

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 elements of interest. In this context, "missing data" refers to the absence of values for certain variables within a dataset, which can arise due to various reasons like experimental limitations or technical constraints (Hornung et al., 2024). On the other hand, "missing data types" specifically refer to situations where different omics data types are not available for all individuals in a study, leading to block-wise missing data patterns (Flores et al., 2023). Both situations, missing data values and missing data types, can hinder the integrative analysis and interpretation of multi-omics datasets.

Missing Data Values FFF: ampliar detalls

Missing Data Types FFF: ampliar detalls

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. Existing imputation methods face challenges in high-dimensional settings, potentially leading to biased findings (Harris et al., 2023). Different strategies, such as combining biospecimen data matrices for imputation, have been explored to preserve correlation structures in multi-biospecimen studies (Dai et al., 2022). Integrative imputation techniques leveraging correlations among multi-omics datasets are essential for accurate downstream analyses in multi-omics studies (Wilson et al., 2022). Novel multi-omics imputation methods have been proposed to integrate multiple correlated omics datasets, improving imputation accuracy and enhancing downstream analysis performance (Song et al., 2020).

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 methods have been proposed to handle missing data in omics datasets, such as imputation techniques(Zhou et al., 2023) (Buyukozkan et al., 2023), decision charts for selecting suitable imputation methods(Kong et al., 2022), and statistical models that directly incorporate missing values into calculations(Kidd et al., 2023). These approaches aim to mitigate the limitations posed by missing data, ensuring robust statistical analysis and interpretation of omics data(Lin et al., 2020). By leveraging advanced techniques like deep learning for multi-omics integration with incomplete data, researchers can enhance the understanding of complex biological processes and improve disease classification accuracy using incomplete multiomics data.

Uncertainty and Sensitivity Analysis: After imputation methods in omics data analysis, uncertainty and sensitivity analysis play crucial roles in assessing the reliability of the results. Various approaches have been proposed to address missing data issues, such as multiple imputation (MI) and nonparametric multiple imputation strategies. MI-MFA, a method combining multiple imputation with multiple factor analysis, has shown promising results in integrating incomplete datasets(Yin & Shi, 2019). Additionally, sensitivity analysis approaches have been developed using standardized sensitivity parameters to evaluate the impact of missingness mechanisms, particularly in cases of missing not at random (MNAR) data. These approaches involve selecting imputing sets based on predictive scores and sensitivity parameters, aiming to provide robust estimates even in the presence of misspecifications(Voillet et al., 2016) (Uranga et al., 2022). Overall, these methods contribute to enhancing the accuracy and reliability of omics data analysis post-imputation.

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.

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, vali-

dating 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.

FFF: PENDENT REVISAR REFS ANTERIORS

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.

FFF: PENDENT REVISAR REFS ANTERIORS

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.

FFF: PENDENT REVISAR REFS ANTERIORS

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

FFF: PENDENT DE PRIMERES NOTES

1.2.7 Motivation for Integrative analysis

FFF: POTSER POSAR-HO AL PRINCIPI DEL CAPITOL???

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.

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

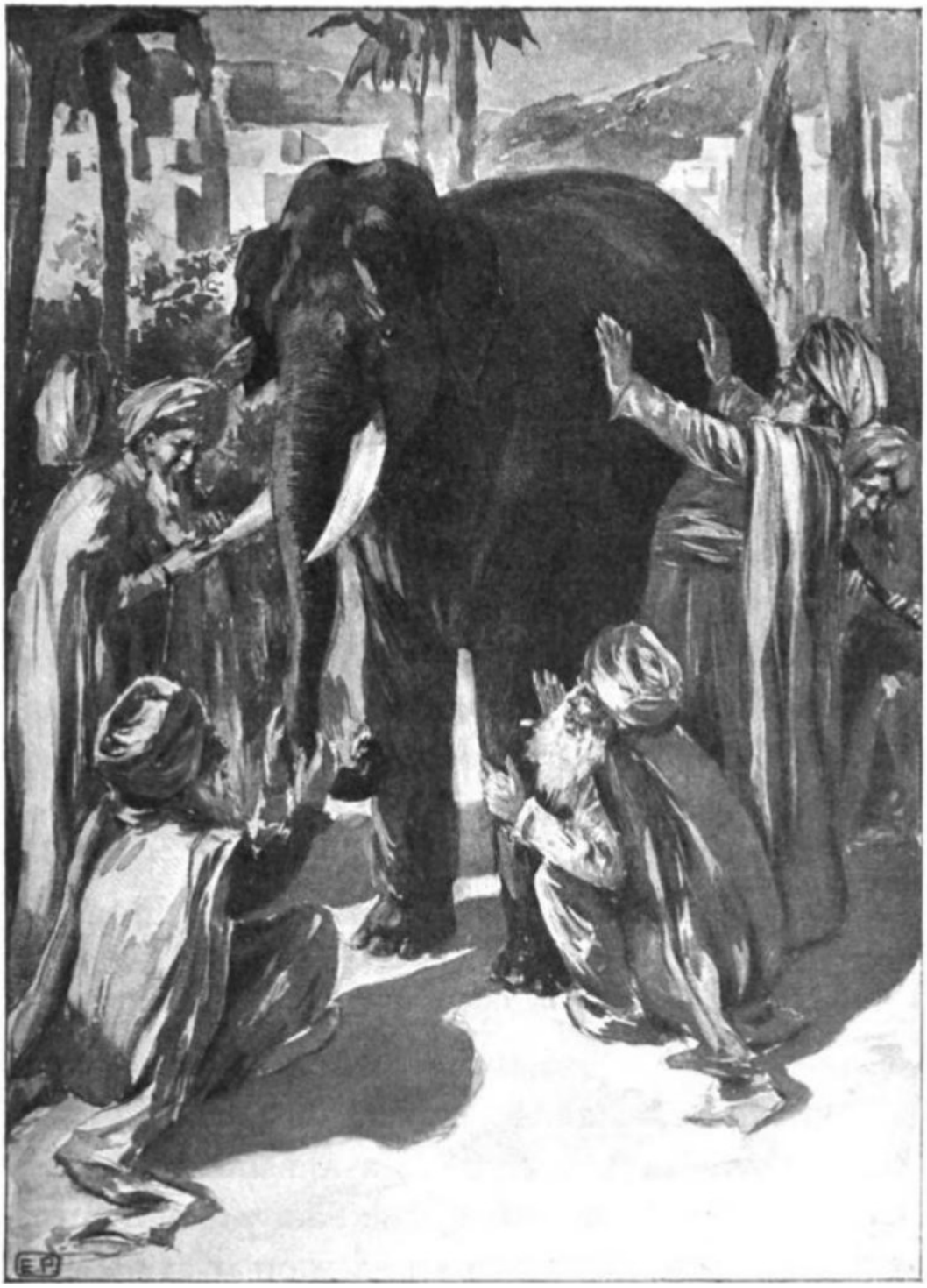

Figure 1.1: The blind men and the elephant. By Illustrator unknown - From The Heath readers by grades, D.C. Heath and Company (Boston), p. 69., Public Domain, https://commons.wikimedia.org/w/index.php?curid=4581263

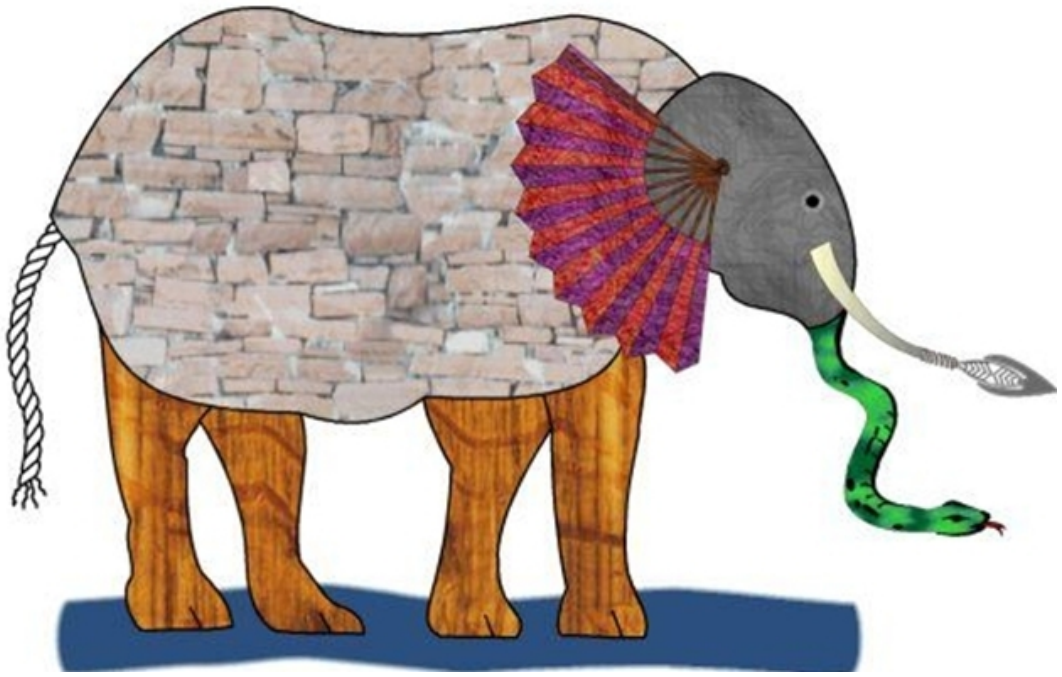

Figure 1.2: Reconstruction of the elephant as the blind men perceive it. Image source: http://doug-johnson.squarespace.com/blue-skunk-blog/2012/12/8/the-blind-men-and-the-elephant.html;jsessionid=30DE43866E1B453471B75CB39688E2CB.v5-web003

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

FFF: La reducció ve més motivada per la necessitat de destacar els aspectes més rellevants i de que aquests siguin més fàcilment interpretables

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

1.2.8 Existing approaches for multi-omics data integration

FFF: Maneres de reduir dimensió amb finalitat integració òmiques

MCIA, RGCCA, MFA... ...(Culhane et al., 2003) ...(Cavill et al., 2016) ...(Vahabi & Michailidis, 2022) ...(Wekesa & Kimwele, 2023) ...(Athieniti & Spyrou, 2023)

1.2.9 Revisió de mètodes de creació de pipelines

FFF: PENDENT INCLOURE PRIMERES NOTES

2
Objectives

Contents

2.1 Working phases (modificar titols)

The motivation for this thesis stems from my work and research experience at Vall d'Hebron Research Institute in Barcelona, where, with the aim of providing a useful tool for the interpretation of omics data in the field of biomedical research, the following phases of work were proposed:

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.

REVISAR i POTSER REESCRIURE AQUESTS PUNTS COM A UN SOL BLOC?

Here is a brief description of the main activities that derived from the initially proposed phases, the methods on which they were based, the objectives with which they were related, as well as some of their results, which will be discussed in more detail in subsequent chapters.

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. (Rohart et al., 2017), 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 (Pucher et al., 2019) and Martin (Martin et al., 2010), going from simple Venn diagrams to more complex, network analysis, software such as some specific R packages (R Core Team, 2022) or Cytoscape (Cline et al., 2007). 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 (Landau, 2021), 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 (Huber et al., 2015), and, finally, to complete objective 4 of this thesis plan, as a web application (Chang et al., 2021) to be used in further steps of the current biomedical research projects in which our collaborators are implied, as well as in future studies.

2.2 Main objectives of this work

In light of the challenges presented in the previous point, the main objectives of this thesis were established as follows:

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

In the context of multi-omics data integration, our proposal relies on the idea that incorporating biological annotations into datasets before proceeding with integrative analysis enriches the outcomes and enhances their biological interpretability. Therefore, augmenting quantitative omics data with contextual biological knowledge will deepen our understanding of complex biological phenomena. To do so, 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 to illuminate hidden patterns and relationships between data sources or

blocks, and the semi-automation capabilities of the Targets R package allow us to build an easy-to-use implementation of the whole process.

3.1 Data Format Review and Quality Assessment

Before initiating the integrative analysis, a meticulous evaluation of data quality and format compatibility was conducted to ensure the reliability of the input datasets. This crucial step aimed to identify and rectify discrepancies, inconsistencies, or errors that could potentially impact subsequent analyses. During this process, datasets spanning various omics technologies, including transcriptomics and proteomics, are selectively acquired from reputable sources and repositories. Emphasis was placed on meticulous source selection to guarantee consistency and adherence to standardized formats. Subsequently, the raw omics data underwent a comprehensive preprocessing phase, addressing issues such as missing values, outliers, and normalization. This preprocessing step was indispensable for enhancing data quality and enabling comparability across diverse datasets. Additionally, a thorough review of data formats encompassing file types, column naming conventions, and units of measurement was conducted. Non-standardized data were systematically transformed into a uniform format to streamline the downstream integration processes. Through these procedures, a robust foundation was established for subsequent integrative omics analysis, ensuring coherence and validity of the synthesized biomedical insights.

Data, whether obtained directly from TCGA or from specific data sets used as examples in specific R packages[data source: http://mixomics.org/mixdiablo/diablo-tcga-case-study/], has to be reviewed by performing a basic descriptive analysis, as is customary in single-omics data studies.

Data source: The Cancer Genome Atlas Network (Network et al., 2012) veure (Koboldt et al., 2012) data original source: https://portal.gdc.cancer.gov/projects/TCGA-BRCA

BRCA subtypes main ref: https://www.pnas.org/doi/full/10.1073/pnas.191367098 veure(Sørlie et al., 2001)

The Cancer Genome Atlas, often abbreviated as TCGA, is a landmark project funded by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) in the United States. It's a joint effort to comprehensively understand the molecular basis of cancer by using genome analysis technologies. TCGA curates a vast collection of real-world data, covering diverse cancer types. This data encompasses omics datasets, such as transcriptomics and proteomics, offering researchers a comprehensive perspective on the molecular landscape of tumors. Stringent quality control and standardization measures ensure the data within TCGA is compatible and readily integrable with other datasets. This facilitates robust multi-omics analyses, and such approach ensures data Compatibility, given that standardized formats eliminate inconsistencies between datasets generated using different platforms or technologies. This also allows for seamless integration of data from various sources, crucial for multi-omics analyses like the one presented in this thesis. The integration of transcriptomics and proteomics data, heavily relies on the ability to combine these kind of datasets for a more holistic understanding. Standardized formats within TCGA remove a significant hurdle in this process. In addition to that, standardized data collection and processing protocols minimize technical variations that could introduce bias into the data, therefore enhancing the reliability and generalizability of findings derived from TCGA records. The public availability of TCGA data empowers researchers worldwide to leverage this rich resource in their investigations and, consequently, this fosters open science collaborations and accelerates advancements in cancer research.

In the vast majority of cases, the information is structured directly in tables or matrices (where, for example, the columns contain patient samples or experimental individuals, while the rows represent the values of the measured features). These matrices can be encapsulated in structures such as Expression Sets or similar, where

information about the omic measurements is accompanied by metadata related to the samples themselves or to the details of the technology used for the analysis.

FFF:Nota d'exemple de dades *Samples in rows (200G, 111P); Features in columns (150S). Tenim dimensions de 200 x 150 per una banda i 111(154 originalment) x 150 per l'altra.*

It is relatively frequent that the datasets available for multi-omics studies present information on the same samples, analyzed by means of two or more different technologies (e.g., microarrays or RNA-seq for mRNA gene expression, plus quantification for proteins) while the information related to the omic molecules analyzed is, obviously, different. Not only that, but it is also common that there is no direct mapping between the different types of features, so that, for example, not all genes analyzed in an RNA expression experiment are unambiguously represented by their corresponding proteins.

The task of recognizing the labels of the molecules analyzed in each dataset and consequently determining which ones are suitable for proceeding with the integrative analysis is not a light one. It often requires to implement a semi-automatic general identification of their names or ID codes, followed by a validation and filtering of the resulting non-obvious cases. If, in addition, the integrative analysis aims to map the different omics in some way at the biological process level (e.g., microRNAs against their target genes), then we are faced with an additional challenge of critical importance for the rest of the process.

Mostrar Figure 3.2 i Figure 3.1 PERO POTSER MILLOR COM A TAULES INTEGRADES AMB MARKDOWN?

	A0FJ	A13E	A0G0	A0SX	A143	A0DA	A0B3	A0I2	A0RT
YWHAE	0.049130778	-0.079982106	-0.032849886	-0.205329492	0.060190211	0.030761714	-0.107861537	0.64984396	-0.013650441
EIF4EBP1	0.447486231	0.605218418	0.894609732	-0.141322924	0.131768992	0.032996799	-0.037124691	-0.52148657	-0.634850633
TP53BP1	0.917834192	0.059101206	0.517044530	-0.313728669	0.330912383	-0.220271002	-0.544743061	-1.60203535	-0.720723295
ARAF	0.022741468	-0.459852981	-0.191821916	-0.074823472	-0.024357467	0.418616650	0.430503500	-0.18714658	-0.374882996
ACACA	-0.086267822	-0.592691835	0.411171898	-0.851480596	0.769751430	-0.714308701	-0.363474049	1.07761482	-1.254491083
ACCB	-0.416624416	-0.062268404	0.825828592	-0.663410436	0.873478702	-0.217526770	-0.269313837	1.58998239	-0.901353585
PRKAA1	0.285270389	-0.275233600	0.067741840	0.029563729	-0.216531821	-0.063065064	-0.077581092	-0.07753959	-0.177636653
ANLN	0.172311102	0.222105981	0.121993985	1.054948103	0.013784220	0.060256895	0.008872461	-0.05187936	-0.041880238
AR	-1.307605693	-1.620475956	-1.077894436	-1.267054694	-0.601327437	-1.208038484	-1.016297633	-0.42122691	-0.952324860
ARID1A	0.505094485	0.339581595	0.227180664	0.355297672	0.544125136	-0.110944799	-0.233223615	-0.35537533	-0.179195256
ASNS	0.811462882	1.181015791	1.950922363	0.607423831	0.538762877	0.311949453	1.138875941	-0.63275876	0.145464752
ATM	-0.495944728	-0.275533386	0.770857796	0.781328690	0.013854306	0.071748319	-0.209624373	-0.92406461	0.833870191
AKT1	-0.001377255	-0.755547887	-0.067397666	0.056726701	0.238114357	0.193712038	-0.301495924	-0.47402849	-0.367759411
ANXA1	-0.092909287	0.194749839	1.252992383	0.575274185	-1.557003586	0.491015188	0.533878400	1.21076392	0.424004827
BRAF	0.476309798	0.143257789	0.224891925	-0.221859607	0.248234872	-0.195445933	-0.036284702	-1.07351919	-0.711215072
BAK1	0.112201063	0.111310840	-0.069962738	-0.036546549	-0.124839115	-0.257300059	-0.115681609	0.87744695	0.183770057
BAX	-0.156538756	-0.205462637	-0.047604780	0.085173319	0.151544397	-0.090041106	-0.041475148	-0.54119284	0.146947246
BCL2	1.060203513	-0.160826453	-1.771917375	0.345023494	-1.588878871	-0.782913123	-1.041432134	0.41442932	0.531749294
BCLX	-0.100950513	-0.171629248	-0.056202128	-0.096473309	-0.140526557	-0.099476757	-0.037510164	0.74769887	-0.275295151
BECN1	-0.019449441	-0.041253352	-0.076969142	0.963238561	-0.219198072	-0.208762350	-0.219048417	-0.65145061	-0.165614955
BID	-0.034821157	-0.298426931	0.073740813	-0.203558523	-0.147058902	-0.035583612	-0.166370155	0.94513564	0.310251463
BCL2L11	0.408337983	-0.442249202	-1.244877548	0.163042908	0.051760204	-0.434277577	-0.290144108	-0.52809090	0.227648621
RAF1	0.108839334	0.403923023	-0.157470172	0.037683828	0.219029836	0.120775889	0.149760561	-0.42642058	-0.284617195
PECAM1	0.096913816	-0.135688779	-0.229473098	0.073040655	-0.037514094	-0.172646324	-0.003427764	-0.57845063	0.081700778
ITGA2	0.056953664	-0.369652041	0.225919578	-0.377061206	0.032209215	-0.279859151	0.064025012	0.20592716	-0.132198583
CDK1	0.391893475	0.431874216	0.170553859	0.487608500	0.356320675	0.057826839	0.201249969	0.15530981	0.238982271
CASP7	-0.209648791	-0.442253479	-0.027768363	0.619517693	-0.113207256	-0.533360022	1.021124691	0.62764170	1.765265466
CAV1	0.533755894	-1.310081134	-2.024819193	-0.105724014	-1.723398601	-1.761346920	-0.396679637	2.68650874	1.696473472
CHEK1	0.160404592	0.223441176	0.376873438	0.004513815	0.227372000	-0.069237193	0.204323902	0.48476037	0.190864987
CHEK2	1.056926875	0.651510308	0.881431094	0.222821482	0.425296288	-0.258980977	0.163256893	-0.70436969	0.476877281
CLDN7	-0.620225952	0.780008039	-0.343776287	0.228050453	0.233078487	0.040042353	-0.287622816	-1.39960030	-1.606562159
COL6A1	-0.869405130	-0.262350291	-0.425013922	-0.159521178	-0.805978550	-0.535507295	-0.513767940	0.97852799	0.835688459
CCNB1	1.516735476	1.025776946	0.977360144	0.569273220	1.368956673	0.071681473	1.237889430	-1.32406936	0.245716912
CCND1	-0.310524605	-0.434476563	-0.226412035	-0.512848244	-0.875023630	0.099692628	-0.359145590	1.01357985	0.247347773
CCNE1	0.987850528	0.249589732	-0.329458663	1.425506793	1.406639282	-0.042159529	0.516007883	0.38413664	0.193681101

Figure 3.1: Example of proteomics input data, viewed as a table in RStudio

```
##            A0FJ     A13E     A0G0     A0SX     A143     A0DA      A0B3     A0I2
## RTN2    4.362183 1.984492 1.727323 4.363996 2.447562 4.770798 3.3520618 1.810382
## NDRG2   7.533461 7.455194 8.079968 5.793750 7.158993 8.748061 5.0984040 3.791965
## CCDC113 3.956124 5.427623 2.227300 3.544866 4.691256 4.305401 0.5932056 2.719169
## FAM63A  4.457170 5.440957 5.543480 4.737114 4.808728 5.307480 5.2175851 4.355919
## ACADS   2.256817 4.028813 2.629855 4.269101 2.442135 3.239909 3.8851534 4.200249
## GMDS    6.017940 4.341692 6.363030 4.001104 7.029723 4.236539 5.9178858 4.830286
## HLA-H   5.006907 6.178668 6.039563 7.087633 5.936138 6.909727 8.0433411 9.130370
## SEMA4A  3.217812 2.864659 5.946028 5.007565 5.901459 6.591109 6.5328925 4.982386
## ETS2    4.734446 5.411029 5.651670 5.902449 6.641225 5.858016 6.3091167 5.304488
## LIMD2   5.099598 4.211397 3.304513 5.479451 5.508654 3.766283 4.1138727 5.149344
```

```
##               A0FJ        A13E        A0G0        A0SX        A143        A0DA        A0B3        A0I2
## YWHAE     0.04913078 -0.07998211 -0.03284989 -0.20532949  0.06019021  0.03076171 -0.107861537  0.64984396
## EIF4EBP1  0.44748623  0.60521842  0.89460973 -0.14132292  0.13176899  0.03299680 -0.037124691 -0.52148657
## TP53BP1   0.91783419  0.05910121  0.51704453 -0.31372867  0.33091238 -0.22027100 -0.544743061 -1.60203535
## ARAF      0.02274147 -0.45985298 -0.19182192 -0.07482347 -0.02435747  0.41861665  0.430503500 -0.18714658
## ACACA    -0.08626782 -0.59269183  0.41117190 -0.85148060  0.76975143 -0.71430870 -0.363474049  1.07761482
## ACCB     -0.41662442 -0.06226840  0.82582859 -0.66341044  0.87347870 -0.21752677 -0.269313837  1.58998239
## PRKAA1    0.28527039 -0.27523360  0.06774184  0.02956373 -0.21653182 -0.06306506 -0.077581092 -0.07753959
## ANLN      0.17231110  0.22210598  0.12199399  1.05494810  0.01378422  0.06025690  0.008872461 -0.05187936
## AR       -1.30760569 -1.62047596 -1.07789444 -1.26705469 -0.60132744 -1.20803848 -1.016297633 -0.42122691
## ARID1A    0.50509449  0.33958160  0.22718066  0.35529767  0.54412514 -0.11094480 -0.233223615 -0.35537533
```

Figure 3.2: Example of gene expression (above) and protein quantification raw data (below), viewed as loaded arrays in R

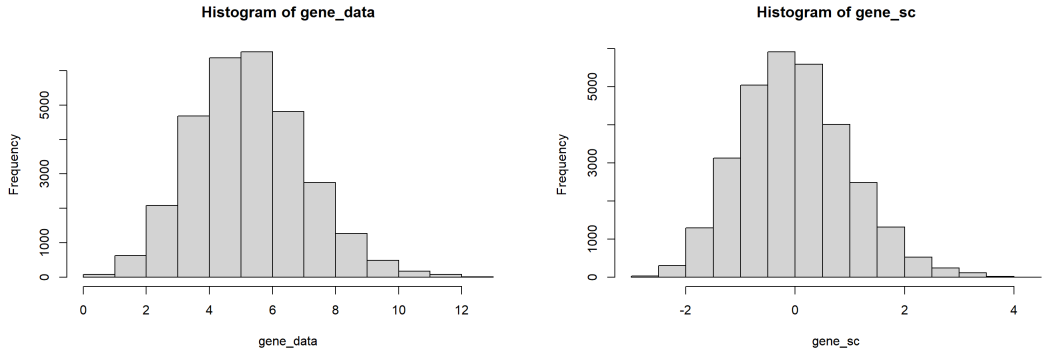

Figure 3.3: Histogram of the gene expression values coming from TCGA-BRCA dataset, before (left) and after (right) data centering

COMENTAR TAMBE REQUERIMENTS DE FORMAT (headers, value types...)

EXPLICAR QUALITY CHECKS APLICATS (grafiques per deteccio d'outliers, data centering...)

El proces s'ha de repetir, obviament, amb tots els input datasets.

3.2 Preprocessing for Integration of Biological Knowledge

The integration of biological knowledge into omics datasets can be achieved through a preprocessing step aimed at expanding the original data matrices with annotations accessed from specialized R libraries, which provided direct access to curated biological databases such as the Gene Ontology(Ashburner et al., 2000),(The Gene Ontology Consortium, 2019) and biochemical pathways information (e.g., KEGG(Kanehisa & Goto, 2000)). This process, that combines the annotation of the most significant biological entities with the quantification and integration of their annotation values to the data matrices, ends up with what we term "Expanded Datasets", which include the original biological features (e.g., gene expression or protein quantification values) as well as new variables coming from the annotation of biological terms. The following steps explain this preprocessing procedure in more detail:

- Selection of biological knowledge sources to feed annotations. Starting from the most commonly used biological knowledge databases in tasks of omics data analysis and interpretation, the goal is to access those that are most complete and applicable to the different types of data that must be annotated. GO and KEGG are excellent choices for that purpose.

- Selection of R packages specialized in the integration of biological information. The choice of the appropriate packages for the integration of biological information will depend on the specific needs of the project. In general, it is important to consider factors such as the type of data that will be integrated, the sources of the data, the integration methods that will be used and the desired level of complexity. In this case, it is appropriate to use R packages that can work with gene and protein identifiers reliably and completely, without adding too much complexity to the process.

- Data-Annotation Mapping. Each omics dataset is mapped to the biological information collected based on its identifiers (for example, gene or protein

names) using the capabilities of the selected R packages. This step facilitates the relationship of each of the elements of the omics data with the biological knowledge entities, creating certain temporary objects that collect the information of these links. So this step allows to relate the elements of omics data with biological knowledge entities, such as genes, proteins, metabolic pathways, etc. The mapping is performed using the identifiers of the elements of the omics data. For example, genes can be identified by their name, their symbol, or their Ensembl ID. Proteins can be identified by their name, their sequence, or their UniProt ID.

- Annotation Integration. The most relevant annotation elements resulting from the previous step can be integrated into the matrix structure of the original omics dataset that has been used for its biological annotation, resulting in an expanded data matrix that combines the initial quantitative omics measurements with new values associated with the biological annotations obtained in the process. This step is implemented by executing new R functions specifically developed for this purpose. The resulting data matrices (which contain the integration of the most relevant biological annotations) are called 'Expanded Matrices', and will be the basis for the subsequent application of integrative analysis methods of the different omics analyzed.

3.2.1 Selection of the sources for biological annotation

Selecting biological information sources for annotations hinges on various criteria: source credibility, information comprehensiveness, content currency, data format standardization, data accessibility, and potentially other relevant considerations like reputation and community acceptance. In the context of the present work, Source Credibility is the reliability and trustworthiness of the source providing the biological information; Comprehensiveness of Information is taken as the breadth and depth of the information contained within the source; Content Currency is the timeliness and up-to-date nature of the information provided; Data Format

Standardization is the adherence to standardized data formats for ease of integration and analysis; and Data Accessibility is considered as the ease of obtaining and accessing the data from each specific source. Other relevant criteria, such as community acceptance, is also a must to be taken into account.

Apuntar que es poden facilitar ja anotacions disponibles prèviament, sempre que compleixin amb el format que s'explica al següent apartat.

From the outset, our method was designed to allow the integration of user-prepared annotations, provided they adhered to the specified formats. This flexibility enabled users to leverage their existing knowledge and annotations, seamlessly incorporating them into our framework for enhanced analysis.

These annotations could be provided in either standard formats or customized according to user preferences. However, it is important to note that customized annotations that deviate from standard nomenclatures may not be compatible with certain functionalities of the proposed solution. Therefore, adhering to standard formats is recommended to ensure optimal compatibility and better integration.

3.2.2 Selection of the annotation packages

PAS QUE ES FA PRACTICAMENT AL MATEIX TEMPS QUE L'ANTERIOR

Destacar criteris de fiabilitat i senzillesa

Apuntar llista o referencia principal important

3.2.3 Biological annotations mapping

FFF: COM VAM PLANTEJAR fer l'anotació biologica. Quines opcions i amb quins metodes estadístics/bioinformatics... FFF: DUBTO SI LO QUE SEGUEIX NO ANIRIA A RESULTATS

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

(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 dataset, see Figure 3.4 for an example) annotated to them,

(ii) a Gene Enrichment Analysis (based on Hypergeometric tests against all GO categories, with FDR correction) is performed in order to retrieve the most relevant annotations to that set of genes/features.(Yu et al., 2012)

Figure 3.4 illustrates an example of a Gene List, where Gene Symbols serve as the standard nomenclature identifiers. These symbols represent unique identifiers for specific genes and are crucial for maintaining consistent gene identification across different species and databases. The use of standardized Gene Symbols ensures clear communication and facilitates data sharing among researchers. While figure 3.5 showcases the outcome of annotating the previous Gene List against the Biological Processes (BP terms) of the Gene Ontology (GO). This process involves assigning GO terms to each gene based on its known biological function. GO terms provide a structured and hierarchical vocabulary for describing gene functions, enabling researchers to effectively organize, analyze, and compare gene functions across various organisms.

```
  [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 3.4: List of gene symbols used as example

Annotated Matrix

Gene Ontology used: **BP**

Min. number of genes required to pass the filter: **8**

Annotated categories: **13** (for *data/mrna.csv*)

Annotated categories: **61** (for *data/prots.csv*)

Shared annotated categories: GO:0000122, GO:0006357, GO:0007155, GO:0007165, GO:0007411, GO:0008285, GO:0019221, GO:0030335, GO:0045893, GO:0045944

(Showing only partial output)

```
tar_read(categ_sums1)
```

```
## GO:0000122 GO:0006357 GO:0007155 GO:0007165 GO:0007411 GO:0008285 GO:0016477 GO:0019221 GO:0030335 GO:0043312
##         11         14         11         21         10          9         10         10          8          8
## GO:0045893 GO:0045944 GO:0055114
##         11         14         10
```

```
tar_read(categ_sums2)
```

```
## GO:0000082 GO:0000122 GO:0000165 GO:0000187 GO:0001525 GO:0001666 GO:0001701 GO:0001934 GO:0006357 GO:0006367
##          8         19         17          8         11          9         10         10         13         10
## GO:0006468 GO:0006915 GO:0006974 GO:0006977 GO:0007050 GO:0007155 GO:0007165 GO:0007169 GO:0007411 GO:0007507
##         24         17         14          8         10         11         30         11          8         12
## GO:0007568 GO:0008283 GO:0008284 GO:0008285 GO:0010468 GO:0010628 GO:0010629 GO:0016032 GO:0016579 GO:0018105
##         10         12         20         16          8         29         12         19         11         16
## GO:0018107 GO:0018108 GO:0019221 GO:0030154 GO:0030335 GO:0032355 GO:0032869 GO:0033138 GO:0033674 GO:0035556
##         11         11         15         11         11         10          9          9          8         17
## GO:0042060 GO:0042127 GO:0042493 GO:0042981 GO:0043065 GO:0043066 GO:0045471 GO:0045892 GO:0045893 GO:0045944
##          9          9         23         10         11         31          8         11         23         30
## GO:0046777 GO:0048538 GO:0050821 GO:0051091 GO:0051897 GO:0070374 GO:0071456 GO:0090090 GO:0098609 GO:1901796
##         10          8          8          8         12          9         11          8          9          9
## GO:2001244
##          8
```

Figure 3.5: Example of basic Go annotation by raw count against GO Biological Processes, setting 8 as minimum number of genes included in the BP entity. Annotation performed separately for gene expression and protein quantification input files

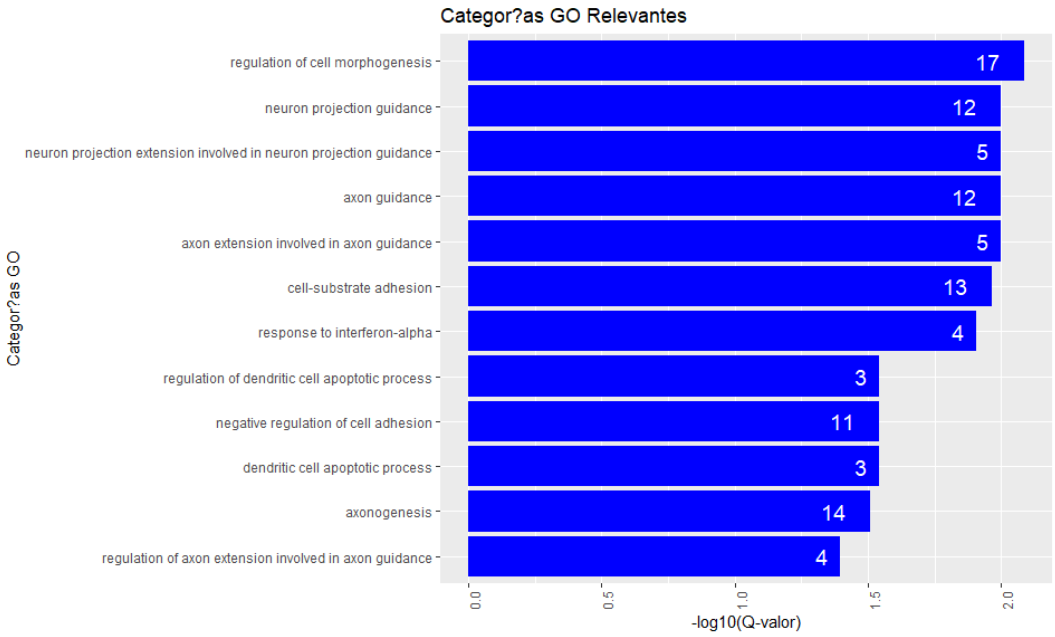

Figure 3.6: Example of results from GO annotation. Results of the biological significance analysis performed with the lists of genes against GO through clusterProfiler

In figure 3.5 we can see the number of Go terms resulting from the the process, given the specified threshold, for both gene expression and protein data sets, as well as the list of GO Ids returned as result (such as "`GO:0007155`", which refers to "cell adhesion" process and is present in the GO terms returned for both data types, with 11 target gene symbols annotated in both cases).

Figure 3.6 presents a graphical representation of the results obtained by processing the same gene list used in the previous examples with the R package clusterProfiler(Yu et al., 2012). This visualization showcases the outcomes of the biological significance analysis performed against the GO categories. clusterProfiler is an R package that automates the process of biological-term classification and the enrichment analysis of gene clusters. The analysis module and visualization module were combined into a reusable workflow, and currently supports three species, including humans, mice, and yeast. In this case, the graphical representation effectively communicates the identified enriched GO terms and their associated significance levels, providing valuable insights into the underlying biological processes associated with the gene list.

FFF: [mostrar formula estadística emprada per les anotacions]

FFF: [punt de millora, que l'anotacio basica pugui ser tb a KEGG]

FFF: COMENTAR AQUI l'opció d'afegir les anotacions com a individus suplementaris enlloc de variables£?

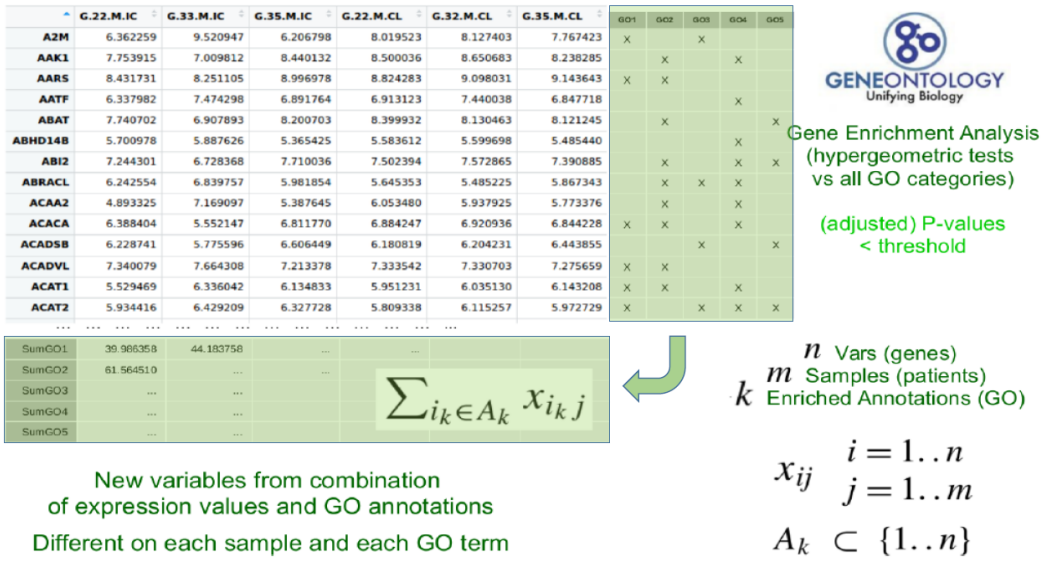

Figure 3.7: Addition of GO terms

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

FFF: [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.

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

FFF: EXPANSIO DE LES MATRIUS (numeritzar anotacions, creació de noves vars a partir de les anotacions)

The process starts already having a couple of datasets from distinct 'omics sources [punt de millora: admetre 3 o + inputs, comentar més tard a Discussion], mapped to gene ids (in the default case, where GO annotation have been performed), containing the results from a selection of differentially expressed genes or most relevant proteins analysis, or similar.

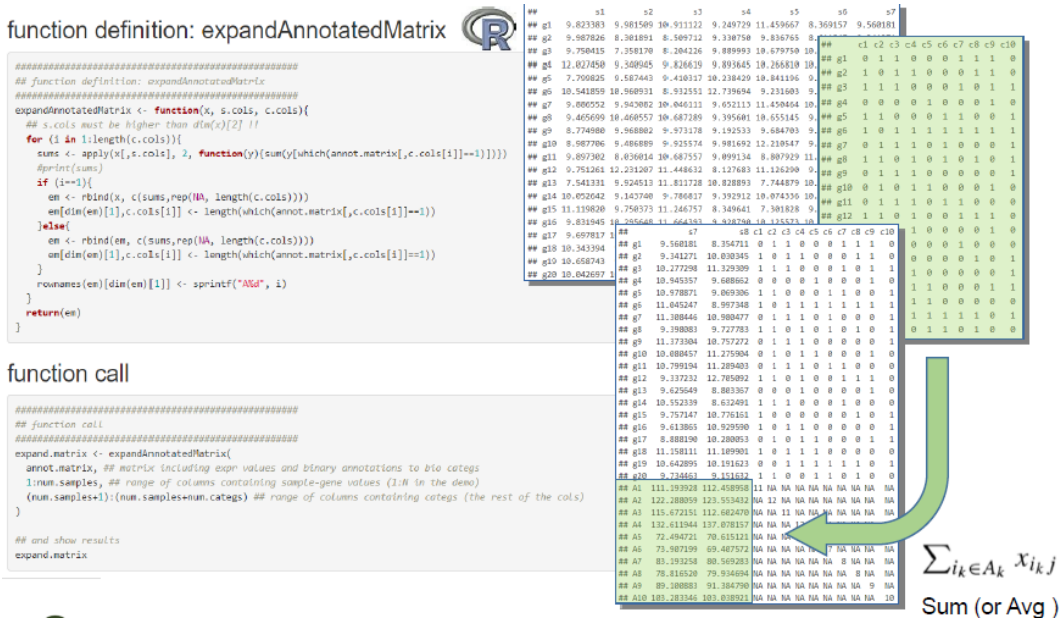

Figure 3.8: Addition of news feats

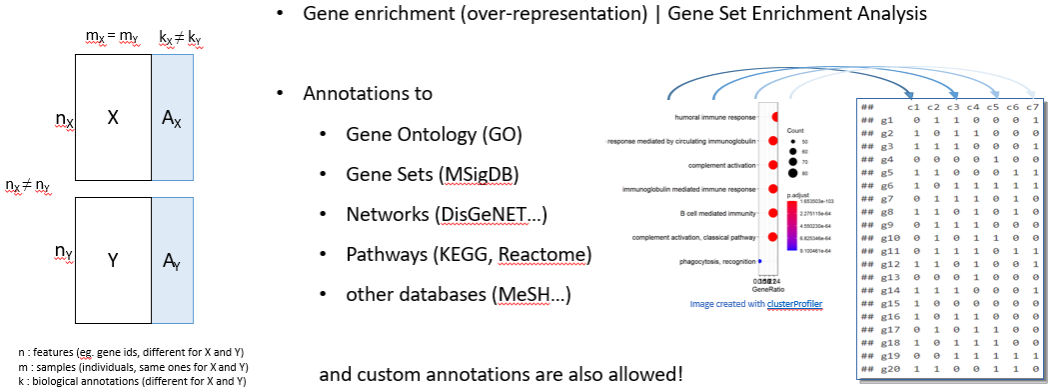

Figure 3.9: Gene enrichment diagram

3.2.4 Annotation Integration

Once the annotations are already computed, mapping each feature of the input dataset to the corresponding biological entity, they can be used to generate new 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 3.10, highlighted in blue) we proceed to generate the Expanded matrices (in green) by casting these annotations

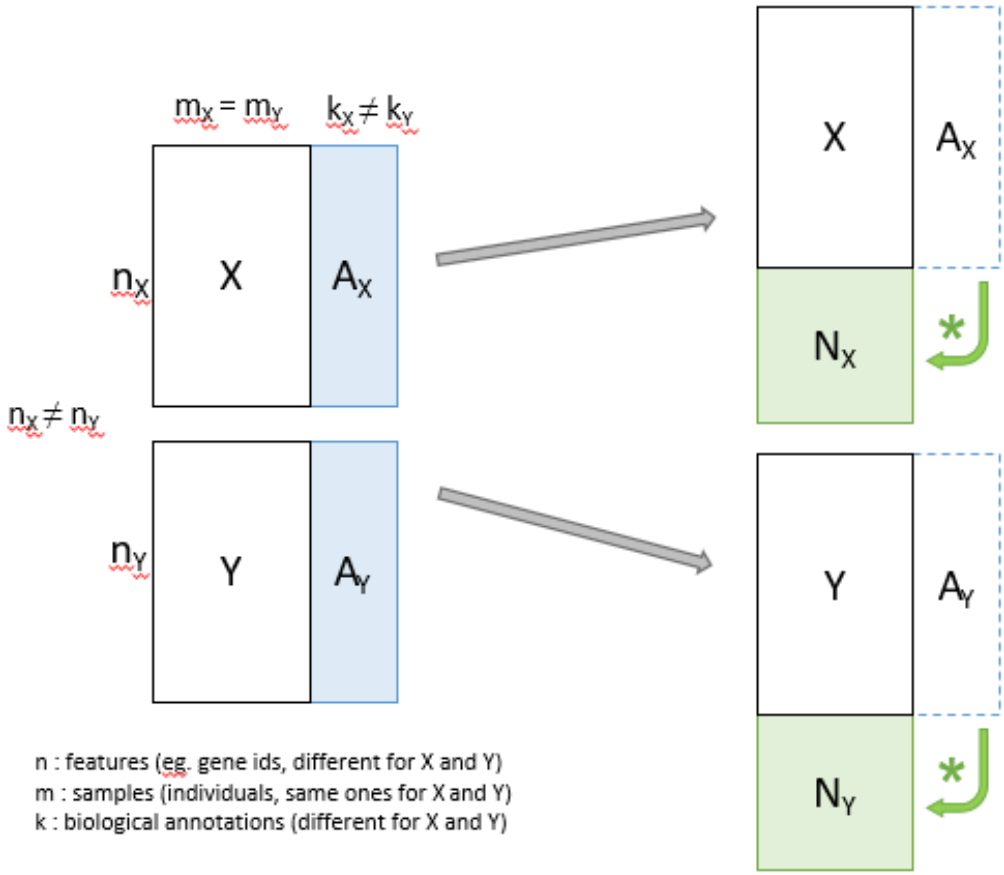

Figure 3.10: Matrix expansion diagram

as numerical values, that is, calculating the average of the numerical expressions 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.

3.3 Integrative Analysis with Joint Dimension Reduction Techniques

To uncover meaningful insights from the expanded datasets and extract relevant information from the integrated omics and biological knowledge, contrasted joint dimension reduction techniques were employed. These techniques enable the si-

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

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

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

Figure 3.11: Addition of new feats (2)

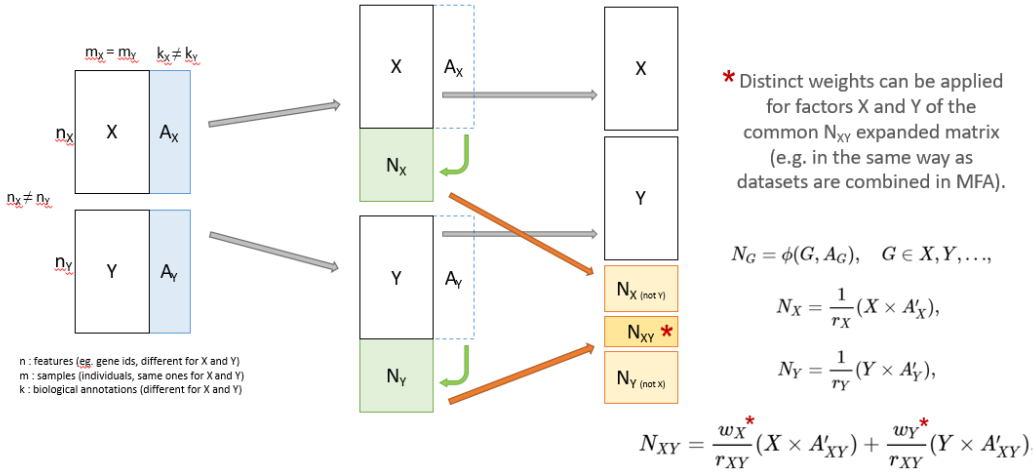

Figure 3.12: Matrix expansion diagram (2)

multaneous 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 data sets.

FFF: 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 (Lovino et al., 2021)
- avantatge del MFA és que podem definir blocs de variables!
- no mirem unicament si guanyem variabilitat, sino tambe si millorem interpretabilitat biologica

3.4 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. FFF: (NO APLICAT ARA PER ARA!)

FFF: PENDENT A AMPLIAR:

- FFF: 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 data sets or computationally intensive tasks.

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.

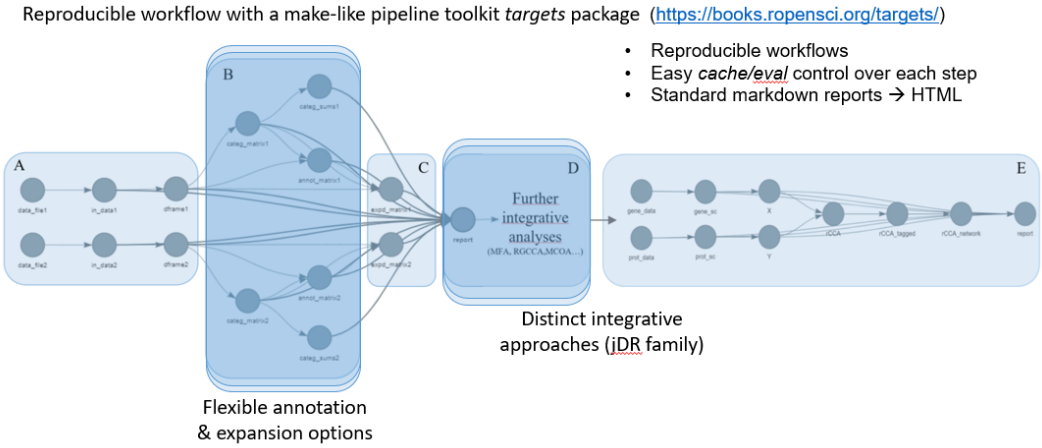

Figure 3.13: Workflow overview

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.

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

Targets workflow diagram (Figure 3.13) showing the steps corresponding with the complete process: The pipeline starts from (A) a couple of 'omics-derived input datasets (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 Mètode proposat per a la integració de info bio

4.2 Packet d'R amb els scripts corresponents

4.3 Workflow (Pipeline) d'anàlisi en amb el paquet 'targets'

4.4 Aplicacions a casos reals 4.3.1 Results from the analysis of human brain tissue samples 4.3.2 Results from the expansion of omics data with biological annotations 4.3.3 Results from the analysis of 150 TCGA-BRCA samples 4.3.4

Results from the application of MFA on TCGA-BRCA data with, and without, expanded data 4.3.5 ...

4.1 Mètode proposat per a la integració de info bio

4.2 Packet d'R amb els scripts corresponents

4.3 Workflow (Pipeline) d'anàlisi en amb el paquet 'targets'

4.4 Aplicacions a casos reals

4.4.1 Results from the analysis of human brain tissue samples

4.4.2 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 3.10 i 3.11 prèvies, de Methods).

4.4.3 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).

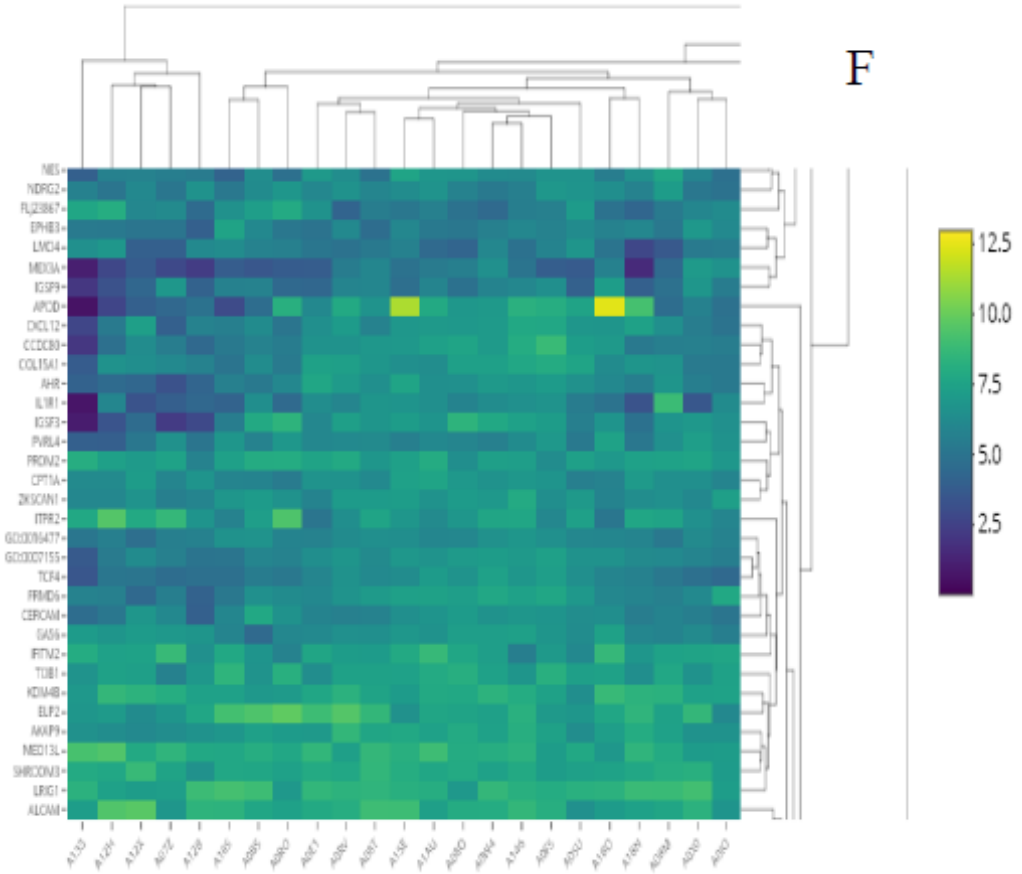

Figure 4.1: Heapmap of an expanded matrix

4.4.4 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.

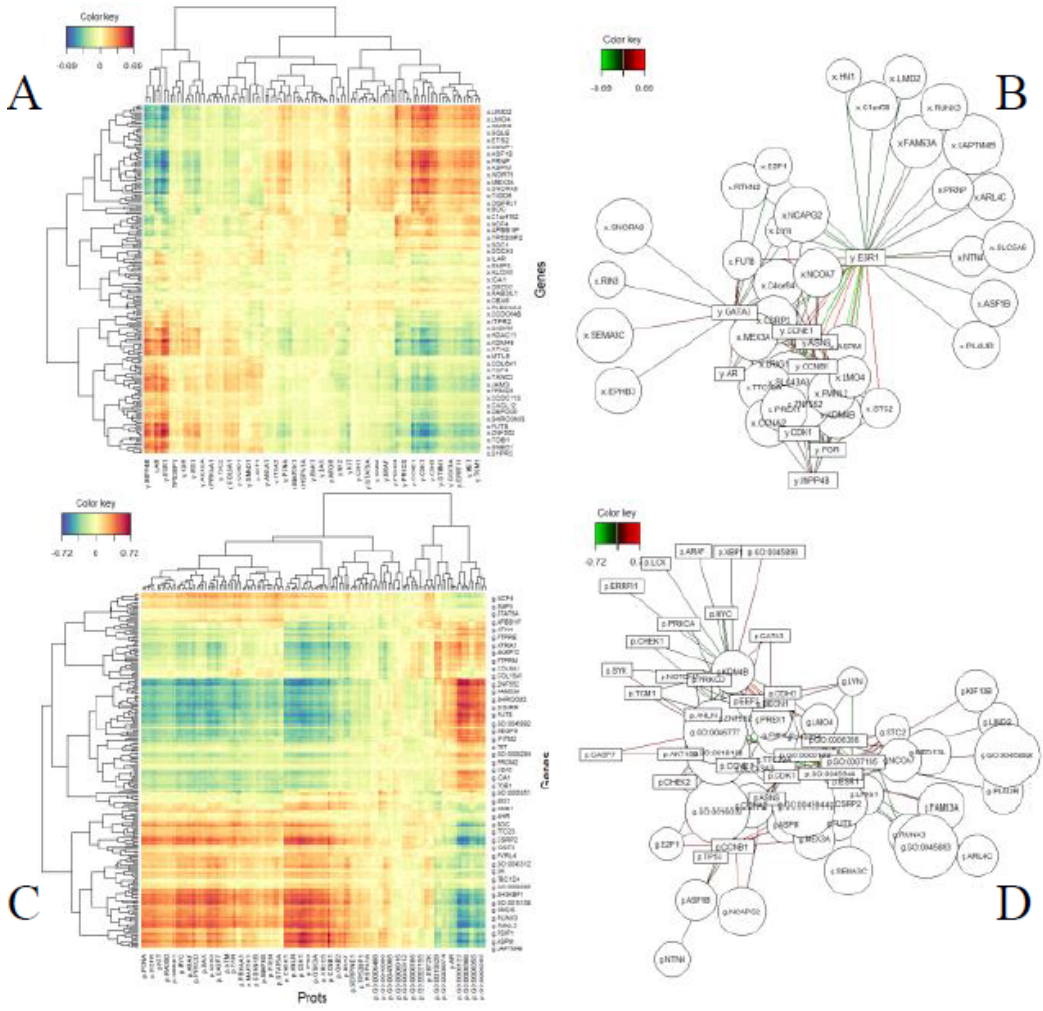

Figure 4.2: BRCA results overview

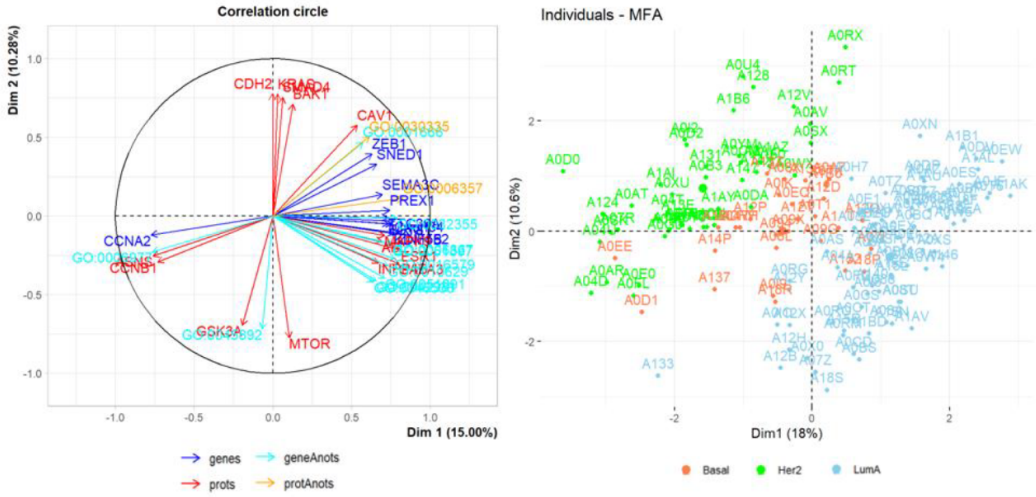

Figure 4.3: BRCA results with MFA

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).

Principals problemes i limitacions de la nostra proposta: * Pendent d'apuntar... * ...

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

- Poder fer servir 3 o més conjunts de dades òmiques
- que l'anotacio basica pugui ser tb a KEGG i no sols a GO
- opció d'afegir les anotacions com a individus suplementaris enlloc de variables
- 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

Athieniti, E., & Spyrou, G. M. (2023). A guide to multi-omics data collection and integration for translational medicine. *Computational and Structural Biotechnology Journal*, *21*, 134–149. https://doi.org/10.1016/j.csbj.2022.11.050

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

Buyukozkan, M., Benedetti, E., & Krumsiek, J. (2023). Rox: A Statistical Model for Regression with Missing Values. *Metabolites*, *13*(1), 127. https://doi.

org/10.3390/metabo13010127

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

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2021). *Shiny: Web application framework for r*. https://CRAN.R-project.org/package=shiny

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

Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., Hanspers, K., Isserlin, R., Kelley, R., Killcoyne, S., Lotia, S., Maere, S., Morris, J., Ono, K., Pavlovic, V., … Bader, G. D. (2007). Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols*, *2*(10), 2366–2382. https://doi.org/10.1038/nprot.2007.324

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

Dai, Z., Bu, Z., & Long, Q. (2022). *Multiple Imputation with Neural Network Gaussian Process for High-dimensional Incomplete Data*. arXiv. https://doi.org/10.48550/arXiv.2211.13297

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

Harris, L., Fondrie, W. E., Oh, S., & Noble, W. S. (2023). Evaluating Proteomics Imputation Methods with Improved Criteria. *Journal of Proteome Research*, *22*(11), 3427–3438. https://doi.org/10.1021/acs.jproteome.3c00205

Hornung, R., Ludwigs, F., Hagenberg, J., & Boulesteix, A.-L. (2024). Prediction approaches for partly missing multi-omics covariate data: A literature review and an empirical comparison study. *WIREs Computational Statistics*, *16*(1), e1626. https://doi.org/10.1002/wics.1626

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

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Ole, A. K., … Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, *12*(2), 115–121. https://doi.org/10.1038/nmeth.3252

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, *28*(1), 27–30. https://doi.org/10.1093/nar/28.1.27

Kidd, J., Raulerson, C. K., Mohlke, K. L., & Lin, D.-Y. (2023). Mediation analysis of multiple mediators with incomplete omics data. *Genetic Epidemiology*, *47*(1), 61–77. https://doi.org/10.1002/gepi.22504

Koboldt, D. C., Fulton, R. S., McLellan, M. D., Schmidt, H., Kalicki-Veizer, J., McMichael, J. F., Fulton, L. L., Dooling, D. J., Ding, L., Mardis, E. R., Wilson, R. K., Ally, A., Balasundaram, M., Butterfield, Y. S. N., Carlsen, R., Carter,

C., Chu, A., Chuah, E., Chun, H.-J. E., ... MD Anderson Cancer Center. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *490*(7418), 61–70. https://doi.org/10.1038/nature11412

Kong, W., Hui, H. W. H., Peng, H., & Goh, W. W. B. (2022). Dealing with missing values in proteomics data. *PROTEOMICS*, *22*(23-24), 2200092. https://doi.org/10.1002/pmic.202200092

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

Landau, W. M. (2021). The targets R package: A dynamic Make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. *Journal of Open Source Software*, *6*(57), 2959. https://doi.org/10.21105/joss.02959

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

Lin, D.-Y., Zeng, D., & Couper, D. (2020). A general framework for integrative analysis of incomplete multiomics data. *Genetic Epidemiology*, *44*(7), 646–664. https://doi.org/10.1002/gepi.22328

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

Lovino, M., Randazzo, V., Ciravegna, G., Barbiero, P., Ficarra, E., & Cirrincione, G. (2021). A survey on data integration for multi-omics sample clustering. *Neurocomputing*, *488*. https://doi.org/10.1016/j.neucom.2021.11.094

Lyngs, U. (2019). Oxforddown: An oxford university thesis template for r markdown. In *GitHub repository*. https://github.com/ulyngs/oxforddown; GitHub. https://doi.org/10.5281/zenodo.3484682

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

Martin, A., Ochagavia, M. E., Rabasa, L. C., Miranda, J., Fernandez-de-Cossio, J., & Bringas, R. (2010). BisoGenet: A new tool for gene network building, visualization and analysis. *BMC Bioinformatics*, *11*(1), 91. https://doi.org/10.1186/1471-2105-11-91

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

Pucher, B. M., Zeleznik, O. A., & Thallinger, G. G. (2019). Comparison and evaluation of integrative methods for the analysis of multilevel omics data: A study based on simulated and experimental cancer data. *Briefings in Bioinformatics*, *20*(2), 671–681. https://doi.org/10.1093/bib/bby027

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

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/journal.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

Song, M., Greenbaum, J., Luttrell, J., Zhou, W., Wu, C., Shen, H., Gong, P., Zhang, C., & Deng, H.-W. (2020). A Review of Integrative Imputation for Multi-Omics Datasets. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.570255

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., Rijn, M. van de, Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lønning, P. E., & Børresen-Dale, A.-L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, *98*(19), 10869–10874. https://doi.org/10.1073/pnas.191367098

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

The Gene Ontology Consortium. (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, *47*(D1), D330–D338. https://doi.org/10.1093/nar/gky1055

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

Uranga, R., Molenberghs, G., & Allende, S. (2022). A multiple regression imputation method with application to sensitivity analysis under intermittent missingness. *Communications in Statistics - Theory and Methods*, *51*(15), 5146–5161. https://doi.org/10.1080/03610926.2020.1834581

Vahabi, N., & Michailidis, G. (2022). Unsupervised Multi-Omics Data Integration Methods: A Comprehensive Review. *Frontiers in Genetics*, *13*. https://www.

frontiersin.org/articles/10.3389/fgene.2022.854752

Voillet, V., Besse, P., Liaubet, L., San Cristobal, M., & González, I. (2016). Handling missing rows in multi-omics data integration: Multiple imputation in multiple factor analysis framework. *BMC Bioinformatics*, *17*(1), 402. https://doi.org/10.1186/s12859-016-1273-5

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

Wekesa, J. S., & Kimwele, M. (2023). A review of multi-omics data integration through deep learning approaches for disease diagnosis, prognosis, and treatment. *Frontiers in Genetics*, *14*. https://www.frontiersin.org/articles/10.3389/fgene.2023.1199087

Wilson, M. D., Ponzini, M. D., Taylor, S. L., & Kim, K. (2022). Imputation of Missing Values for Multi-Biospecimen Metabolomics Studies: Bias and Effects on Statistical Validity. *Metabolites*, *12*(7), 671. https://doi.org/10.3390/metabo12070671

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

Yin, P., & Shi, J. Q. (2019). Simulation-based sensitivity analysis for non-ignorably missing data. *Statistical Methods in Medical Research*, *28*(1), 289–308. https://doi.org/10.1177/0962280217722382

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS : A Journal of Integrative Biology*, *16*(5), 284–287. https://doi.org/10.1089/omi.2011.0118

Zhou, W., Zhao, C., Liu, A., Zhang, X., Cao, X., Ding, Z., Sha, Q., Shen, H., & Deng, H.-W. (2023). *CLCLSA: Cross-omics Linked embedding with Contrastive Learning and Self Attention for multi-omics integration with incomplete multi-omics data*. https://doi.org/10.21203/rs.3.rs-2768563/v1