

# 1

## Introduction

### 1.1 Content of the introductory text (WIP)

The general concept of Data Integration can be defined as the combination of data residing in different sources in order to provide the users with a unified view of these data [1]. However, the practical meaning of the term Integration may vary from, for instance, the computational combination of data, to the combination of studies performed independently, the simultaneous analysis of multiple variables on multiple datasets, or any possible approach for homogeneously querying heterogeneous data sources. Therefore, in many cases, an integrative analysis may be preferable than a simple combination of data from distinct sources. Integrative analysis allows not only for the combination of heterogeneous data, but also for the combined use of these data in order to get the most relevant information and, what is better, to be able to extract some information that could not be unveiled by the separated analysis of each of the original data types.

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 [2,3,4]. 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 [5,6]. 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 [7,8,9]. Meng [10], Cavill [11], Wu [12], Subramanian [30], Krassowski [31], and Cantini [32], 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 [12]. 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) [13]. Fellenberg [14] introduces a way to integrate Gene Ontology information with Correspondence Analysis to facilitate the interpretation of microarray data. De Tayrac et al. [15] 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 [16]. Meng and Culhane [10] 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. [17], 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 to improve its biological interpretability from the point of view of the biomedical researchers.

## **1.2 Background/State of the Art**

Falta desenvolupar punts

### **1.2.1 Omics data analyses**

3 problemes essencials (veure projecte recerca Alex):

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

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

### 1.2.2 Integrative analyses

Allows the combination of distinct omics data.

The blind men and the elephant [https://en.wikipedia.org/wiki/Blind\\_men\\_and\\_an\\_elephant](https://en.wikipedia.org/wiki/Blind_men_and_an_elephant)

Interpretability is a weak point of most multi omics approaches.

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

### 1.2.3 Review of existing approaches for multi-omics data integration

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

## 1.3 Content from template

Welcome to `oxforddown` [[@lyngsOxforddown2019](#)], a thesis template for R Markdown that I created when writing [my own PhD thesis](#) at the University of Oxford. This template allows you to write in R Markdown, while formatting the PDF output with the beautiful and time-tested [OxThesis LaTeX template](#). The sample content is partly adapted from [thesisdown](#) .

Hopefully, writing your thesis in R Markdown will provide a nicer interface to the OxThesis template if you haven't used TeX or LaTeX before. More importantly, *R Markdown* allows you to embed chunks of code within your thesis and generate plots and tables directly from the underlying data, avoiding copy-paste steps. This gets you into the habit of doing reproducible research, which will benefit you long-term as a researcher, and also help anyone that is trying to reproduce or build upon your results down the road.

## 1.4 Why use it?

*R Markdown* creates a simple and straightforward way to interface with the beauty of LaTeX. Packages have been written in **R** to work directly with LaTeX to produce nicely formatting tables and paragraphs. In addition to creating a user friendly interface to LaTeX, *R Markdown* allows you to read in your data, analyze it and to visualize it using **R**, **Python** or other languages, and provide documentation and commentary on the results of your project.

Further, it allows for results of code output to be passed inline to the commentary of your results. You'll see more on this later, focusing on **R**. If you are more into **Python** or something else, you can still use *R Markdown* - see '[Other language engines](#)' in Yihui Xie's *R Markdown: The Definitive Guide*.

Using LaTeX together with *Markdown* is more consistent than the output of a word processor, much less prone to corruption or crashing, and the resulting file

is smaller than a Word file. While you may never have had problems using Word in the past, your thesis is likely going to be about twice as large and complex as anything you've written before, taxing Word's capabilities.

## **Who should use it?**

Anyone who needs to use data analysis, math, tables, a lot of figures, complex cross-references, or who just cares about reproducibility in research can benefit from using *R Markdown*. If you are working in 'softer' fields, the user-friendly nature of the *Markdown* syntax and its ability to keep track of and easily include figures, automatically generate a table of contents, index, references, table of figures, etc. should still make it of great benefit to your thesis project.