UNIVERSITAT DE BARCELONA

# Integrative Analysis of Omics Data with Biological Knowledge in Translational Medicine

**by**

**Ferran Briansó Castilla**

**Annual summary report of the Thesis Research Plan submitted in partial fulfillment for the degree of Doctor of Philosophy in the**

**Departament de Genètica, Microbiologia i Estadística**

**Thesis Director and Tutor:**

**Àlex Sánchez Pla**

**November 2020**

## BACKGROUND

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 et al. [10], Cavill et al. [11] and Wu et. al. [12] are good reviews of the state of the art of using multivariate methods for Integrative 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. Besides this, these methods show strong limitations. Most of them can only rely on the *Gene Ontology* and, what is worst, they require the data to be *multiply rectangular*, that is, they need to have a common dimension, that means that the individuals on which different measurements are taken have to be the same, and this is often an unrealistic assumption. In this thesis the use of both classical *GO Terms* and more flexible *Gene Sets* 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.

## PROJECT OBJECTIVES

The main objectives of this work are the following:

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

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

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

**4. To implement the knowledge acquired with this work into the appropriate bioinformatics tools, e.g. R packages and 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.

## METHODOLOGY

Distinct **working phases**, with the corresponding steps, have to be followed in order to achieve the objectives explained above.

1. On one hand, we have been working on the application of integrative *multi-omics* methods to (I) the analysis of specific, but not publicly available, data sets provided by research units from our affiliation center, Vall d'Hebron Research Institute, and other research institutions that we collaborate with [29,31,36] and (II) to the integrative analysis of larger data sets from public data bases, such as Breast Cancer samples from the TCGA project [18,19].
2. Other steps of this thesis plan are the 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. A final part will be the implementation of the methods developed for this study in the appropriate bioinformatics tools, such as R packages and a web-based application, to facilitate their use in biomedical research projects.

Here follows a list with a brief description of the **main five steps to be performed**, the methods in which they are initially based, and the objectives which they are related to:

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 a publication in 2018[32]. 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 [20]. 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 [21] and mogsa [22], and (II) Regularized Canonical Correlation Analysis with Sparse Partial Least Squares regression (sPLS), provided by mixomics R package [23]. This work had been presented at some meetings [30,33,34,35,37] and in an already published extended abstract's series book [29]. 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 data bases, such as distinct *omics* data from 150 samples from the TCGA-BRCA collection. This work is already finished, and complies with objectives 3 and 2.

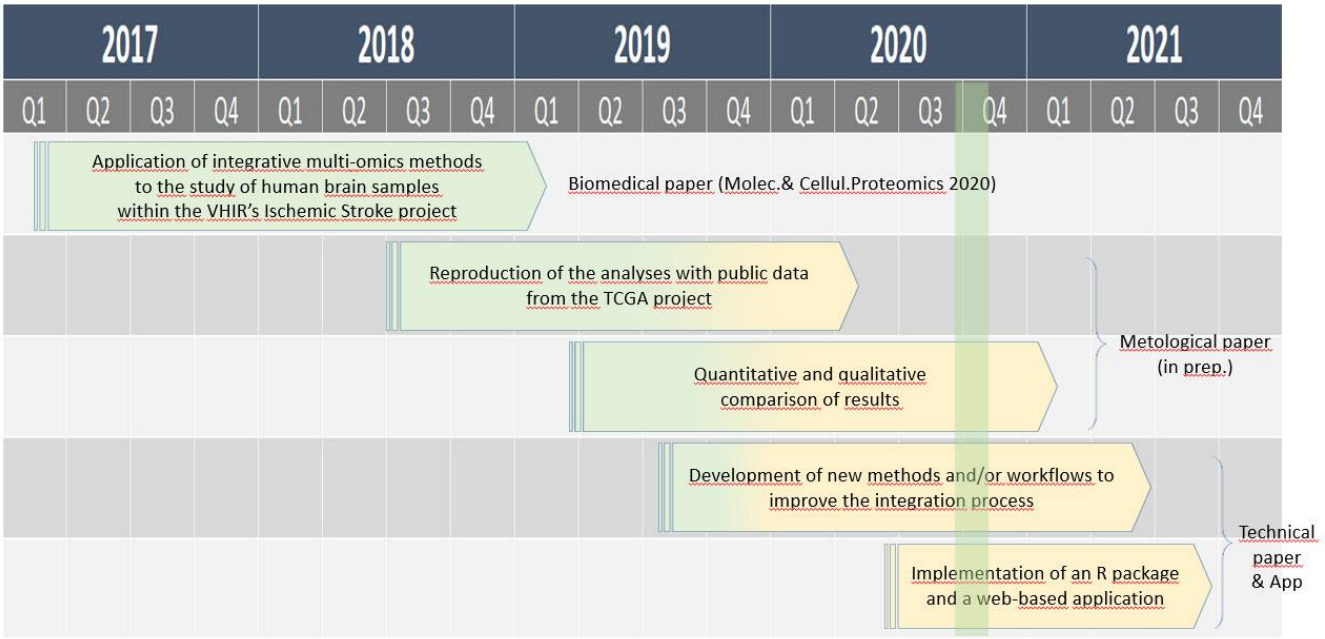
3) Use of all the data sets analyzed up to this point to make a comparison of results between the main implemented methods, and eventually some others, which is the aim of objective 1. This will be done with quantitative and qualitative comparison and visualization methods, such as those explained by Thallinger [24] or Martin [25], going from simple Venn diagrams to more complex, network analysis, software such as some specific R packages [20] or Cytoscape [26].

4) Development of new methods and/or workflows in order to improve and/or combine the benefits from the selected approaches, especially those ones allowing the addition of biological significance to the integration process. This will consist of algorithms and/or mathematical modeling of the approach derived and is intended to comply with the objective number 2 of this working plan.

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

## WORKING PLAN

According to the main steps explained in previous section, here follows a chronogram with the estimated periods on which each part of the work has been, or is scheduled to be, performed:




In addition to that, our goal is to have a published paper for each one of the three main phases explained in the Methodology section: the results from the integrative analysis of brain samples affected by ischemic stroke (biomedical paper: already published in *Molecular & Cellular Proteomics* August 2020); the proposal of new methods to improve the integrative analysis (methodological paper, in prep.); and the implementation of these methods to a software package and a web-based application (at least a technical publication, 2021-2022).

# REFERENCES


1) Lenzerini, M. Data integration: A theoretical perspective. PODS, Madison, Wisconsin, USA, 2002.
2) Cisek, K. Nephrol Dial Transplant. 2015
3) Wang, K. Expert Rev Proteomics. 2014
4) Wang, H. Front Med. 2014
5) Wanichthanarak, K. Biomark Insights. 2015
6) Gomez-Cabrero, D. BMC Syst Biol. 2014
7) Wheelock, ÅM. Mol Biosyst. 2013
8) Lé Cao, K-A. Bioinformatics. 2009
9) Culhane, AC. BMC Bioinformatics. 2003
10) Meng, C. Brief Bioinformatics. 2016
11) Cavill, R. Brief Bioinformatics. 2016
12) Wu, C. High-Throughput. 2019
13) Ashburner, M. Nature Genetics. 2000
14) Busold, CH. Bioinformatics. 2005
15) de Tayrac, M. BMC Genomics 2009
16) Huang, DW. Nucl Acids Res. 2009
17) Tyekucheva, S. Genome Biology 2011
18) TCGA Research Network: http://cancergenome.nih.gov/
19) TCGA-BRCA Project: https://portal.gdc.cancer.gov/projects/TCGA-BRCA
20) R Development Core Team. 2008
21) Culhane, AC. Bioinformatics. 2005
22) Meng, C. bioRxiv. 2016
23) Lé Cao, K-A. 2016 https://CRAN.R-project.org/package=mixOmics
24) Pucher, BM. Brief Bioinform. 2018
25) Martin, A. BMC Bioinformatics. 2010
26) Cline, MS. Nature Protocols. 2007
27) Huber, W. Nature Methods. 2015
28) Shiny by RStudio: http://shiny.rstudio.com/


## RELATED PUBLICATIONS


29) CJ. Rodríguez-Hernández, S. Mateo-Lozano, M. García, C. Casalà, F. Briansó, N. Castrejón, E. Rodríguez, M. Sunol, AM. Carcaboso, C. Lavarino, J. Mora, and C. de Torres. **2016** ***Cinacalcet inhibits neuroblastoma tumor growth and upregulates cancer-testis antigens.*** **Oncotarget,** 7 (13):16112–16129, March 2016. ISSN 1949-2553. doi: 10.18632/oncotarget.7448.

30) F. Briansó, T. García-Berrocoso, J. Montaner, and A. Sánchez-Pla. **2017** ***Integrative Analysis of Transcriptomics and Proteomics Data for the Characterization of Brain Tissue after Ischemic Stroke.*** In: Ainsbury E., Calle M., Cardis E., Einbeck J., Gómez G., Puig P. (eds) Extended Abstracts Fall 2015. Trends in Mathematics, vol 7. Birkhäuser, Cham

31) S. Rodriguez-Fernandez, I. Pujol-Autonell, F. Briansó, D. Perna-Barrull, M. Cano-Sarabia, S. Garcia-Jimeno, A. Villalba, A. Sanchez, E. Aguilera, F. Vázquez, J. Verdaguer, D. Maspoch, and M. Vives-Pi. **2018** ***Phosphatidylserine-Liposomes Promote Tolerogenic Features on Dendritic Cells in Human Type 1 Diabetes by Apoptotic Mimicry.*** **Front. Immunol,** 9:253, February 2018. doi: 10.3389/fimmu.2018.00253

32) A. Simats, L. Ramiro, T. García-Berrocoso, F. Briansó, R. Gonzalo, L. Martín, A. Sabé, N. Gill, A. Penalba, N. Colomé, A. Sánchez-Pla, F. Canals, A. Bustamante, A. Rosell, J. Montaner. **2020** ***A mouse brain-based multi-omics integrative approach reveals potential blood biomarkers for ischemic stroke.*** **Molecular & Cellular Proteomics** August 31, 2020, mcp.RA120.002283; doi: 10.1074/mcp.RA120.002283


## OTHER SCIENTIFIC COMMUNICATIONS

33) XXVIIIth International Biometric Conference IBC2016 (Jul.2016) **Poster presentation.** ***Integrative Analysis of Transcriptomics and Proteomics Data for the Characterization of Brain Tissue After Ischemic Stroke***. F. Briansó, T. García-Berrocoso, J. Montaner, and A. Sánchez.

34) The 15th European Conference on Computational Biology ECCB (Sep.2016) **Poster presentation.** ***Multivariate Methods for the Integrative Analysis of Transcriptomics and Proteomic Data in a Study on Ischemic Stroke.*** F. Briansó, T. García-Berrocoso, J. Montaner, and A. Sánchez-Pla.

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

36) 28th Symposium on Cerebral Blood Flow, Metabolism and Function (Apr.2017) **Short talk.** ***Perivascular macrophages attract neutrophils to the brain after ischemia.*** J. Pedragosa, A. Salas-Pérdomo, M. Gallizioli, R. Cugota, F. Briansó, F. Pérez-Asensio, A. Gieryng, B. Kaminska, F. Miró-Mur, and AM. Planas.

37) 28th Symposium on Cerebral Blood Flow, Metabolism and Function (Apr.2017) **Poster presentation.** ***Integrative analysis of transcriptomics and proteomics data for the molecular characterization of human brain after ischemic stroke***. T. García-Berrocoso, A. Simats, F. Briansó, V. Llombart, A. Hainard, A. Sánchez-Pla, JC Sanchez, and J. Montaner.