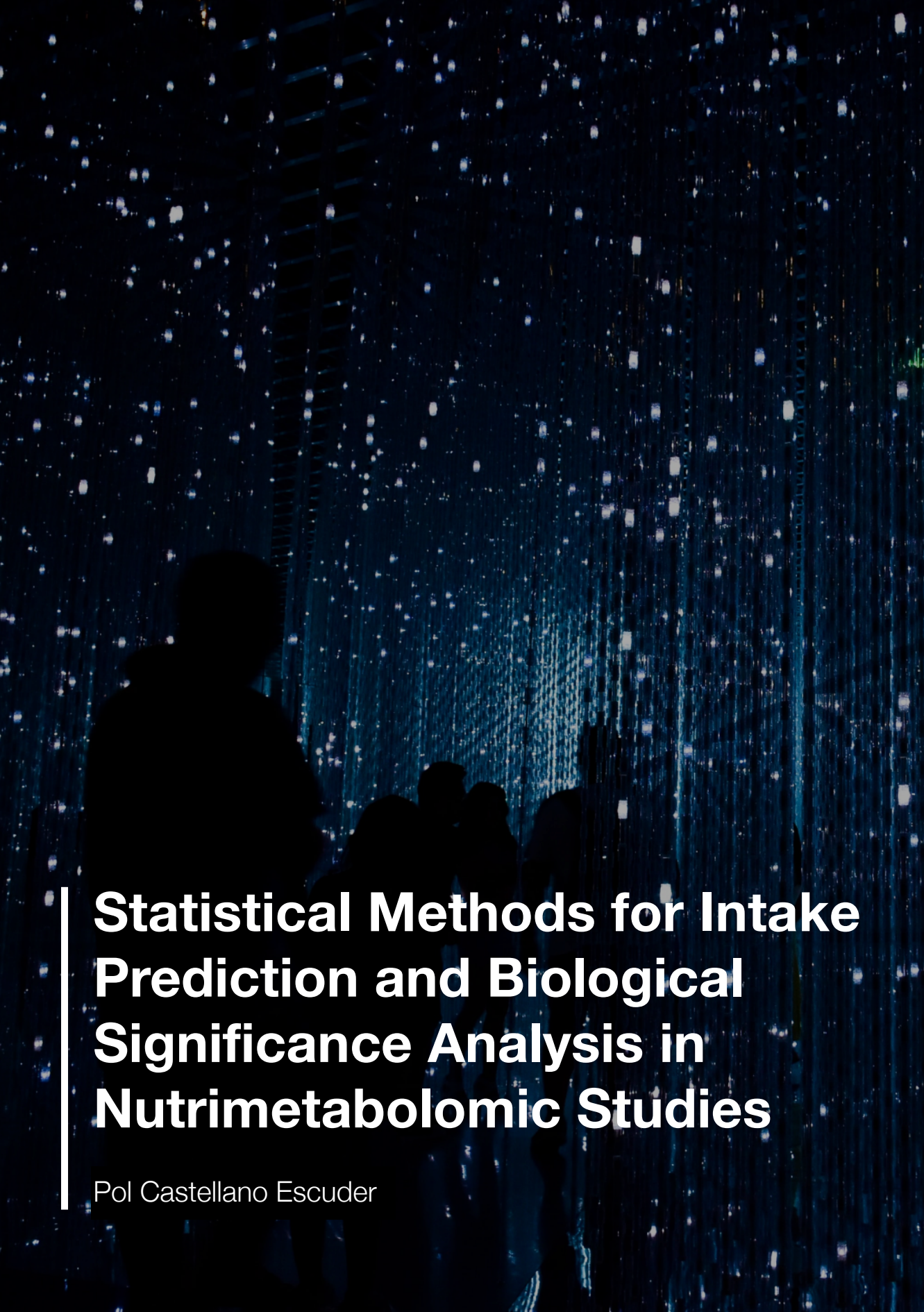
UNIVERSITAT DE BARCELONA

Statistical methods for intake prediction and biological significance analysis in nutrimetabolomic studies

Pol Castellano Escuder



Statistical Methods for Intake Prediction and Biological Significance Analysis in Nutrimetabolomic Studies

Pol Castellano Escuder

UNIVERSITAT DE BARCELONA

PROGRAMA DE DOCTORAT EN BIOMEDICINA

ÀREA DE RECERCA EN BIOINFORMÀTICA

UNIVERSITAT DE BARCELONA

Statistical Methods for Intake Prediction and Biological Significance Analysis in Nutrimetabolomic Studies

Mètodes estadístics per la predicció de la ingesta i l'anàlisi de la significació biològica en estudis de nutrimetabolòmica

Memòria presentada per **Pol Castellano Escuder**

per optar al grau de doctor per la Universitat de Barcelona

Departament de Genètica, Microbiologia i Estadística

Departament de Nutrició, Ciències de l'Alimentació i Gastronomia

Directors de la tesi

Dr. Alex Sánchez Pla (*Tutor*)

Dra. Cristina Andrés Lacueva

Juny 2021

To my parents:

Núria Escuder Vidal
Tomàs Castellano de Solà

Acknowledgements

"Success is stumbling from failure to failure with no loss of enthusiasm."
Winston S. Churchill

"The only way to write good code is to write tons of shitty code first. Feeling shame about bad code stops you from getting to good code."
Hadley Wickham

I am absolutely convinced that no matter how hard the effort, how many hours of dedication, even no matter how high the individual skills; all the great achievements of life always depend strongly on the conditions, context and above all, on the people around you.

This thesis is an example where this idea becomes more than obvious, and I can only restate that. It is obvious that there are many hidden hours of work and sacrifice behind this thesis, but even so, I truly believe that none of this would have been possible without all the people who have sided with me, not only scientifically talking, but in every possible way. For this reason, it is truly an honor for me to address these words of gratitude to everyone who has contributed in some way to this personal achievement.

First and foremost, I would like to express my gratitude to my thesis supervisors, **Prof. Alex Sánchez Pla** and **Prof. Cristina Andrés Lacueva**. I am enormously grateful to them for believing in me and giving me the opportunity to do a doctoral thesis, making my dream come true. Thanks for advising me, guiding me, correcting my scientific work, and for making this thesis a huge learning process that has far exceeded my expectations.

I would like to extend these thanks to the professors, colleagues and friends of the Statistics and Bioinformatics group, for everything they have taught me, which is most of the statistical knowledge I have today. Thanks **Prof. Francesc Carmona**, **Prof. Esteban Vegas**, **Prof. Ferran Reverter**, **Prof. Antonio Miñarro**, and **Prof. Marta Cubedo**.

I would also like to express my gratitude to my colleagues and friends of the Biomarkers and Nutritional & Food Metabolomics group, where I have spent most of the time and with whom we have shared countless moments, from lab meetings to trips, dinners, and beers. Without them it would have been impossible to do this work, and I am grateful to say that I have finished this thesis with much more than lab colleagues. I would especially like to thank **Magalí Palau**, for guiding me in the first steps of the thesis; **Maria Cristina Cadena**, for her smile and her joy that always helped to create an unbeatable atmosphere in the lab; and **Nicole Hidalgo**, who after more than three years has become a sister and an indispensable support to me.

It is also my deep personal desire to thank all the members of GRBIO, from whom I have learned a lot and with whom I have shared fantastic moments, from research seminars to the most picturesque moments. I would like to specifically express my thanks to **Marta Bofill** and **Guillermo Villacampa**, for always being there, supporting each other, discussing a multitude of statistical topics and sharing incredible moments, becoming great friends, within and outside the work environment.

Thanks to all the people I met in Aberystwyth. Thanks to **Prof. John Draper** for welcoming me in his lab, and also thanks to **Tom Wilson** and **Mandy Lloyd** for their guidance during my stay abroad.

I would also like to dedicate a special thanks to all the people who at the time made me feel passionate about science, advised me, and guided me in the decisions that made me, in part, write these lines today. My most sincere thanks, **Josep Jiménez**, **Silvia Ribó** and **Judith Cebrià**.

I would like to extend these thanks to all my friends, both inside

and outside science, who have given me unquestionable support in carrying out this personal achievement. Thanks to **Marçal Yll** and **Núria Catasús**, with whom I have shared science since we started our biology and biochemistry studies at 18 years old. I also thank my friends **Enric Martí** and **Adrià Hernández**, who have always been there to support and listen to me, especially in the final stretch of this work. I would also like to express my gratitude to **Damià**, for all his help and for making my path easier, especially in the last year of the thesis, which coincided with the COVID-19 pandemic.

Finally, I would like to give my deepest gratitude to my family, my great referents, those who have always been by my side. To my mother, an example of perseverance, perfectionism and self-improvement. To my father, from whom I do not stop learning and who has taught me how to face all the challenges of life. And my sister, **Carlota**, who has always given me unconditional support in all my personal challenges and who I am sure, will achieve everything she sets out to do in life. And last but certainly not least my thanks and love to **Montse**, my partner, who patiently supported plenty of holidays, weekends, and even vacations devoted to this work.

Thanks to all those I have not mentioned, who have contributed in some manner to this achievement. Thank you very much.

Abstract

This thesis is the product of three and a half years working on the complex world of metabolomics and nutrition. All the work presented here is focussed on the problems arising from associating and integrating metabolomics data with nutritional or dietary data. This issue has been approached using both observational and interventional studies and from a mainly bioinformatic point of view, proposing different methods and tools to reduce the complexity of nutrimetabolomics data analysis.

Thus, this work consists of four chapters divided into three parts, in addition to a summary of the content of the entire thesis in Catalan, the references, and the appendices. The first part consists of a global introduction, where the fundamental concepts needed for the correct understanding of the thesis are reviewed, as well as basic concepts about metabolomics and nutrition, the state of the art of the nutrimetabolomics field, and the fundamentals of biological significance analyses, among others. Then, this first part ends with a brief definition of the objectives of this work. In the second part, the results of this thesis are carefully presented and discussed. The results are presented in a compact format, with each section being a summary of a scientific publication. These results include the develompent of an ontology that defines the relationships between dietary metabolites and foods, the development of an open source tool for metabolomics data analysis, the development of an open source tool for nutrimetabolomics enrichment analysis, other open source tools developed in the context of this work, and a section with different publications where the methods and tools developed have been applied. Then, all these individual results are discussed together, providing a global and unified context where all the developments of this thesis are related. Lastly, the third part of this thesis presents

the conclusions, contextualizing all the obtained results within the main objective of the thesis: contribute to the improvement of the integration and interpretation of nutrimetabolomics data. Additionally, in the appendices, the published results and some extra information used in carrying out this research are presented.

Finally, although this thesis is made up of contents from the fields of metabolomics, nutrition, bioinformatics and biostatistics, it has been written for a wide scientific audience, trying to be as comprehensible as possible for any profile of researchers, avoiding unnecessary complexities and always following the transversal objective of the thesis.

I hope you find it useful but, above all, that you enjoy reading it.

Table of Contents

Glossary **1**

I Introduction and Objectives **5**

Chapter 1: Introduction **7**
1.1 Metabolomics 8
1.1.1 Metabolome 8
1.1.1.1 Human metabolome 9
1.1.2 Metabolome profiling techniques 10
1.1.2.1 Mass spectrometry 11
1.1.2.2 NMR spectroscopy 14
1.1.3 Metabolome profiling approaches 14
1.1.3.1 Untargeted metabolomics 14
1.1.3.2 Targeted metabolomics 15
1.2 Nutrimetabolomics 16
1.2.1 Nutritional studies 17
1.2.1.1 Interventional studies 17
1.2.1.2 Observational studies 18
1.2.2 Dietary assessment techniques 18
1.2.2.1 Dietary recalls 19
1.2.2.2 Food frequency questionnaires 20
1.2.2.3 The need for a complementary approach 21
1.3 Biomarkers 21
1.3.1 Dietary and health biomarkers 22
1.3.1.1 Types of dietary biomarkers 23
1.4 Data analysis in nutrimetabolomics 23
1.4.1 Statistical modeling 25
1.4.1.1 Linear models 25

1.4.1.2 Generalized linear models 26
1.4.1.3 Generalized additive models 26
1.4.2 Data mining 27
1.4.2.1 Principal Components Analysis 27
1.4.2.2 Partial Least Squares 29
1.4.3 Statistical learning 31
1.4.3.1 The Lasso 31
1.4.3.2 Random forests 32
1.5 Ontologies 35
1.5.1 The gold standard: The Gene Ontology 37
1.5.2 Ontologies in metabolomics 38
1.5.2.1 ChEBI: Chemical Entities of Biological Interest 39
1.5.3 Ontologies in nutrition 41
1.5.3.1 FoodOn: Food Ontology 41
1.5.3.2 ONS: Ontology for Nutritional Studies 43
1.5.4 Ontologies in nutrimetabolomics 44
1.6 Biological significance analysis 45
1.6.1 Biological significance analysis methods 47
1.6.1.1 Over Representation Analysis 48
1.6.1.2 Gene Set Enrichment Analysis 49
1.6.2 Biological significance analysis in nutrimetabolomics 51

Chapter 2: Objectives 53
2.1 Main objective 53
2.2 Specific objectives 53

II Results and Discussion 55

Thesis directors report 57

Chapter 3: Results 61
3.1 Methodological and software developments 62
3.1.1 Paper 1: Food-Biomarker Ontology 62
3.1.1.1 Background 62
3.1.1.2 Aim 62
3.1.1.3 Results 62

3.1.1.4 Conclusion 64
3.1.2 Paper 2: POMAShiny 65
3.1.2.1 Background 65
3.1.2.2 Aim 66
3.1.2.3 Results 66
3.1.2.4 Conclusion 67
3.1.3 Paper 3: The fobitools framework 69
3.1.3.1 Background 69
3.1.3.2 Aim 70
3.1.3.3 Results 70
3.1.3.4 Conclusion 73
3.2 Application of developed tools 74
3.2.1 Paper 4: Assessing adherence to healthy dietary habits through the urinary food metabolome . . 74
3.2.1.1 Background 74
3.2.1.2 Aim 74
3.2.1.3 Study design 75
3.2.1.4 Results 75
3.2.1.5 Conclusion 76
3.2.2 Paper 5: The food-related serum metabolome associates with later cognitive decline in older subjects 76
3.2.2.1 Background 77
3.2.2.2 Aim 77
3.2.2.3 Study design 77
3.2.2.4 Results 79
3.2.2.5 Conclusion 79
3.3 Software 79
3.3.1 R/Bioconductor packages 79
3.3.1.1 POMA 80
3.3.1.2 fobitools 80
3.3.2 Graphical User Interfaces 81
3.3.2.1 POMAShiny 81
3.3.2.2 fobitoolsGUI 82
3.3.2.3 POMAcounts 83

Chapter 4: Discussion 85

III Conclusions **95**

IV Resum en català **99**

Agraïments **101**

Chapter 5: Introducció **105**
5.1 Metabolòmica 106
5.1.1 Metaboloma 106
5.1.1.1 Metaboloma humà 106
5.1.2 Tècniques d'obtenció de perfils de metabolòmics 107
5.1.2.1 Espectrometria de masses 107
5.1.2.2 Ressonància magnètica nuclear 108
5.1.3 Estratègies d'obtenció de perfils de metabolòmics 108
5.2 Nutrimetabolòmica 109
5.2.1 Estudis nutricionals 110
5.2.2 Mètodes per a l'assessorament dietètic 111
5.3 Biomarcadors 112
5.4 Anàlisi de dades nutrimetabolòmiques 113
5.4.1 Modelització estadística 113
5.4.2 Mineria de dades 114
5.4.3 Aprenentatge estadístic 115
5.5 Ontologies 115
5.6 Anàlisi de la significació biològica 117
5.6.1 Mètodes d'anàlisi de la significació biològica . . 118
5.6.2 Anàlisi de la significació biològica en nutrimetabolòmica 119

Chapter 6: Objectius **121**
6.1 Objectiu principal 121
6.2 Objectius específics 121

Chapter 7: Resultats **123**
7.1 Desenvolupaments metodològics i de programari 124
7.1.1 Article 1: L'ontologia de biomarcadors i aliments 124
7.1.2 Article 2: POMAShiny 125
7.1.3 Article 3: L'entorn de treball fobitools 126
7.2 Aplicació de les eines desenvolupades 126

7.2.1 Article 4: Avaluació de l'adherència a hàbits dietètics saludables mitjançant el metaboloma alimentari en orina . . . 126
7.2.2 Article 5: El metaboloma serològic relacionat amb els aliments s'associa amb un detreiorament cognitiu tardà en individus d'edat avançada . . 128
7.3 Programari . . . 129
7.3.1 Paquets de Bioconductor . . . 129
7.3.1.1 POMA . . . 129
7.3.1.2 fobitools . . . 130
7.3.2 Interfícies gràfiques . . . 130
7.3.2.1 POMAShiny . . . 131
7.3.2.2 fobitoolsGUI . . . 131
7.3.2.3 POMAcounts . . . 132

Chapter 8: Discussió . . . 133

Chapter 9: Conclusions . . . 143

V References 147

VI Appendices 159

Appendix A . . . 161
A.1 Thesis publications . . . 161
A.1.1 Paper 1: Food-Biomarker Ontology . . . 161
A.1.2 Paper 2: POMAShiny . . . 170
A.2 Thesis software . . . 187
A.2.1 POMA use case . . . 187
A.2.2 fobitools use case . . . 209

Appendix B . . . 225
B.1 Other publications . . . 225
B.1.1 Paper 6: A polyphenol-rich diet causes increase in the gut microbiota metabolite indole 3-propionic acid in older adults with preserved kidney function 225
B.1.2 Paper 7: Apolipoprotein E and sex modulate fatty acid metabolism in early cognitive decline 227

B.1.3 Paper 8: A mixture of four dietary fibres ameliorates adiposity, and improves metabolic profile and intestinal health in cafeteria-fed obese rats: an integrative multi-omics approach 228
B.2 Other software 229
B.2.1 Lheuristic 229
B.2.2 Covid19Explorer 230

List of Tables

1.1 The advantages and limitations of NMR spectroscopy and MS spectrometry as an analytical tool for metabolomics research adapted from Emwas, 2015. . . 13
1.2 A 2×2 table for assessing over-representation analysis (from Goeman & Bühlmann, 2007). 49

3.1 Statistical methods provided in POMAShiny. *Methods that allow the use of covariates. 68
3.2 Example of dietary free-text annotation with fobitools package. 73
3.3 Clinical and demographic characteristics of the discovery and validation case-control samples of the D-CogPlast study. 78

List of Figures

1.1 *"The Omics Cascade"* (Narad & Kirthanashri, 2018). . 8
1.2 The human metabolomes (Scalbert et al., 2014). 9
1.3 The HPLC-MS equipment used throughout this work. . 12
1.4 Untargeted versus targeted metabolomics studies (Schrimpe-Rutledge, Codreanu, Sherrod, & McLean, 2016). 15
1.5 A 24-Hour dietary recall. 19
1.6 A food frequency questionnaire, illustrating bread, savoury biscuits and breakfast cereals (Mulligan et al., 2014). 20
1.7 Key steps of a metabolomics study. 24
1.8 The first two principal components for the ST000336 data set provided in the POMA package. The red circles indicate control subjects and the blue triangles indicate DMD subjects. 28
1.9 First two PLS-DA components for the ST000336 data set provided in the POMA package. The red circles indicate control subjects and the blue triangles indicate DMD subjects. 29
1.10 *Bagging* methodology scheme. 32
1.11 The scheme of random forest algorithm. The final prediction is obtained by taking a majority vote of the predictions from all the trees in the forest (Hongdong et al., 2019). 33
1.12 Mean Decrease in Gini of the top 10 selected features for the ST000336 data set provided in the POMA package. 34
1.13 Growth of GO annotated scientific publications over time (from http://geneontology.org). 37

1.14 The structure of GO for the *hexose biosynthetic process* term (from http://geneontology.org). 38
1.15 Growth of ChEBI annotated compounds over its releases (from https://www.ebi.ac.uk/chebi/). 40
1.16 The structure of ChEBI for the *phloretin* term. 40
1.17 The structure of FoodOn for the *apple* term (Dooley et al., 2018). 42
1.18 The ONS hierarchical structure. The terms in green boxes are ONS-specific terms, while terms in other color boxes are imported from existing ontologies (Vitali et al., 2018). 43
1.19 Gene lists derived from diverse *omics* data undergo pathway enrichment analysis, using ORA and GSEA methods, to identify pathways that are enriched in the experiment (adapted from Reimand et al., 2019). 48
1.20 A GSEA overview illustrating the method. A) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the "gene tags" (location of genes from a set S within the sorted list). B) Plot of the running sum for S in the data set, including the location of the maximum enrichment score (ES) and the leading-edge subset (Subramanian et al., 2005). 50

3.1 The structure of FOBI for the *apple* term (Castellano-Escuder et al., 2020). 64
3.2 POMAShiny's workflow (Castellano-Escuder et al., 2021). 66
3.3 FOBI sub-network corresponding to the annotated terms in the Table 3.2. This network has been generated with the fobitools package. 72
3.4 Cook2Health project study design. 75
3.5 Screenshot of the POMAShiny *Home* page. 82
3.6 Screenshot of the fobitoolsGUI *Enrichment Analysis* page. 83
3.7 Screenshot of the POMAcounts *Home* page. 84

B.1 MaPLE project study design (Guglielmetti et al., 2020). 227
B.2 Screenshot of the Lheuristic *Home* page. 230
B.3 Screenshot of the Covid19Explorer *Home* page. 231

Glossary

AHEI-2010 Alternative Healthy Eating Index 2010

BSA Biological Significance Analysis

CD Cognitive Decline

ChEBI Chemical Entities of Biological Interest

D-CogPlast Dietary modulators of Cognitive ageing and brain Plasticity

Da Dalton (unified atomic mass unit)

DE Differentially Expressed

DMD Duchenne Muscular Dystrophy

DR Dietary Recall

EDA Exploratory Data Analysis

EIT European Institute of Innovation and Technology

FAIR Findable, Accessible, Interoperable and Reusable

FC Fold Change

FDR False Discovery Rate

FFQ Food Frequency Questionnaire

FOBI Food-Biomarker Ontology

FooDB Food Database

FoodBAll Food Biomarkers Alliance

FoodOn Food Ontology

GAM Generalized Additive Model

GC-MS Gas Chromatography-Mass Spectrometry

GLM Generalized Linear Model

GO Gene Ontology

GSEA Gene Set Enrichment Analysis

GUI Graphical User Interface

HMDB Human Metabolome Database

HPLC High-Performance Liquid Chromatography

ID Identifier

KEGG Kyoto Encyclopedia of Genes and Genomes

LASSO Least Absolute Shrinkage and Selector Operator

LC-MS Liquid Chromatography-Mass Spectrometry

Limma Linear Models for Microarray and RNA-Seq Data

LM Linear Model

MaPLE Microbiome mAnipulation through Polyphenols for managing Leakiness in the Elderly

ML Machine Learning

MS Mass Spectrometry

MSEA Metabolite Set Enrichment Analysis

MSI Metabolomics Standards Initiative

NMR Nuclear Magnetic Resonance

OBI Ontology for Biomedical Investigations

OBO Open Biological and Biomedical Ontology

ONS Ontology for Nutritional Studies

OPLS-DA Orthogonal Partial Least Squares Discriminant Analysis

ORA Over Representation Analysis

OS Operating System

OWL Web Ontology Language

PCA Principal Components Analysis

PCR Principal Components Regression

PLS Partial Least Squares

PLS-DA Partial Least Squares Discriminant Analysis

RDF Resource Description Framework

RF Random Forests

sPLS-DA Sparse Partial Least Squares Discriminant Analysis

VIP Variable Importance in the Projection

XML eXtensible Markup Language

Part I

Introduction and Objectives

Chapter 1

Introduction

Understanding the link between nutrition and health is one of the major goals of modern nutrition. Thanks to the great advances in the field of metabolomics in recent years, this high-throughput technique has become increasingly an indispensable ally for nutrition sciences, being nutritional metabolomics (or nutrimetabolomics) a key tool for exploring the relationships between diet and health and for predicting food intake through metabolomic profiles, among others.

However, many of the relationships between metabolites and foods are not yet fully clear and are subject to discussion, requiring further in-depth studies in this area. This thesis focuses on the exhaustive study of these relationships between metabolites and diet in order to better understand their complexity and contribute to the improvement and simplification of nutrimetabolomics data analysis as well as to the improvement of the biological interpretation of its results.

To achieve this goal, this work proposes different bioinformatic tools designed for the problems arising from the analysis of these data. Different tools and resources such as an ontology that defines the relationships between metabolites and foods, a statistical analysis tool for metabolomics data analysis, and tools for biological significance analysis[1] in nutrimetabolomic studies are presented below.

[1]This concept is carefully detailed in "Biological significance analysis" section.

1.1 Metabolomics

In life sciences, the suffix "*-omics*" refers to the collective characterization and quantification of pools of biological molecules that translate into the structure, function, and dynamics of an organism. Thus, the different high-throughput study disciplines are classified as different "*omics*" in the world of biology, being the genomics, transcriptomics, proteomics, and metabolomics the main and most studied *omics*, respectively.

Sequentially and ordered from "*top-to-down*" (Figure 1.1), genomics studies the structure and function of the genome and its different genes, transcriptomics studies the different transcripts produced by the genome (e.g., mRNA, rRNA, tRNA), proteomics studies proteins formed by the translation of transcripts, and metabolomics studies the small molecule (< 900 Da) metabolic products (e.g., lipids and vitamins) of biological systems (often derived from protein-mediated biochemical reactions). In recent years, the growing field of metabolomics is focusing on the analysis of many hundreds of metabolites in complex specimens that include biofluids, tissues, and cells.

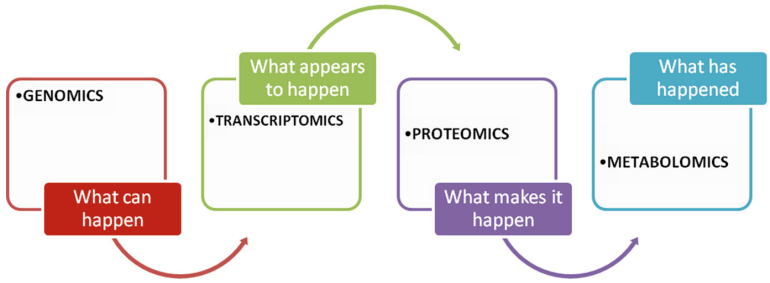


Figure 1.1: *"The Omics Cascade"* (Narad & Kirthanashri, 2018).

1.1.1 Metabolome

The metabolome is the collection of all low molecular weight molecules (metabolites) present in a biological system. It is a result of the biochemical reactions being catalyzed by the proteins of the proteome (Figure 1.1) and determines the final phenotype of the organism. The number of the different compounds in the metabolome varies depending

on the organism but it is constantly changing due to all the chemical reactions occurring in the organism (Færgestad et al., 2009).

1.1.1.1 Human metabolome

Specifically, this thesis focuses on the human metabolome, composed of all metabolites in the human body. However, all these metabolites which constitute the human metabolome can come from four well-defined sources that are described below (Figure 1.2).

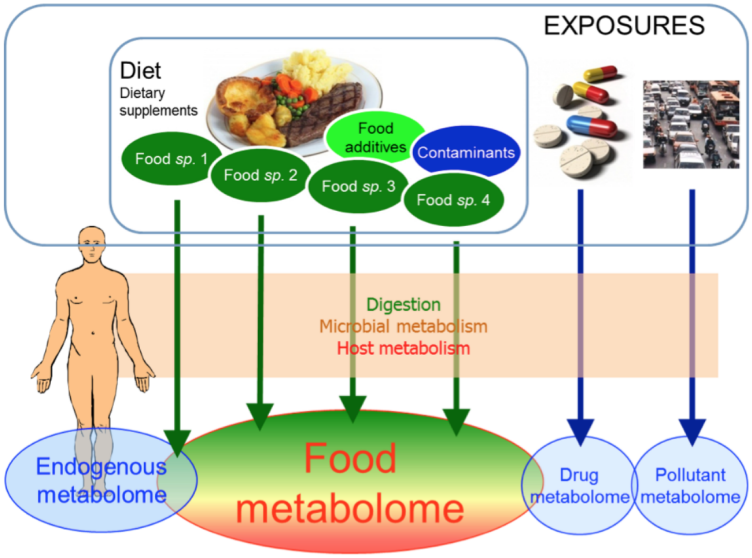


Figure 1.2: The human metabolomes (Scalbert et al., 2014).

- **Food metabolome**

During the last 15 years, different definitions have been proposed for this type of metabolome (Cevallos-Cevallos, Reyes-De-Corcuera, Etxeberria, Danyluk, & Rodrick, 2009; Fardet et al., 2008; Wishart, 2008), however, the most accepted one today is the following:

"*The food metabolome is defined as the part of the human metabolome directly derived from the digestion and biotransformation*

of foods and their constituents" (Scalbert et al., 2014).

According to this definition, the food metabolome does not consider those compounds present in the nature of foods, but of compounds derived from the absorption, digestion and biochemical transformations that food undergoes after ingestion. This type of human metabolome is composed of more than 25000 compounds and is extremely complex and variable depending on diet. This thesis focuses mainly on this type of metabolome.

- **Endogenous metabolome**

The endogenous metabolome is the collection of compounds present in the human metabolome that are naturally produced by the body (e.g., amino acids, organic acids, fatty acids, pigments). Like the food metabolome, this metabolome can also be influenced by diet, being altered those endogenous metabolites and metabolic pathways affected by diet.

- **Drug metabolome**

The drug metabolome is the collection of xenobiotic compounds from derivatives of drugs and/or nutraceuticals present in the human metabolome.

- **Pollutant metabolome**

The pollutant metabolome is composed of metabolic derivatives of environmental pollution.

1.1.2 Metabolome profiling techniques

The most common used techniques used for metabolome profiling are mass spectrometry and NMR spectroscopy.

1.1.2.1 Mass spectrometry

Mass spectrometry (MS) is an analytical technique that measures the mass-to-charge ratio of ions. The results are typically presented as a mass spectrum, a plot of intensity as a function of the mass-to-charge ratio. MS is used in many different fields and is applied to pure samples as well as complex mixtures.

These spectra are used to determine the elemental or isotopic signature of a sample, the masses of particles and molecules, and to elucidate the chemical identity or structure of molecules and other chemical compounds.

In a typical MS procedure, a sample, which may be solid, liquid, or gaseous, is ionized. This may cause some of the sample's molecules to break into charged fragments or simply become charged without fragmenting. These ions are then separated according to their mass-to-charge ratio. The ions are detected by a mechanism capable of detecting charged particles, such as an electron multiplier. Results are displayed as spectra of the signal intensity of detected ions as a function of the mass-to-charge ratio. The atoms or molecules in the sample can be identified by correlating known masses (e.g., known metabolite) to the identified masses or through a characteristic fragmentation pattern.

An important enhancement to MS is using it in tandem with different chromatographic techniques:

- **Gas chromatography–mass spectrometry**

Gas chromatography–mass spectrometry (GC-MS) is used for the analysis of volatile molecules. The sample first goes through the gas chromatograph where high-resolution separation of volatile organic compounds in a mixture is accomplished in the gas phase. This stream of separated compounds is fed online into the ion source, a metallic filament to which voltage is applied. This filament emits electrons which ionize the compounds. The ions can then further fragment, yielding predictable patterns. Intact ions and fragments pass into the mass spectrometer's analyzer and are detected (Emwas, 2015).

- **Liquid chromatography-mass spectrometry**

Liquid chromatography-mass spectrometry (LC-MS) comprises two powerful analytical tools, high-performance liquid chromatography (HPLC) and MS. When combined, LC-MS represents a very powerful analytical tool for the separation, identification, and quantification of molecules in a mixed sample. The HPLC technique separates molecules first based on different physical and chemical properties such as molecular size, charge, polarity, and affinity toward other molecules. Once the analytes are separated, they pass through the mass spectrometer analyzer where they are detected based on the mass-to-charge ratio, and the intensity of each resultant line corresponds to relative concentration of each molecule (Emwas, 2015). It differs from GC-MS in that the mobile phase is liquid, usually a mixture of water and organic solvents, instead of gas.

All data used in this thesis have been obtained via HPLC-MS.

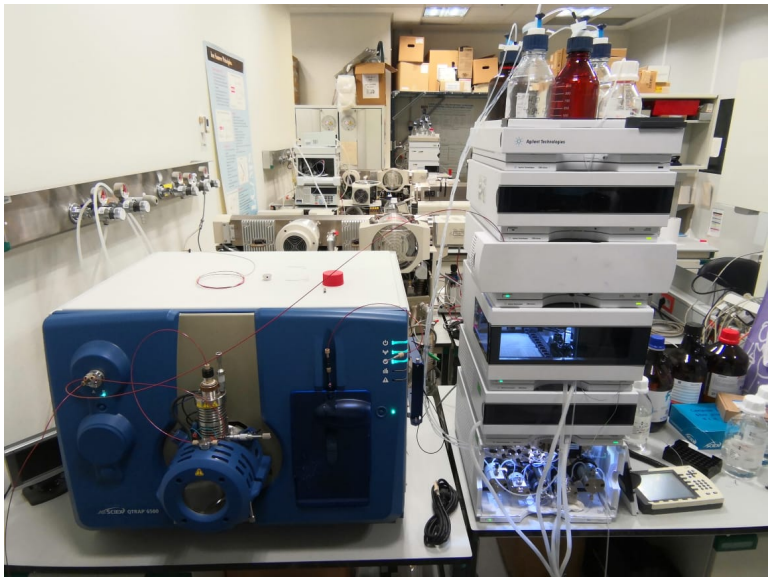


Figure 1.3: The HPLC-MS equipment used throughout this work.

Table 1.1: The advantages and limitations of NMR spectroscopy and MS spectrometry as an analytical tool for metabolomics research adapted from Emwas, 2015.

	NMR	Mass spectrometry
Sensitivity	Low	High
Selectivity	Nonselective analysis	Both selective and nonselective (targeted and nontargeted) analyses
Sample measurement	All metabolites can be detected in one measurement	Need different chromatography techniques for different classes of metabolites
Sample recovery	Nondestructive, sample can be recovered, restored, reused	Destructive but need a small amount of sample
Reproducibility	Very high	Moderate
Sample preparation	Minimal sample preparation	More demanding; needs different columns and optimization of ionization conditions
Number of detectable metabolites	40–200	> 500
Target analysis	Not relevant for targeted analysis	Superior for targeted analysis

1.1.2.2 NMR spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy is a spectroscopic technique to observe local magnetic fields around atomic nuclei. The sample is placed in a magnetic field and the NMR signal is produced by excitation of the nuclei sample with radio waves into nuclear magnetic resonance, which is detected with sensitive radio receivers. The intramolecular magnetic field around an atom in a molecule changes the resonance frequency, thus giving access to details of the electronic structure of a molecule and its individual functional groups.

This technique has been used to identify proteins, metabolites, and other complex molecules. Besides identification, NMR spectroscopy provides detailed information about the structure, dynamics, reaction state, and chemical environment of molecules. NMR spectra are unique, well-resolved, analytically tractable and often highly predictable for small molecules.

However, both MS and NMR techniques present different advantages and limitations. Some of these advantages and limitations are shown in Table 1.1.

1.1.3 Metabolome profiling approaches

Metabolome profiling approaches can be divided into two different groups (Figure 1.4); untargeted metabolomics, an intended comprehensive analysis of all the measurable analytes in a sample (including chemical unknowns), and targeted metabolomics, the measurement of defined groups of chemically characterized and biochemically annotated metabolites (Roberts, Souza, Gerszten, & Clish, 2012).

1.1.3.1 Untargeted metabolomics

Untargeted (or discovery-based) metabolomics focuses on global detection and relative quantification of small molecules in a sample, including chemical unknowns.

This type of approach must be coupled to advanced chemometric techniques to reduce the extensive data sets generated into a smaller set of manageable signals (Schrimpe-Rutledge, Codreanu, Sherrod, & McLean, 2016). Untargeted analysis offers the opportunity for novel target discovery, as coverage of the metabolome is only restricted by the methodologies of sample preparation and the inherent sensitivity and specificity of the analytical technique employed. However, the principal challenges of untargeted analysis lie in the protocols and time required to process the extensive amounts of raw data generated, the difficulties in identifying and characterizing unknown small molecules, the reliance on the intrinsic analytical coverage of the platform employed, and the bias towards detection of high-abundance molecules (Schrimpe-Rutledge et al., 2016).

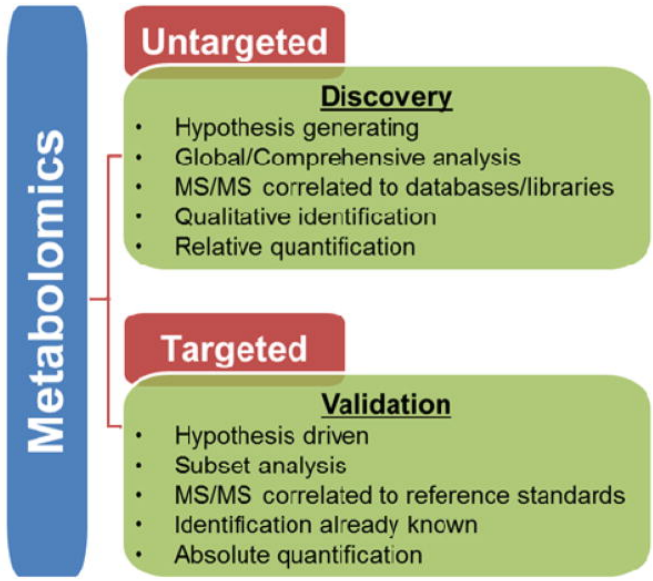


Figure 1.4: Untargeted versus targeted metabolomics studies (Schrimpe-Rutledge, Codreanu, Sherrod, & McLean, 2016).

1.1.3.2 Targeted metabolomics

In contrast, targeted (or validation-based) metabolomics focuses on measuring well-defined groups of metabolites, with opportunities for absolute quantification.

Targeted metabolomics aims to measure a predefined groups of biochemically characterized and interpreted metabolites (a metabolome subset). This reduced coverage of the metabolome means that targeted metabolomics is reliant on a *priori* knowledge of metabolites and their biochemical pathways, which hinders the discovery of novel metabolic targets. Moreover, through the use of isotopically labeled standards, targeted metabolomics can be made quantitative. The clear definition of the species measured reduces the likelihood of analytical artifacts progressed to ensuing experiments and data analysis (Roberts et al., 2012).

Due to the specificity of the data collected in a targeted metabolomics experiment, data processing and analysis tends to be less labor intensive when compared to untargeted metabolomics, as it is not necessary to identify chemical unknowns.

All data used in this thesis have been obtained via targeted metabolomics.

1.2 Nutrimetabolomics

Recent advances in high-throughput metabolomic approaches have provided an improved understanding of altered metabolic pathways, new gene functions, or the regulation of important enzymes. At the same time, the integration of metabolomics with nutritional science (nutritional metabolomics or nutrimetabolomics) enhances current clinical and research practices by providing a deeper insight into the relationships between various metabolites and health status (Ulaszewska et al., 2019).

In nutrimetabolomics, the aim is usually to investigate perturbations of the human metabolome by specific diets, foods, nutrients, microorganisms, or bioactive compounds.

As metabolomics is fundamentally phenotype-driven, nu-

trimetabolomics provides better and more individualized biomarkers[2] than other techniques and is expected to furnish better indicators of dietary effects on a target population or patients. Ultimately, the intertwining of nutrition and metabolomics in nutrimetabolomics aims to achieve personalized prognostic and diagnostic nutrition, making nutrimetabolomics one of the most promising avenues for improving the nutritional care and dietary treatment of patients in the future (Ulaszewska et al., 2019).

1.2.1 Nutritional studies

Nutritional studies can be divided into two main groups: interventional studies and observational (or cohort) studies.

1.2.1.1 Interventional studies

In controlled interventional studies, participants consume food items of interest in a single intake (acute study) or in repeated meals over a period of time (studies of medium or long term), from a few days up to 6 months or more (Scalbert et al., 2014). In acute studies, biofluids are collected over a time period of up to 24 hours after consumption of the food of interest and compared with participants consuming a control food (Scalbert et al., 2014). Urine is the preferred biofluid in this type of studies (Tebani & Bekri, 2019).

One limitation of these studies is the fact that the biomarkers identified may not be sufficiently specific for the test food in population studies, because regular diets may include other foods containing precursors of the same biomarkers. For instance, in a cross-sectional analysis of a whole-diet intervention study it was only possible to verify ~23% of potential biomarkers observed in previous-meal studies (Andersen et al., 2014).

In the context of this thesis, two interventional studies have been used to carry out most of the work presented later in Chapter 3. These are the EIT Health "Cook2Health" project and the "MaPLE" project

[2]The concept of "*biomarker*" is carefully detailed in "Biomarkers" section.

(Guglielmetti et al., 2020), respectively.

1.2.1.2 Observational studies

Observational studies can also play an important role in biomarker discovery. Low (or non-consumers) and high consumers are selected from food intake data collected by using FFQs, DRs or other dietary assessment techniques.[3] In this case, metabolomic profiles are compared between these subgroups to unveil potential dietary biomarkers that are reflective of habitual intake, as long as these biomarkers have a sufficient half-life in the organism or that the foods are regularly consumed (Scalbert et al., 2014).

Despite the potential of these studies, it has to be taken into account that many foods consumed are or can be highly correlated and there is a risk for identifying biomarkers that are not specific to the particular food of interest unless their identity and specific occurrence in the considered foods are established. Thus, these studies are mainly association studies and do not allow causal inference (Praticò et al., 2018). Intervention studies are, therefore, needed to validate the potential metabolite as a specific food intake biomarker.

This thesis also includes work done from an observational study. This is the D-CogPlast project, which has led to paper 5, presented later in the "Results" section.

1.2.2 Dietary assessment techniques

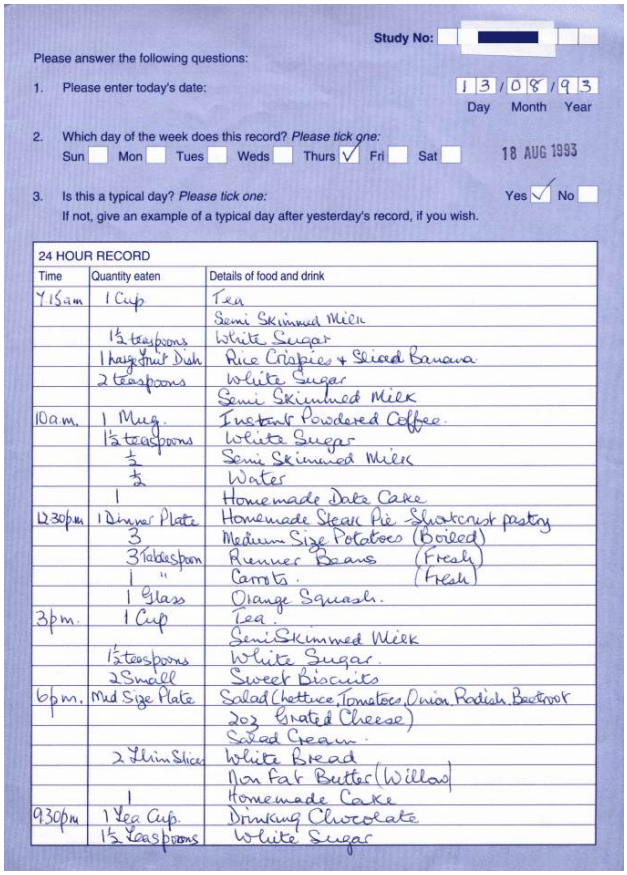
For decades, nutritional research has been a crucial pillar to unveil diet–health relationships at both individual and population scale. However, consistency, validation and reproducibility of the dietary assessments have been the great challenges (Tebani & Bekri, 2019). Despite all known drawbacks of these type of approaches, 24-hour dietary recalls (DRs) and food frequency questionnaires (FFQs) have been the most widely used approaches for assessing diet during the last

[3]Dietary assessment techniques are described below.

years (Park et al., 2018).

1.2.2.1 Dietary recalls

Dietary recalls include a structured collection of detailed information about food intake during the previous 24 hours.



Study No: [redacted]

Please answer the following questions:

1. Please enter today's date: 13 / 08 / 93 (Day / Month / Year)

2. Which day of the week does this record? *Please tick one:*
Sun ☐ Mon ☐ Tues ☐ Weds ☐ Thurs ☑ Fri ☐ Sat ☐ 18 AUG 1993

3. Is this a typical day? *Please tick one:* Yes ☑ No ☐
If not, give an example of a typical day after yesterday's record, if you wish.

24 HOUR RECORD

Time	Quantity eaten	Details of food and drink
7.15am	1 Cup	Tea
		Semi Skimmed Milk
	1½ teaspoons	White Sugar
	1 Large Fruit Dish	Rice Crispies + Sliced Banana
	2 teaspoons	White Sugar
		Semi Skimmed Milk
10a.m.	1 Mug.	Instant Powdered Coffee.
	1½ teaspoons	White Sugar
	½	Semi Skimmed Milk
	½	Water
	1	Homemade Date Cake
12.30pm	1 Dinner Plate	Homemade Steak Pie Shortcrust pastry
	3	Medium Size Potatoes (Boiled)
	3 Tablespoon	Runner Beans (Fresh)
	1 "	Carrots. (Fresh)
	1 Glass	Orange Squash.
3pm.	1 Cup	Tea.
		Semi Skimmed Milk
	1½ teaspoons	White Sugar.
	2 Small	Sweet Biscuits
6pm.	Med Size Plate	Salad (Lettuce, Tomatoes, Onion, Radish, Beetroot
		2oz Grated Cheese)
		Salad Cream.
	2 Thin Slices	White Bread
		Non Fat Butter (Willow)
	1	Homemade Cake
9.30pm	1 Tea Cup.	Drinking Chocolate
	1½ Teaspoons	White Sugar

Figure 1.5: A 24-Hour dietary recall.

Dietary recalls consists of an open form, where participants can report all kinds of meals and recipes they have eaten during the previous day (Figure 1.5). This makes the DR preprocessing step prior to their utilization quite long and complicated, sometimes leading to a loss of accuracy and specificity. In addition, one of the major other

drawbacks of this approach is that all these recalls are subjected to inter-day nutritional variation (Garden, Clark, Whybrow, & Stubbs, 2018).

1.2.2.2 Food frequency questionnaires

On the other hand, FFQs are dietary assessment tools delivered as a questionnaire (usually self-administered) to estimate frequency and, in some cases, portion size information about food and beverage consumption over a specified period of time, typically the past month, three months, or year. Unlike a DR, a FFQ is a closed questionnaire, often from 80 to 120 items (including foods and beverages).

FFQs are the preferred option in large-scale studies of diet and health as despite both DR and FFQ methods require an intensive preparation before implementation, the managing and processing of a well-validated FFQ is rather smooth for large studies (Tebani & Bekri, 2019).

However, this type of dietary assessment also presents certain drawbacks such as inaccurate estimation of portion size, socially distorted responses, lack of objectivity, poor accuracy of self-reported data, and errors in food intake composition, which hinder the FFQ and make it prone to record errors (Freedman, Schatzkin, Midthune, & Kipnis, 2011; Thompson & Subar, 2013).

FOODS AND AMOUNTS	AVERAGE USE LAST YEAR								
BREAD AND SAVOURY BISCUITS (one slice or biscuit)	Never or less than once/month	1-3 per month	Once a week	2-4 per week	5-6 per week	Once a day	2-3 per day	4-5 per day	6+ per day
White bread and rolls						✓			
Brown bread and rolls				✓					
Wholemeal bread and rolls	✓								
Cream crackers, cheese biscuits		✓							
Crispbread, eg. Ryvita		✓							
CEREALS (one bowl)									
Porridge, Readybrek				✓					
Breakfast cereal such as cornflakes, muesli etc.					✓				

Figure 1.6: A food frequency questionnaire, illustrating bread, savoury biscuits and breakfast cereals (Mulligan et al., 2014).

Additionally, FFQs can also be used to calculate different nutritional and health scores, for example, the Alternate Healthy Eating Index-2010 (AHEI-2010) (Chiuve et al., 2012).

The AHEI-2010 was developed as an alternative measure of diet quality to identify future risk of diet-related chronic disease (Leung et al., 2012; Leung, Epel, Ritchie, Crawford, & Laraia, 2014; Wang et al., 2014). The AHEI-2010 was originally developed on the basis of the FFQ (Chiuve et al., 2012). However, previous studies have also used DRs to compute the AHEI-2010 scores (Leung et al., 2012, 2014; Wang et al., 2014).

This index has been used to assess adherence to healthy dietary habits through the urinary food metabolome in paper 4, presented later in the "Results" section.

1.2.2.3 The need for a complementary approach

The growing field of metabolomics has enabled in-depth exploration of the food-related metabolome and the identification of potential food intake biomarkers (or dietary biomarkers) for objectively assessing dietary intake. These biomarkers may provide actionable information to fill in the gaps of self-reported dietary assessment methods given the tight connection between the metabolite production, food intake, microbiome, and health status (Bekri, 2016).

1.3 Biomarkers

The use of biomarkers in basic and clinical research as well as in clinical practice has become so commonplace that their presence as primary endpoints in clinical trials is now accepted almost without question. In the case of specific biomarkers that have been well characterized and repeatedly shown to correctly predict relevant clinical outcomes across a variety of treatments and populations, this use is entirely justified and appropriate. In many cases, however, the "validity" of biomarkers is assumed where, in fact, it should continue to be evaluated and

reevaluated (Strimbu & Tavel, 2010).

The term "*biomarker*" refers to a broad subcategory of medical signs (objective indications of medical state) which can be measured accurately and reproducibly. Medical signs stand in contrast to medical symptoms, which are limited to those indications of health or illness perceived by patients themselves. There are several more precise definitions of biomarkers in the literature, and they fortunately overlap considerably. In 1998, the National Institutes of Health Biomarkers Definitions Working Group defined a biomarker as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" (Strimbu & Tavel, 2010).

1.3.1 Dietary and health biomarkers

Dietary biomarkers enable accurate and objective dietary assessment, which can mitigate the inherent misreporting errors of traditional self-reported methods, and thus provide a closer and more reliable overview of the interplay between diet and health. Therefore, robust and informative biomarkers are needed to understand diet and lifestyle interaction in health and disease (Trepanowski & Ioannidis, 2018; Zeevi et al., 2015). However, two main challenges hinder the application and use of nutritional biomarkers; 1) there is no consensus on nutritional biomarker definition, its assessment and use, and 2) even the well-validated biomarkers lack consistency to support strong recommendations (Tebani & Bekri, 2019).

Nevertheless, one widely accepted dietary biomarker definition was proposed by the FoodBAll initiative (`https://foodmetabolome.org`), classifying dietary biomarkers as:

"*Specific measurements within the body that accurately reflect the intake of a food constituent or food*".

And they added:

"*These markers are measured in biofluids such as blood and urine,*

and include natural food constituents such as vitamins and fatty acids, in addition to certain food additives like iodine in milk or food contaminants like polychlorinated biphenyls in fatty fish. Dietary biomarkers can be used to measure nutritional status and food intake, to find associations between diet and disease outcomes, and to monitor dietary changes in populations".

1.3.1.1 Types of dietary biomarkers

Dietary biomarkers can be divided into two groups:

- **Biomarkers of intake (exogenous biomarkers)**

These biomarkers are those that come directly from diet, that is, those that can be found in the food metabolome. Those metabolites derived from the food metabolism and detectable in biofluids can be potential biomarkers of intake.

- **Biomarkers of effect (endogenous biomarkers)**

These biomarkers are those that are not in the food metabolome but can be altered by diet. Therefore, they are endogenous metabolites (from the endogenous metabolome) that can be altered by the ingestion of certain foods. In this way, biomarkers of effect allow to detect the consumption of certain foods, not from the compounds of these foods or their derivatives, but from reactions within the body of metabolites already present before consuming them.

1.4 Data analysis in nutrimetabolomics

Often, data analysis is one of the critical points in nutrimetabolomic studies. Data analysis is the process of systematically applying statistical and/or logical techniques to extract biological interpretations of the data. This process is composed of different preprocessing operations (e.g., missing value imputation, normalization), statistical

methods, and biological significance analysis[4] procedures (Figure 1.7).

However, there are a variety of methods to carry out data analysis processes, each with its different applications, advantages and disadvantages. For instance, machine learning methods tend to give good results in terms of predictive modeling and biomarker discovery, however, they are often hard to interpret and make it difficult to use complementary study information (e.g., confounding factors). On the other hand, other simpler statistical approaches, such as linear models, do not usually provide such satisfactory results in terms of prediction but allow the use of confounding factors and are often more interpretable. Thus, data analysis can be very complex, requiring researchers to choose and adapt the best methodologies for each type of study.

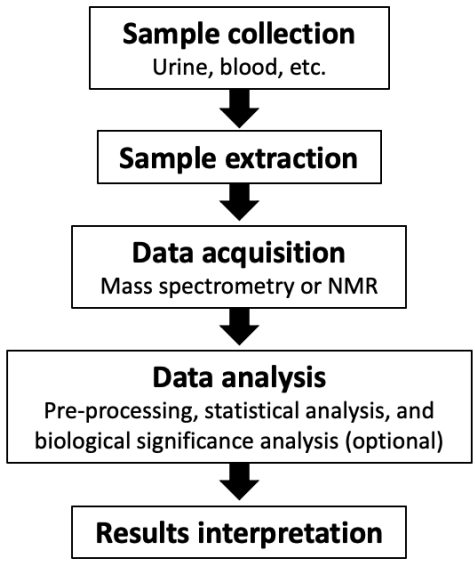


Figure 1.7: Key steps of a metabolomics study.

The POMAShiny web-based tool developed in the context of this thesis focuses on this section, providing different methods for an

[4]This concept is carefully detailed in "Biological significance analysis" section.

intuitive and user-friendly metabolomics data analysis (see Table 3.1).

1.4.1 Statistical modeling

Classical statistics provide researchers with a well-established set of tools for addressing several research questions, such as comparing the response to different treatments or modeling the effects of a set of variables on the concentration of a metabolite. Statistical models can be constructed with high flexibility and, in the process of model building and checking, it is possible to include external knowledge from the experimental design or knowledge of confounding variables (Ulaszewska et al., 2019). This kind of models are usually high interpretable models. Often, due to multiple collinearities or lack of relevant information in individual variables, only a few features are included in the model, facilitating its interpretation.

The most widely used statistical modeling methods in the field of nutrimetabolomics are linear models (LMs), generalized linear models (GLMs), and generalized additive models (GAMs).

1.4.1.1 Linear models

Linear models describe a continuous response variable as a function of one or more predictor features. Linear regression is a statistical method used to create a linear model (Equation 1.1). The model describes the relationship between a dependent variable Y (also called "response") as a function of one or more independent variables X_p (called "predictors").[5]

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \tag{1.1}$$

LMs are the basis for comparing outcome averages (both for numerical and categorical variables) across subpopulations and they can be extended to cope with a variety of more complex study designs. When the outcome variable is restricted (e.g., count or binary) the analytical framework should be extended to GLMs (Ulaszewska et al., 2019).

[5]Note that throughout this work, the number of variables or features is indicated with p.

1.4.1.2 Generalized linear models

GLMs are a framework for modeling response variables that are bounded or discrete. These models are used when modeling positive quantities (e.g., populations) that vary over a large scale, when modeling categorical data with a fixed number of choices that cannot be meaningfully ordered, or when modeling ordinal data (e.g., ratings on a scale from 0 to 5), where the different outcomes can be ordered but the quantity itself may not have any absolute meaning.

In the presence of grouped data, LMs and GLMs should be made hierarchical (mixed models) to fully take advantage of the groups defined in the study design. Such grouping includes repeated measures (both parallel and crossover), blocked designs, and multilevel data (Ulaszewska et al., 2019). The most used GLMs in nutritional and nutrimetabolomic studies are the logistic regression (Equation 1.2), for two-group studies (e.g., control and nutritional intervention), and the multinomial logistic regression, for studies with more than two groups (e.g., a study with four different treatments).

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p, \tag{1.2}$$

where $X = (X_1, ..., X_p)$ are p predictors.

1.4.1.3 Generalized additive models

Generalized additive models provide a general framework for extending a standard LM by allowing non-linear functions of each of the variables, while maintaining additivity (Equation 1.3) (James, Witten, Hastie, & Tibshirani, 2013). Just like LMs, GAMs can be applied with both quantitative and qualitative responses.

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + ... + f_p(x_{ip}) + \epsilon_i \tag{1.3}$$

It is common in nutritional studies to consider the time variable in statistical models. When the interest is to model trends in time, GAMs are a possible option to deal with this issue. However, due to their data driven nature, they require a vast amount of time points to provide

reliable results (Ulaszewska et al., 2019).

Unfortunately, the application of statistical models in multivariate omics can be difficult, with the major issue being unstable modeling arising from multiple feature collinearities (which can be present in both targeted and untargeted metabolomic experiments). Consequently, a preliminary step of feature selection is often necessary, which may result in loss of information (Ulaszewska et al., 2019).

1.4.2 Data mining

Data mining methods focus on the analysis of several features at a time, taking into account the different relationships between them. These methods can provide information about the structure of the data and different internal relationships that would not be observed with classical statistical models. However, the interpretation of these methods can be more complex.

The most widely used data mining methods in the field of nutrimetabolomics are principal components analysis (PCA) and partial least squares (PLS), with some of its variants.

1.4.2.1 Principal Components Analysis

PCA is an unsupervised method for dimension reduction of a n x p data matrix X. This method is either done by calculating the data covariance matrix and performing eigenvalue decomposition on this covariance matrix without considering any response variable Y.

The first principal component direction of the data is that along which the observations vary the most. For example, consider Figure 1.8, which shows the two first principal components for the ST000336 example data set provided in the POMA Bioconductor package. This data set consists of a targeted metabolomics study where 31 urine metabolites were measured from boys with Duchenne Muscular Dystrophy (DMD) and controls. Figure 1.8 suggests that subjects with positive scores on the first component tend to be DMD subjects,

while subjects with negative scores on the first component are more associated with non-DMD subjects. Otherwise, the second component seems to have no (or very little) association with the presence of the disease (note that the first component explains the 55.43% of the global variability while the second component only explains the 8.15%).

- **Principal Components Regression**

PCR approach involves constructing the first M principal components and using them as the predictors in a linear regression model. The key idea is that often a small number of principal components allow to explain most of the variability in the data, as well as the relationship with the response. In this case, is assumed that the directions in which $X_1, ..., X_p$ show the most variation are the directions that are associated with Y (James et al., 2013).

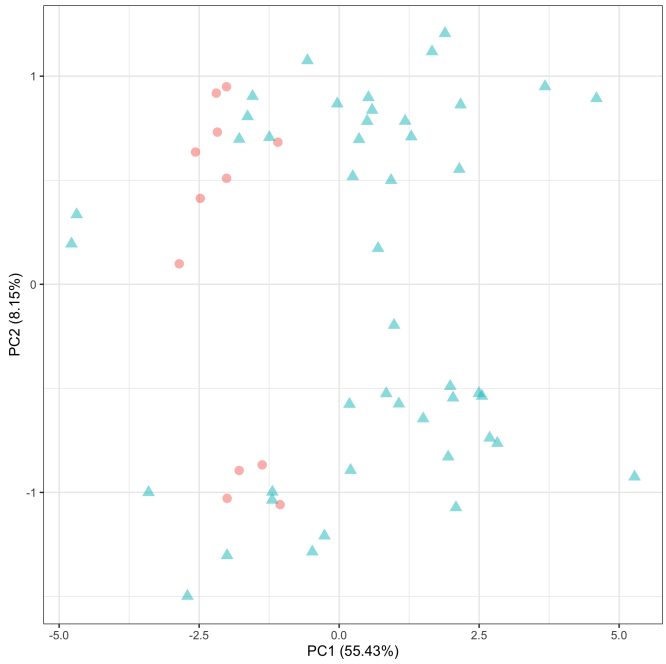


Figure 1.8: The first two principal components for the ST000336 data set provided in the POMA package. The red circles indicate control subjects and the blue triangles indicate DMD subjects.

1.4.2.2 Partial Least Squares

Partial Least Squares (PLS) is a supervised alternative to PCR. It is also a dimension reduction method, which first identifies a new set of features that are linear combinations of the original features, and then fits a LM via least squares using these new M features (James et al., 2013). Unlike PCR, PLS identifies these new features in a supervised way, that is, using the response Y. PLS method attempts to find directions that help explain both the response and the predictors (or features).

- **Partial Least Squares Discriminant Analysis**

PLS-DA is a PLS extension used when the response Y is categorical.

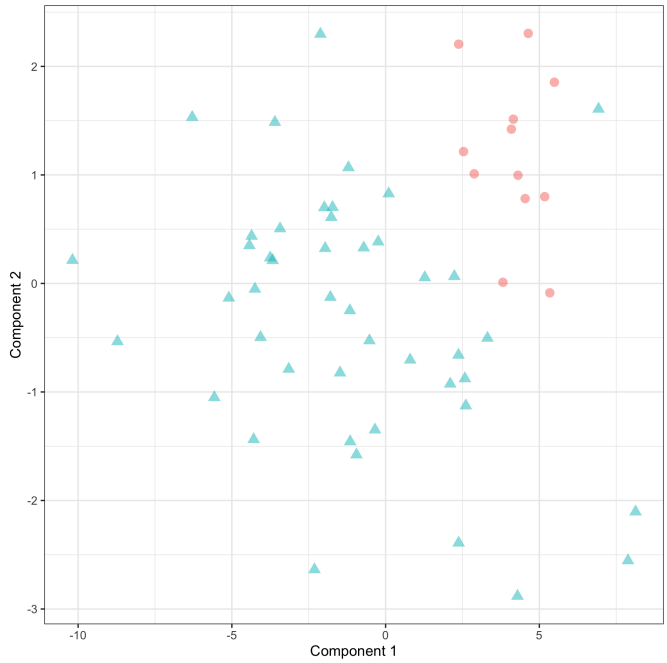


Figure 1.9: First two PLS-DA components for the ST000336 data set provided in the POMA package. The red circles indicate control subjects and the blue triangles indicate DMD subjects.

Figure 1.9 shows the two first components of a PLS-DA. The data used to generate this plot is exactly the same as the data used in the

PCA plot (Figure 1.8). PLS-DA plot shows a much smaller overlap between DMD subjects and controls than the overlap in the Figure 1.8. This difference is due to the use of the response variable Y information in the calculation of the components.

- **Sparse Partial Least Squares Discriminant Analysis**

Sparse PLS-DA performs variable selection and classification in a one-step procedure (Lê Cao, Boitard, & Besse, 2011). sPLS-DA is a special case of sparse PLS, where ℓ_1 penalization[6] is applied on the loading vectors associated to the data matrix X.

- **Orthogonal Partial Least Squares Discriminant Analysis**

OPLS-DA is a PLS extension where the data is separated into predictive and uncorrelated information. This leads to improved diagnostics, as well as more easily interpreted visualization. However, these changes only improve the interpretability, not the predictivity, which is equivalent to classification using standard PLS-DA. The corresponding predictive scores and loading vectors are therefore less subject to orthogonal variation. Variation that is unrelated to the class response Y is described in the orthogonal components (Boccard & Rutledge, 2013).

Note that while PCA is often used as a tool for data visualization (both samples and features) and unsupervised data exploration, PLS methods are more used for classification and feature selection purposes, respectively.

These multivariate methods can return attractive results, which unfortunately cannot be generalized for the all nutrimetabolomic studies. In order to avoid this situation, a thorough validation of these methods (internal cross-validation, permutation analysis or external cross-validation) is absolutely mandatory (Ulaszewska et al., 2019).

[6]This concept is described later in "The Lasso" section.

1.4.3 Statistical learning

In recent times, predictive statistical learning methods are receiving increasing attention for the nutrimetabolomics research. The increased use of these methods is due to the need to identify dietary biomarkers or panels of dietary biomarkers with predictive capacity in order to achieve the ambitions objective of predicting food intake from urine, blood, or serum metabolites, and not just identify those metabolites associated with certain dietary patterns or foods. Nowadays, this aim has become one of the most important goals in the field of nutrimetabolomics.

However, these methods may also present some drawbacks, such as the high number of samples required, or the less intuitive interpretation of the output provided by these methods, which is generally less easy to interpret than classical statistical models or PLS methods family, respectively.

1.4.3.1 The Lasso

As discussed above, a large number of p features are usually quantified in nutrimetabolomics studies. This means that a previous step of feature selection is often required, as modeling all features at once can be a tedious process, that may not be a problem for the accuracy of the prediction, but it can hinder the model interpretation.

The lasso is a regularized regression method very similar to least squares procedure when fitting a linear model, with the difference that the coefficients are estimated using a technique that shrinks them towards zero. This method uses a ℓ_1 penalty, which has the effect of forcing some of the coefficient estimates to be exactly equal to zero, selecting only the most important features and removing the less important ones (with the lower effects) (James et al., 2013). The lasso coefficients minimize the quantity:

$$\min_{\beta_0,\beta} \left\{ \frac{1}{N} \|y - \beta_0 - X\beta\|_2^2 \right\} \text{ subject to } \|\beta\|_1 \leq t, \tag{1.4}$$

where N is the number of observations and t is a prespecified free parameter that determines the amount of regularization.

Hence, the lasso performs a feature selection (or a *best subset* selection), purpose for which this method is often used instead of predictive purposes. As a result, models generated from the lasso are generally much easier to interpret than those produced using all p features in the study. This method produces sparse models, that is, models that involve only a subset of features (James et al., 2013).

This methodology has been widely used in the context of this thesis, being implemented as a resource in the POMA and POMAShiny tools, and being used for the data analysis in papers 4 and 5, presented later in the "Results" section.

1.4.3.2 Random forests

Random forests (Breiman, 2001) is one of the most popular and widely used machine learning methods. It is a flexible supervised learning algorithm and is part of tree-based methods (or decision trees), a collection of approaches for regression and classification problems.

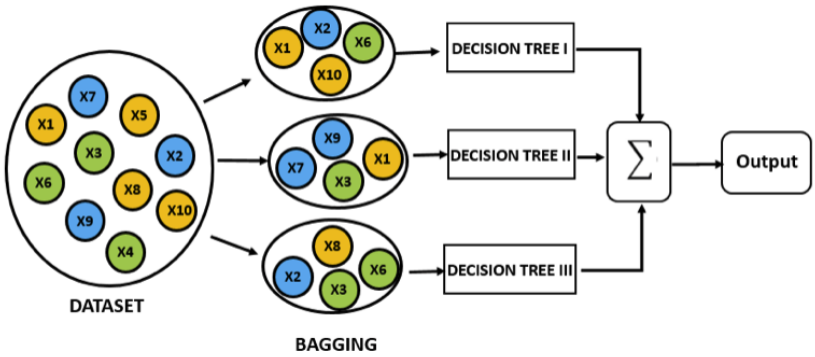


Figure 1.10: *Bagging* methodology scheme.

Random forests is a modification of bagging[7] (Figure 1.10) that builds a large collection of *de-correlated* trees, and then averages them

[7]Bagging is a technique used to reduce the variance of predictions by combining the results of several classifiers, each of them modeled with different subsets taken

(Friedman, Hastie, Tibshirani, & others, 2001).

In contrast to single decision trees, which are built on an entire data set and use all features, random forests is a collection of decision trees that randomly selects observations (or samples) and specific features to build multiple decision trees and then averages the results. Focusing on classification problems, after building a large number of trees, each tree "votes" a class (group label) and the class that receives the most votes by simple majority is the selected or predicted class (Figure 1.11).

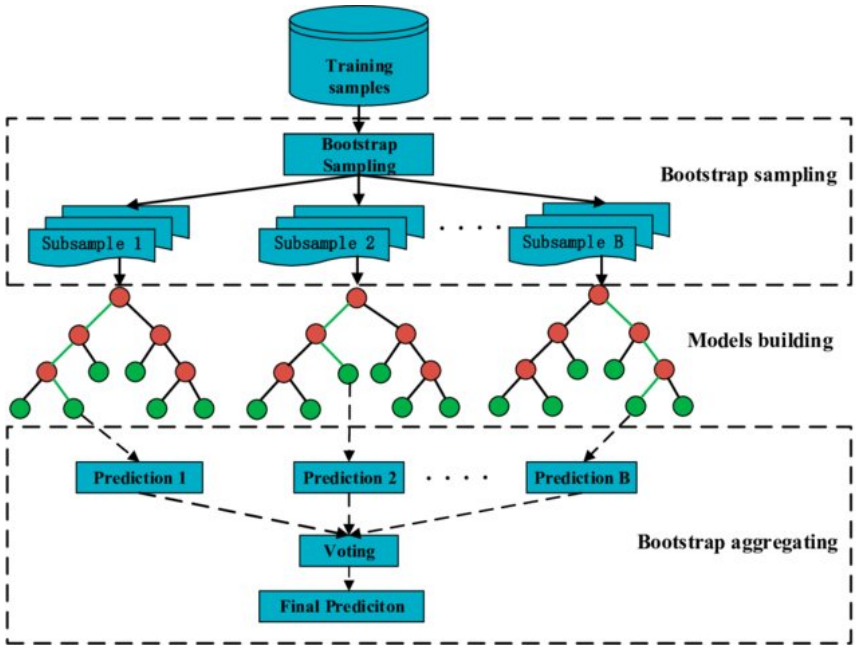


Figure 1.11: The scheme of random forest algorithm. The final prediction is obtained by taking a majority vote of the predictions from all the trees in the forest (Hongdong et al., 2019).

The Gini Impurity is a metric used in decision trees to determine how to split the data into smaller groups. This metric measures how often a randomly chosen record from the data set used to train the model will be incorrectly labeled (e.g., if half of the records in a group

from the same population.

are "A" and the other half of the records are "B", a record randomly labeled based on the composition of that group has a 50% chance of being labeled incorrectly).

In random forest, Gini Importance can be leveraged to calculate Mean Decrease in Gini, which is a measure of feature importance. Mean Decrease in Gini is an effective a measure of how important a feature is for estimating the value of the target variable across all of the trees that make up the forest. A higher Mean Decrease in Gini indicates higher feature importance.

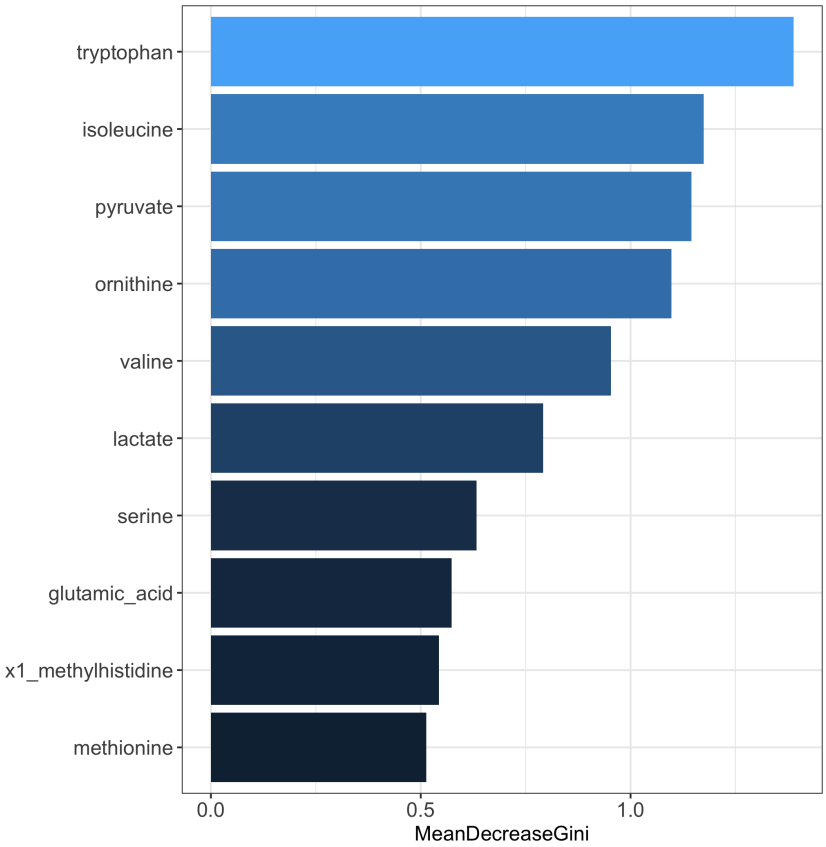


Figure 1.12: Mean Decrease in Gini of the top 10 selected features for the ST000336 data set provided in the POMA package.

Random forests has a variety of applications, such as image

classification and biomarker discovery. Consequently, this method is increasingly used in several *omics* studies, including nutrimetabolomics (Erban et al., 2019; Ulaszewska et al., 2019).

This methodology has been implemented as a resource (for classification problems) in the POMA and POMAShiny tools, presented later in the "Results" section.

Beyond the presented methods lasso and random forests, other approaches have also been proposed for prediction and data mining purposes in *omics* studies, such as support vector machines (SVM), genetic algorithms, and k-nearest neighbors (k-NN), among others. However, these methods are not explained here as they go beyond the scope of this thesis.

1.5 Ontologies

The growing emergence of high-throughput analytical techniques in the life sciences over the past three decades has created significant challenges in data management. Currently, one of the main problems that researchers face lies in the question: *where are these data and how can we use them?* Unfortunately, the heterogeneity of storage platforms, data formats and privacy requirements of some of them often hinders their widespread access and use.

In this vein, the creation of ontologies, defined as the "*specification of a representational vocabulary for a shared domain of discourse - definitions of classes, relations, functions and other objects*" (Kramer & Beißbarth, 2017), is of vital importance to help analyze, annotate and homogenize these large and complex data sets (Hoehndorf, Schofield, & Gkoutos, 2015; Schlegel, Ruttenberg, & Elkin, 2015). This is a major issue within the FAIR Guiding Principles for scientific data management and stewardship (Wilkinson et al., 2016), which aim to improve the findability, accessibility, interoperability and reusability of data. In particular, ontologies play a central role in the interoperability concept (Kramer & Beißbarth, 2017; Noy, McGuinness, & others, 2001), which establishes that

"*(meta)data has to use a formal, accessible, shared and broadly applicable language for knowledge representation*" (Wilkinson et al., 2016).

Thus, a proper ontology should provide a formal representation of knowledge in a certain domain (e.g., nutrition, metabolomics), in a way that people and computers can understand the concepts it contains and learn about the domain that is being represented (Rubin, Shah, & Noy, 2008). Ontologies are typically structured within a knowledge hierarchy where concepts are connected by standardized semantic relationships (e.g., "is a", "ingredient of") formally specifying knowledge relations such as generalizations of specifications of the domain of interest (Vitali et al., 2018).

The OWL (Web Ontology Language) is a knowledge representation language for authoring ontologies. The OWL is designed to be used by applications that need to process the content of information instead of just presenting information to humans. This language facilitates greater machine interpretability of web content than that supported by XML, and RDF by providing additional vocabulary along with a formal semantics.

However, ontologies can be expressed also using the OBO (Open Biological and Biomedical Ontology) format. The OBO project is a coordinated international collaborative effort to define standards and methodologies for the development of ontologies and controlled vocabularies in the life sciences, which defines the OBO format for ontology serialization.[8] The OBO project also provides the OBO Foundry (`http://www.obofoundry.org/`), a web registry for orthogonal and interoperable domain reference ontologies, which includes some of the most popular ontologies in the life sciences, such as the Gene Ontology, ChEBI, and FoodOn.

[8]Process of translating a data structure or object state into a format that can be stored or transmitted and reconstructed later.

1.5.1 The gold standard: The Gene Ontology

The Gene Ontology (or commonly called *GO*) is an annotation resource created and maintained by a public consortium whose main goal is to provide a "*dynamic, controlled vocabulary that can be applied to all eukaryotes even as knowledge of gene and protein roles in cells is accumulating and changing*" (Ashburner et al., 2000). This ontology has become one of the most successful resources used for many different purposes (e.g., annotation, biological interpretation) in biomedical research (Figure 1.13). Currently, the Gene Ontology contains more than 44.000 terms, divided into three specific sub-domains: *Biological process*, *Molecular function*, and *Cellular component*.

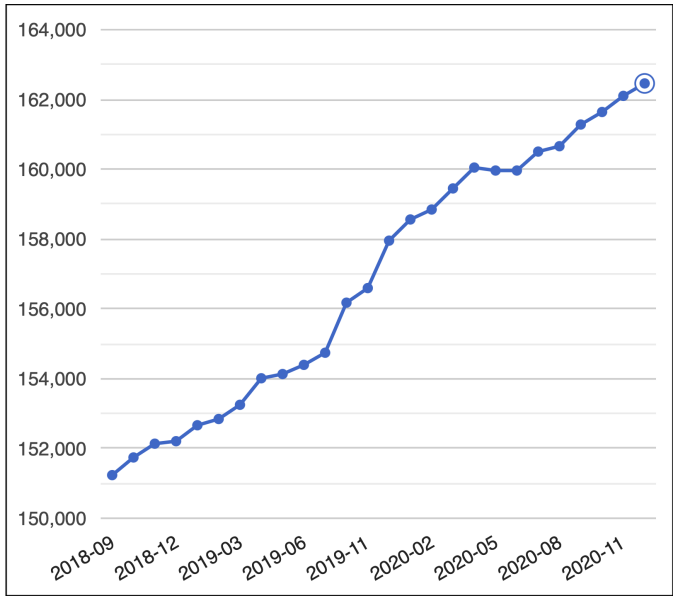


Figure 1.13: Growth of GO annotated scientific publications over time (from http://geneontology.org).

An example of the GO structure is illustrated in the Figure 1.14. The structure of GO can be described in terms of a graph, where each GO term is a node, and the relationships between the terms are edges between the nodes. GO, like other ontologies, has a hierarchical structure, where *child* terms are more specialized than their *parent* terms, but unlike a strict hierarchy, a term may have

more than one parent term. For instance, the biological process term *hexose biosynthetic process* has two parents, *hexose metabolic process* and *monosaccharide biosynthetic process*, showing that *biosynthetic process* is a subtype of *metabolic process* and a *hexose* is a subtype of *monosaccharide* (`http://geneontology.org`).

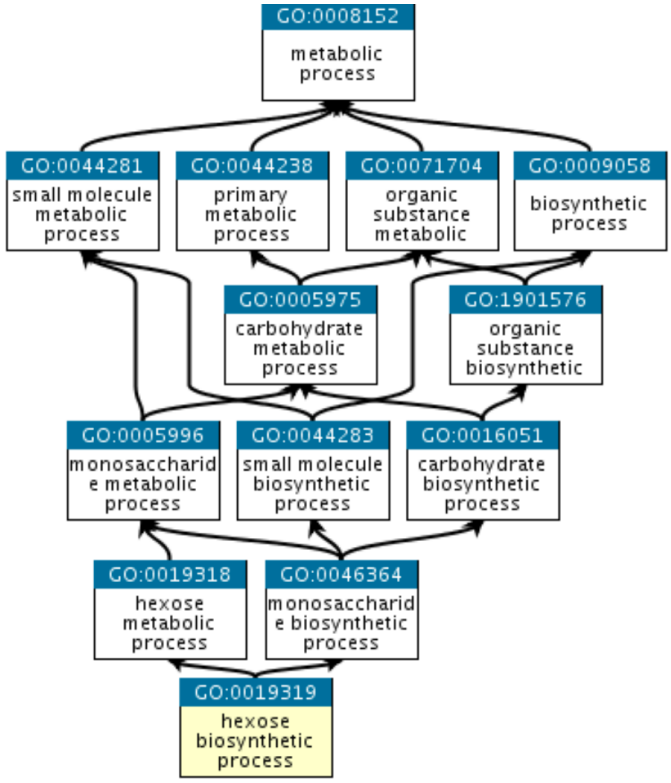


Figure 1.14: The structure of GO for the *hexose biosynthetic process* term (from http://geneontology.org).

1.5.2 Ontologies in metabolomics

In 2007, the Metabolomics Standards Initiative (MSI) (Sansone, Fan, et al., 2007) highlighted the importance of ontologies in the field of metabolomics (Sansone, Schober, et al., 2007). Its aim was to address the challenges associated with interpreting and integrating experimental process and data across disparate sources in metabolomics experiments by developing a common semantic

framework to enable metabolomics-user communities to consistently annotate the experimental process and to enable meaningful exchange of data sets (Sansone, Schober, et al., 2007).

Later in 2015, Schlegel et al. reported again that "*the application of ontologies to metabolomics can improve the consistency of study data and can help link data using relationships that extend the computational capacity of the study data and enrich that knowledge source with a myriad of nationally available data to help fuel hypothesis driven laboratory based research*" (Schlegel et al., 2015).

Today, the most well-known and used ontology in the field of metabolomics (to describe compounds) is the ChEBI ontology, encompassing a large number of chemicals and metabolites grouped into their corresponding chemical classes.

1.5.2.1 ChEBI: Chemical Entities of Biological Interest

ChEBI (`https://www.ebi.ac.uk/chebi/`) ontology is the reference ontology for describing chemical compounds of biological interest in terms of their chemical structures, chemical categories and roles (Degtyarenko et al., 2007).

This ontology contains more than 58.000 fully annotated compounds (being most of them manually annotated) and continues to grow over time (Figure 1.15).

An example of the ChEBI structure is illustrated in the Figure 1.16. Here, the term *phloretin* has only one parent, that is *dihydroalcones*, however, this second term has two parents, *carbonyl compound* and *flavonoids*, indicating that *phloretin* is a *carbonyl compound* and a *flavonoid(s)*, and belonging all the terms mentioned, to the *organochalcogen compound* class.

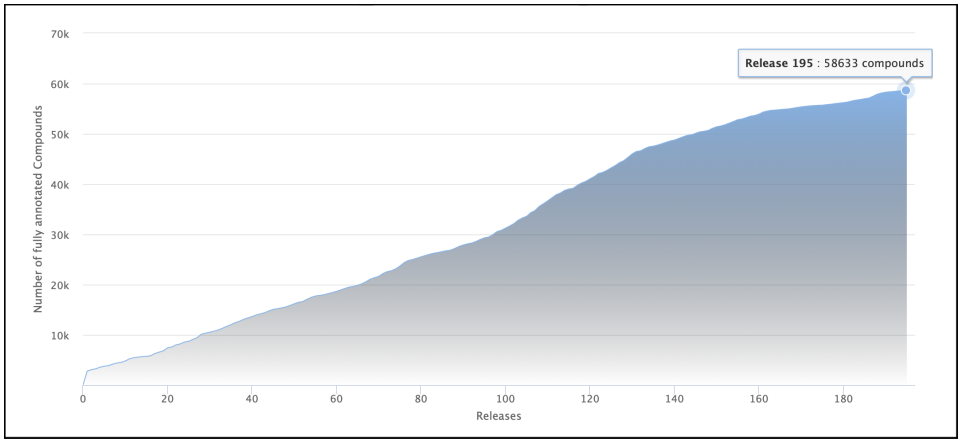


Figure 1.15: Growth of ChEBI annotated compounds over its releases (from https://www.ebi.ac.uk/chebi/).

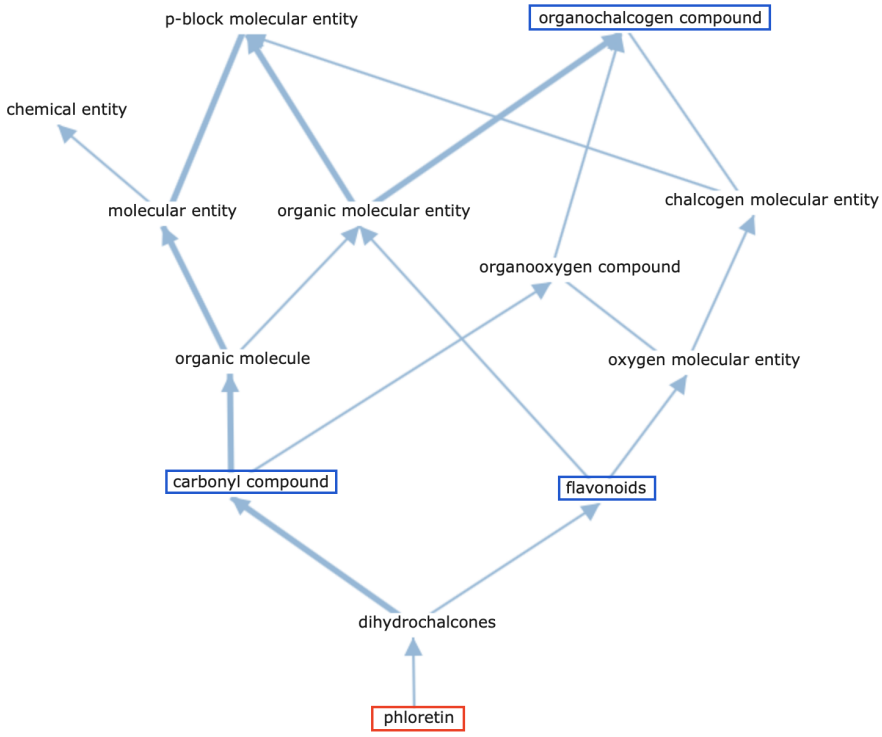


Figure 1.16: The structure of ChEBI for the *phloretin* term.

1.5.3 Ontologies in nutrition

Nutritional research largely relies on accurate dietary assessment, which is of great relevance to evaluate food intake and dietary habits. Dietary assessment also helps in understanding the association between nutrition and health status. Nutritional research is often conducted by using two complementary approaches: (i) self-reporting methods (e.g., food frequency questionnaires, dietary recalls) and (ii) the measurement of dietary biomarkers using a variety of analytical chemistry techniques, including metabolomics (Scalbert et al., 2014; Ulaszewska et al., 2019). With regard to traditional dietary assessment tools, it should be noted that subjective self-reports generate very complex textual data, containing types and quantities of foods and recipes in very diverse and heterogeneous formats that depend on the country/region, socio-demographic factors, etc.

For this reason, the creation of ontologies in the field of nutrition is extremely important, being of great help to jointly analyze different nutritional studies, or to define standardized nutritional terms in these studies.

The nutrition field has more specific ontologies than the field of metabolomics, being FoodOn the most prominent and well-known ontology in the field, followed by other very useful ontologies such as the ONS.

1.5.3.1 FoodOn: Food Ontology

FoodOn is the reference ontology in the field of nutrition. This ontology provides a controlled vocabulary to name all parts of animals, plants, and fungi which can bear a food role for humans and domesticated animals, as well as derived food products and the processes used to make them (Dooley et al., 2018).

The main aim of FoodOn is to help with the data harmonization issues that arise from the large vocabulary used to define foods and food-related concepts around the world. With more than 29.000 terms, FoodOn covers several basic raw food source ingredients, packaging methods, cooking methods, and preservation methods. Moreover, this

ontology also includes an upper-level consisting of a variety of product type schemes under which food products can be categorized (Figure 1.17).

As shown in Figure 1.17, the term *apple (whole, raw)* is linked by different properties to a large number of related terms. For example, the *apple (whole, raw)* term *develops from part of* the *apple tree as food source* term, and the *apple pie* term is an *apple food product*, which in turn derive from the *apple (whole, raw)* itself.

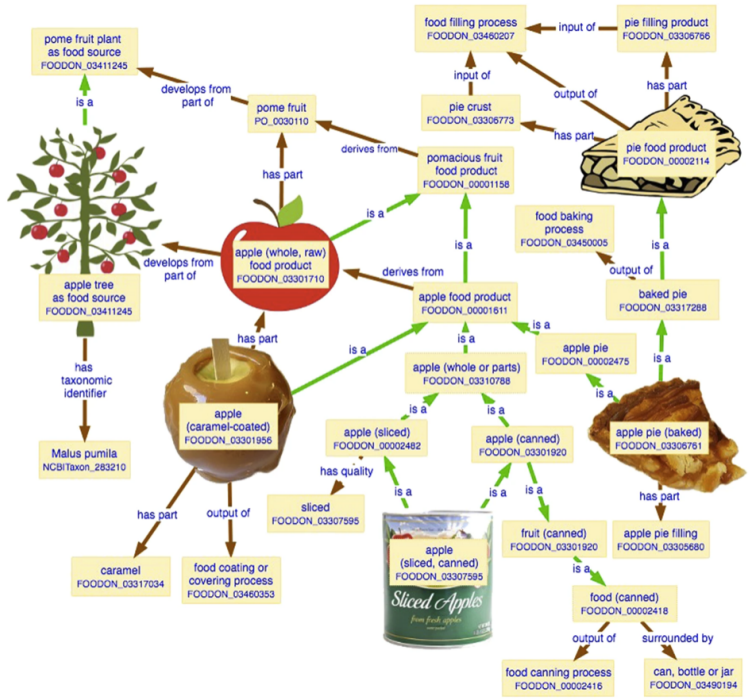


Figure 1.17: The structure of FoodOn for the *apple* term (Dooley et al., 2018).

This ontology is of great importance for the subsequent development of the FOBI ontology, presented later in the "Results" section.

1.5.3.2 ONS: Ontology for Nutritional Studies

Like the FoodOn, the ONS is part of the OBO Foundry. The ONS is a formal ontology for the description of nutritional studies. This ontology aims to establish an ontological framework that can help nutrition researchers by selecting the appropriate terms from the wide range of existing ontologies and creating the necessary terms that are still missing in the field (Vitali et al., 2018).

The ONS provides nutrition researchers of several nutritional-related, unified, and standardized terms collected in a single ontology, without having to resort to numerous ontology sources.

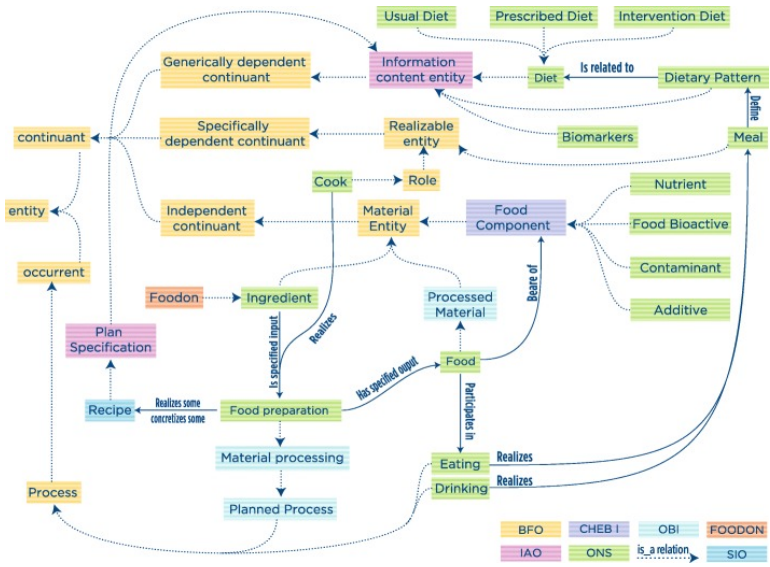


Figure 1.18: The ONS hierarchical structure. The terms in green boxes are ONS-specific terms, while terms in other color boxes are imported from existing ontologies (Vitali et al., 2018).

As shown in Figure 1.18, the ONS is made up of ONS-specific terms and terms of other ontologies. For instance, terms *Dietary pattern*, *Diet*, or *Biomarkers* are specific from ONS, while other terms belong to other ontologies. For example, *Food Component* term is adopted from ChEBI, and *Material processing* term is adopted from OBI.

1.5.4 Ontologies in nutrimetabolomics

Although most existing ontologies have been specifically designed for a single domain, there are also some others composed of interconnected sub-ontologies, thus enabling users to establish relationships among different concepts or domains. For instance, ChEBI is organized in two sub-ontologies: (i) "Molecular Structure", in which molecular entities are classified according to structure and (ii) "Subatomic Particle", which classifies particles smaller than atoms. Another example is the Gene Ontology, including three independent sub-ontologies: (i) "Biological process", referred to a biological objective to which the gene or gene product contributes; (ii) "Molecular function", defined as the biochemical activity of a gene product; and (iii) "Cellular component", which refers to the place in the cell where a gene product is active (Ashburner et al., 2000). In this regard, nutritional research also generates large amounts of complex and inter-related data coming from self-reporting methods and metabolomics experiments. Therefore, an interconnected set of sub-ontologies would be particularly useful for defining relationships between both metabolomics and diet, since no specific ontology exists for the field of nutrimetabolomics.

Currently, different complete and useful databases provide information on metabolites and foods, including Exposome-Explorer (Neveu et al., 2016), Phenol-Explorer (Rothwell et al., 2013), PhytoHub (`http://phytohub.eu/`), and Food Database (FooDB) (`http://foodb.ca/`) - all of which contain rich information about food constituents and food metabolites, defining in some cases the presence or absence of these metabolites in certain foods and in other cases, defining the concentrations of these metabolites in certain foods.

However, relationships between foods and their associated metabolites are extremely complex and the way they are described varies tremendously across these databases. This lack of commonality and the lack of a common, hierarchical structure makes data comparison and data searching quite difficult. Therefore, the development of a comprehensive ontology to clearly define the relationships between nutritional (food composition) and metabolomics (food associated metabolite or biomarker) data is needed (Maruvada et al., 2020). This ontology could have multiple practical applications in

nutrimetabolomics, being the annotation of terms using a consistent and standardized nomenclature the most basic one, but of great importance in this research field due to the inherent complexity and heterogeneity of the data managed (e.g., multiple names/synonyms to define the same food/metabolite). Additionally, other potential applications of this ontology would be the ability to perform different biological significance analyses and to conduct semantic similarity analyses (e.g., to establish novel associations between foods and metabolites).

In the context of this work, an ontology named FOBI (Food-Biomarker Ontology) has been created with the aim of providing a common language to describe the many complex relationships in nutrimetabolomics research. This new ontology will allow users to integrate dictionaries and analyze these two types of data independently or together in a consistent and homogeneous way.

The FOBI ontology is presented later in the "Results" section.

1.6 Biological significance analysis

Unlike the concept of "*statistical significance*", the "*biological significance*" concept can be treated from different perspectives and defined in different ways. For instance, this concept can be treated from a clinical point of view, referring to an effect that has a noteworthy impact on health, or from an *omics* point of view, referring to the biological interpretation or biological relevance of statistical differences in *omics* experiments. In this thesis, this concept is treated exclusively from an *omics* perspective.

Biological significance analysis (BSA), also known as *enrichment analysis*, *pathway enrichment analysis*, or *functional enrichment analysis*, denotes any method that benefits from biological pathway or network information to gain insight into a biological system (Creixell et al., 2015; Reimand et al., 2019). In other words, these type of analyses integrate the existing biological knowledge (from

different biological sources such as databases and ontologies) and the statistical results of *omics* studies, obtaining a deeper understanding of biological systems. Since BSA uses the results derived from statistical analysis, it is the last step of an *omics* data analysis process (Figure 1.7).

These methods can be of great help for interpreting the results and generating new hypotheses (Marco-Ramell et al., 2018). However, these methods are not always necessary, as they depend on the aim of the study. For instance, if the aim of the study is to find a panel of dietary biomarkers that accurately predict the intake of a food, the role of these metabolites in the body may not be of interest to researchers, since these predictive features are the result of the study itself. In this case, a BSA may not add any additional value to the study. Otherwise, if the aim of the study is to evaluate the impact of a specific food or diet on the metabolism, a BSA can be of great help in understanding the interactions and metabolic pathways that are being altered in the organism, as these may not be obvious just by observing the results derived from statistical analysis.

In most *omics* studies, the output of statistical analysis is usually a list of features selected as *statistically significant*[9] or *statistically relevant*[10] according to a predefined statistical criteria. BSA methods use these selected features to explore associated biologically relevant pathways, diseases, etc., depending on the nature of the input feature list (e.g., genes, metabolites) and the source used to extract the biological knowledge (e.g., GO, KEGG, FOBI). Thus, the input of BSA is usually a feature list (in some cases accompanied by the *fold change*[11] and/or the statistical significance of each feature to rank the

[9]In statistical hypothesis testing, a result is statistically significant when it is very unlikely to have occurred given the null hypothesis (H_0). The significance level of the study, α, is the probability of rejecting the H_0 when it is true, and the p-value, p, is the probability of obtaining the effect observed in a sample or larger, when the H_0 is true. The result is statistically significant when $p \leq \alpha$.

[10]The term *statistically relevant* is used when selecting features of interest without properly using a hypothesis contrast, therefore, without using p-values. For example, when using the VIP value in PLS-DA or the Mean Decrease in Gini in random forests, features can be ranked according to their importance on the study outcome but it is not correct to speak of "*significance*", then it will be called "*relevance*".

[11]Fold change (FC) is a measure that describes how much a quantity changes between two measurements. It is defined as the ratio between the two quantities;

list) and the output is usually a list of biological pathways with their associated statistical significance (Figure 1.19).

Therefore, these methods allow researchers to move from lists of genes or metabolites to metabolic pathways and/or diseases associated with those lists, and consequently, to the study design. Moreover, due to the large amount of easy-to-use web-based tools available for this purpose, BSA has become a very good ally for *omics* data analysis.

1.6.1 Biological significance analysis methods

Due to the large number of applications of these methods and its wide use in the *omics* field, different approaches for BSA have been proposed in the recent years. Currently, the most popular used approaches for BSA are the over-representation analysis (ORA) and the gene set enrichment analysis (GSEA), with its variants for other fields such as the metabolite set enrichment analysis (MSEA) (Xia & Wishart, 2010).

Some other powerful methods, such as pathway topology analysis, have been proposed to perform BSA in *omics* studies, however, this work will focus exclusively on the two most commonly used methods for this purpose: ORA and GSEA.

Figure 1.19 shows these two most common approaches to perform BSA, with their different required inputs. Considering this scheme, ORA requires a gene list with no order, just selected genes. On the other hand, GSEA requires a ranked list of genes, together with the metric used to rank them, such as p-value or FC. In both cases, ORA and GSEA, the output of the analysis is a table with the enriched pathways together with their associated statistical significance.

for quantities A and B, the fold change of B relative to A is B/A.

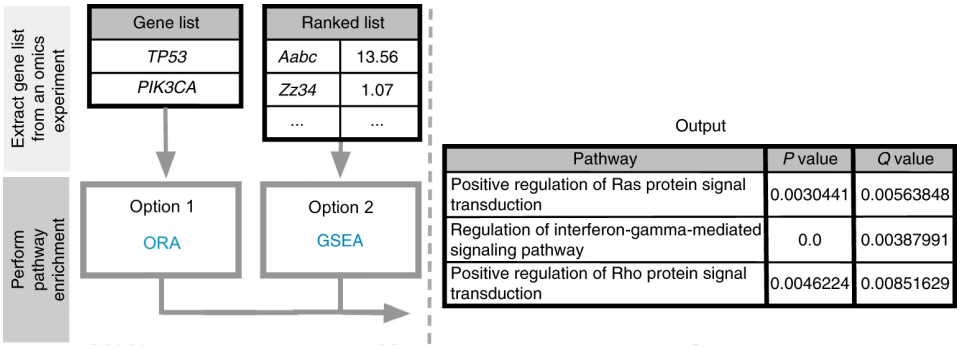


Figure 1.19: Gene lists derived from diverse *omics* data undergo pathway enrichment analysis, using ORA and GSEA methods, to identify pathways that are enriched in the experiment (adapted from Reimand et al., 2019).

1.6.1.1 Over Representation Analysis

ORA is one of the most used methods to perform BSA in *omics* studies due to its simplicity and easy understanding. This method statistically evaluates the fraction of features (e.g., metabolites or genes) in a particular pathway found among the set of features statistically selected (Khatri, Sirota, & Butte, 2012). Thus, ORA is used to test if certain groups of features (e.g., gene sets defined in the GO) are represented more than expected by chance given a feature list.

The p-value for over representation of the gene or metabolite set among the statistically significant features -or differentially expressed (DE) genes in the Table 1.2- is subsequently calculated using a test for independence in the 2×2 table (Table 1.2). Different tests have been proposed for testing this independence, including the χ^2 test, the hypergeometric test (Fisher's exact test) and the binomial z-test for proportions. Each of these tests is equivalent to a procedure that finds the H_0 of a test statistic by randomly reassigning genes to the labels for being in the gene set and for being DE. The differences are in the choice of the test statistic and whether the random reassignment is done with or without replacement (Goeman & Bühlmann, 2007).

Table 1.2: A 2×2 table for assessing over-representation analysis (from Goeman & Bühlmann, 2007).

	DE genes	**Non-DE genes**	**Total**
In gene set	p_{GD}	p_{GD^c}	p_G
Not in gene set	p_{G^cD}	$p_{G^cD^c}$	p_{G^c}
Total	p_D	p_{D^c}	p

However, ORA has a number of limitations. The most important one is the need of using a certain threshold or criteria to select the feature list. This means that features do not meet the selection criteria must be discarded. For example, excluding from the analysis those features with p-values (if this is the criterion used) greater than 0.05 and including those features with p-values lower or equal than 0.05. According to this criterion, features with p-values of 0.051 would be excluded and the features with p-values of 0.049 would be accepted, being this a nonsense in many scenarios.

The second big limitation of ORA is that this method assumes independence of pathways and features. In ORA, is assumed that each feature is independent of the other features and each pathway (or set) is independent of the other pathways (or sets). However, biology is a complex network of interactions and it is a misconception that different genes or metabolites are independent of each other, as well as that the different metabolic pathways are also independent (Khatri et al., 2012).

1.6.1.2 Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) methodology was proposed for the first time in 2005, with the aim of improving the interpretation of gene expression data. The main purpose of GSEA is to determine whether members of a gene set S tend to occur toward the top (or bottom) of the gene list L, in which case the gene set is correlated with the phenotypic class distinction (Subramanian et al., 2005).

This type of analysis basically consists of three key steps (Subramanian et al., 2005):

The first step consists on the calculation of an enrichment score (ES). This value indicates the degree to which a set S is over-represented at the extremes (top or bottom) of the entire ranked gene list L (Figure 1.20A). The ES is calculated by walking down the list L, increasing a running-sum statistic when a gene is found in S and decreasing it when a gene is not found in S. The magnitude of the increment depends on the correlation of the gene with the phenotype. The ES is the maximum deviation from zero encountered in the random walk (Figure 1.20B).

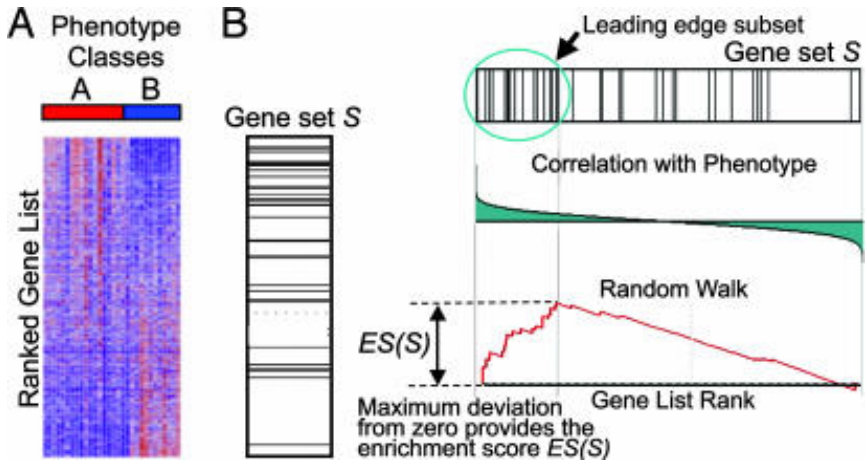


Figure 1.20: A GSEA overview illustrating the method. A) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the "gene tags" (location of genes from a set S within the sorted list). B) Plot of the running sum for S in the data set, including the location of the maximum enrichment score (ES) and the leading-edge subset (Subramanian et al., 2005).

The second step is the estimation of significance level of ES. The statistical significance (nominal p-value) of the ES is estimated by using an empirical phenotype-based permutation test that preserves the complex correlation structure of the gene expression data. The phenotype labels (L) are permuted and the ES of the S is recomputed for the permuted data, which generates a null distribution for the ES. The empirical, nominal p-value of the observed ES is then calculated relative to this null distribution. The permutation of class labels (groups) preserves gene-gene correlations and, thus, provides a

more biologically reasonable assessment of significance than would be obtained by permuting genes.

Finally, the third step consist on the adjustment for multiple hypothesis testing. When an entire database of gene sets is evaluated, the estimated significance level is adjusted for multiple hypothesis testing. First, the ES is normalized for each gene set to account for the size of the set, yielding a normalized enrichment score (NES). Then, the proportion of false positives is controlled by calculating the FDR[12] corresponding to each NES.

In 2010, a modification of the GSEA methodology was presented for metabolomics studies. This method was called Metabolite Set Enrichment Analysis (MSEA) and its main aim was to help researchers identify and interpret patterns of human and mammalian metabolite concentration changes in a biologically meaningful context (Xia & Wishart, 2010). MSEA is currently widely used in the metabolomics community and it is implemented and freely available at the known MetaboAnalyst web-based tool (Chong et al., 2018; Xia & Wishart, 2010).

As can be seen, GSEA approach is more complex than the ORA methodology, both in terms of methodological aspects and understanding of the method. Despite being widely used and covering some limitations of ORA, the use of ORA still prevails over GSEA, probably because of the mentioned complexity.

1.6.2 Biological significance analysis in nutrimetabolomics

Currently, BSA in nutrimetabolomics studies is conducted by non-specific tools (e.g., ORA and MSEA methods conceived for general metabolomics studies). These tools are available on different powerful and widely used platforms for metabolomics data analysis, such as the

[12]The False Discovery Rate (FDR) is a statistical approach used in multiple hypothesis testing to correct for multiple comparisons. It is defined as the expected proportion of type I errors (false positives).

MetaboAnalyst web-based tool (Chong et al., 2018).

Often, these enrichment analysis methods use databases such as KEGG (`https://www.genome.jp/kegg/`) or REACTOME (`https://reactome.org`) to obtain the information of different metabolite sets and metabolic pathways (from endogenous metabolome). However, the use of these generic databases, focused on metabolic pathways of different organisms, may exclude from analyses those exogenous compounds derived from diet and other substances not directly involved in metabolic pathways (e.g., food metabolome and drug metabolome).

As explained before, nutrimetabolomics aims to study both the endogenous compounds altered by diet, and the exogenous compounds derived from food metabolism. Therefore, current methods provide a solution for only the 50% of this purpose; the endogenous metabolites.

This current limitation claims for the development of new tools and methods that allow researchers to perform robust enrichment analyses using lists of metabolites derived from nutrimetabolomics studies, considering both exogenous and endogenous metabolites and their relationships with foods.

For this reason, different methods have been developed in the context of this thesis to make possible the new concept of "food enrichment analysis". These methodologies will allow researchers to obtain enriched food groups based on metabolite lists instead of classical enriched metabolic pathways.

This block constitutes one of the main objectives of this thesis and all methods developed for this purpose have been integrated within the fobitools framework, presented later in the "Results" section.

Chapter 2

Objectives

This thesis focuses on the development of methods and tools to advance in the discovery of the complex relationships between diet and food-derived metabolites, as well as to facilitate the interpretation of the results in nutrimetabolomic studies. All developed methods and tools will be published and/or distributed as open source resources, accompanied whenever possible by a web interface to make them more accessible to the community.

2.1 Main objective

The main objective of this thesis is the deeply study of the relationships between the food metabolome and food intake in the context of different nutrimetabolomic studies. This is a transversal objective and all specific objectives depend on it.

2.2 Specific objectives

Complementary to the main objective, the following list of specific objectives is considered:

1. The development of an ontology that defines clearly the relationships between metabolites and foods using a comprehensive and standardized common language.

    (a) The development of an open source tool that allows the easy query and use of the ontology created in the specific

objective 1. This tool will enable features such as:

i. Annotation of dietary free-text data.

ii. Biological significance analysis in nutrimetabolomic studies.

(b) Provide an open source GUI resource for the tool developed in the previous specific objective (a) to make it more accessible to the scientific community.

2. The development of a tool for statistical analysis of metabolomics data that includes alternative/complementary statistical methods to help improve the biomarker discovery process in the context of the nutrimetabolomic studies.

   (a) Provide an open source GUI resource for the tool developed in the specific objective 2 to make it more accessible to the scientific community.

3. The application of the developed tools in the specific objectives 1 and 2 to real nutrimetabolomic studies.

   (a) Identify metabolites or groups of metabolites associated with the AHEI-2010 healthy diet index.

   (b) Identify metabolites or groups of metabolites associated with disease risk or health status.

Part II

Results and Discussion

Thesis directors report

UNIVERSITAT DE BARCELONA

Dr. Alex Sánchez Pla
Departament de Genètica, Microbiologia i Estadística
Facultat de Biologia
asanchez@ub.edu

Dra. Cristina Andrés Lacueva
Departament de Nutrició, Ciències de l'Alimentació i Gastronomia
Facultat de Farmàcia
candres@ub.edu

INFORME DEL DIRECTOR DE TESI DEL FACTOR D'IMPACTE DELS ARTICLES PUBLICATS. En cas que es presenti algun treball fet en coautoria, caldrà incloure també un INFORME DEL DIRECTOR de la tesi (signat), en què s'especifiqui exhaustivament quina ha estat la participació del doctorand/a en cada article i si algun dels coautors d'algun dels treballs presentats en la tesi doctoral ha utilitzat, implícitament o explícitament aquests treballs per a l'elaboració de la tesi doctoral.

El **Dr. Alex Sánchez Pla** i la **Dra. Cristina Andrés Lacueva,** directors d'aquesta tesi doctoral, titulada "**Statistical Methods for Intake Prediction and Biological Significance Analysis in Nutrimetabolomic Studies**"

INFORMEN

Que aquesta tesi doctoral està elaborada com a compendi de les 2 publicacions següents:

1. **Pol Castellano-Escuder**, Raúl González-Domínguez, David S. Wishart, Cristina Andrés-Lacueva, Alex Sánchez-Pla (2020). *FOBI: An ontology to represent food intake data and associate it with metabolomic data*. Database, 2020.

   - Factor d'impacte (actual i dels últims 5 anys) i posicionament (quartil i decil): **[IF = 2.59 (3.66); Q2, D3]**. Ocupa la posició **15** de **59** dins de la categoria "Mathematical & Computational Biology".

2. **Pol Castellano-Escuder**, Raúl González-Domínguez, Francesc Carmona-Pontaque, Cristina Andrés-Lacueva, Alex Sánchez-Pla (2021). *POMAShiny: a user-friendly web- based workflow for metabolomics and proteomics data analysis*. PLOS Computational Biology, *Acceptat*.

   - Factor d'impacte (actual i dels últims 5 anys) i posicionament (quartil i decil): **[IF = 4.7 (5.26); Q1, D2]**. Ocupa la posició **6** de **59** dins de la categoria "Mathematical & Computational Biology" i la posició **9** de **77** dins de la categoria "Biochemical Research Methods".

A la primera publicació, l'estudiant de doctorat Pol Castellano Escuder va dur a terme tota la programació de l'ontologia FOBI, així com el desenvolupament de la interfície gràfica de visualització de l'ontologia presentada a l'article. En aquest treball, l'estudiant també va redactar majoritàriament el primer esborrany del manuscrit.

A la segona publicació, el doctorand va realitzar tota la part de programació tant del paquet de Bioconductor POMA, com de la interfície gràfica POMAShiny. Aquests dos desenvolupaments inclouen també la programació d'estructures de testeig i validació, de diferents casos d'ús i del lloc web de l'eina. En aquest treball, l'estudiant també va redactar el primer esborrany del manuscrit.

Pel que fa als altres articles inclosos a la secció de resultats, aquests han estat enviats a revistes científiques però encara no han estat acceptats, pel que no s'inclouen en aquest informe.

UNIVERSITAT DE BARCELONA

Dr. Alex Sánchez Pla
Departament de Genètica,
Microbiologia i Estadística
Facultat de Biologia
asanchez@ub.edu

Dra. Cristina Andrés Lacueva
Departament de Nutrició, Ciències
de l'Alimentació i Gastronomia
Facultat de Farmàcia
candres@ub.edu

A la publicació 3, el doctorand va tenir una implicació complerta com a la publicació 2, realitzant tota la part de programació tant del paquet de Bioconductor fobitools, com de la interfície gràfica fobitoolsGUI. Aquests dos desenvolupaments inclouen també la programació d'estructures de testeig i validació, de diferents casos d'ús i del lloc web de l'eina. En aquest treball, l'estudiant també va redactar el primer esborrany del manuscrit.

Donat que tots els desenvolupaments publicats en aquesta tesi són projectes de codi obert, la participació del doctorand en les diferents publicacions és fàcilment traçable a través l'historial de contribucions de GitHub.

- FOBI: https://github.com/pcastellanoescuder/FoodBiomarkerOntology/commits/master
- POMA: https://github.com/pcastellanoescuder/POMA/commits/master
- POMAShiny: https://github.com/pcastellanoescuder/POMAShiny/commits/master
- fobitools: https://github.com/pcastellanoescuder/fobitools/commits/master
- fobitoolsGUI: https://github.com/pcastellanoescuder/fobitoolsGUI/commits/master

A les publicacions 4 i 5, el doctorand ha realitzat la major part de les anàlisis estadístiques, controls de qualitat, neteja de dades, validació i interpretació dels resultats, i en el cas de l'article 4, ha contribuït en bona part a l'escriptura del manuscrit.

Els altres articles enviats a revistes científiques i presentats a l'annex B d'aquest treball (articles 6,7 i 8), han resultat de contribucions científiques on el doctorant, fent servir en la majoria dels casos les eines computacionals desenvolupades en la seva tesi doctoral, ha realitzat una part de les anàlisis estadístiques.

Dr. Alex Sánchez Pla
Departament de Genètica,
Microbiologia i Estadística
Universitat de Barcelona

Dra. Cristina Andrés Lacueva
Departament de Nutrició, Ciències
de l'Alimentació i Gastronomia
Universitat de Barcelona

Chapter 3

Results

All publications presented in this thesis have been submitted to international peer-reviewed journals. In this chapter, each publication is presented as a brief summary of the paper content, presenting the background, aim, results, and conclusions of the work, together with the journal impact factor, quartile and decile (if the manuscript has already been accepted).

The publications are divided into methodological developments and applications of the developed methods. In the methodological papers the main contributions fulfill the objectives of the thesis while in the application papers, the developed methodology and tools are applied to discover and identify metabolites associated with diet (dietary biomarkers), AHEI-2010, and health status or disease.

The last section of this chapter is devoted to all software, both R packages and GUIs, developed in the context of this thesis.

Those publications and developed tools not directly related to the contents of the thesis are also included in the Appendix B.

3.1 Methodological and software developments

3.1.1 Paper 1: Food-Biomarker Ontology

Pol Castellano-Escuder, Raúl González-Domínguez, David S. Wishart, Cristina Andrés-Lacueva, Alex Sánchez-Pla (2020). *FOBI: An ontology to represent food intake data and associate it with metabolomic data.* Database, 2020.

- Journal impact factor: **2.593**
- Journal quartile/decile: **Q2/D3** (15 of 59 - Mathematical & Computational Biology)

3.1.1.1 Background

As seen in section "Ontologies in nutrimetabolomics", the complexity and heterogeneity of metabolomics and nutritional data often hinder their analysis and integration. Moreover, the relationships between foods and their associated metabolites are extremely complex and the way they are described varies tremendously across different databases. This lack of commonality makes imperative the development of a comprehensive ontology to clearly define these relationships (Maruvada et al., 2020).

3.1.1.2 Aim

The aim of this work was to develop an ontology to describe the many complex relationships between diet-derived metabolites and foods through a standardized common language.

3.1.1.3 Results

The ontology developed is called FOBI (Food-Biomarker Ontology) and describes foods and their associated metabolite entities in a hierarchical way using a formal naming system, category definitions, properties, and relations between nutritional and metabolomics data. FOBI is composed of two interconnected sub-ontologies. One is

a "Food Ontology" consisting of raw foods and multi-component foods while the second is a "Biomarker Ontology" containing food intake biomarkers classified by their chemical classes. These two sub-ontologies are conceptually independent but interconnected by different properties. This allows data and information regarding foods and food biomarkers to be visualized in a bidirectional way, going from metabolomics to nutritional data or vice versa. Potential applications of FOBI ontology include the annotation of foods and biomarkers using a well-defined and consistent nomenclature, the standardized reporting of metabolomics workflows (e.g., metabolite identification, experimental design), and the ability to perform different biological significance analyses in nutrimetabolomic studies.

Figure 3.1 illustrates the FOBI architecture considering the *apple* term as an example. According to this, apple can be a raw food with the following relationships: "apple *is_a* pomaceous fruit food product *is_a* plant fruit food product *is_a* Fruits and vegetables *is_a* Food" (the property *is_a* is represented by blue arrows). In addition, apple can also be an ingredient in multi-component foods such as apple pie, so that "apple *IsIngredientOf* apple pie *is_a* bakery product *is_a* multi-component food *is_a* Food" as well "apple pie *Contains* apple" (the properties *IsIngredientOf* and *Contains* are represented by orange arrows). Considering phloretin and 5-(3',4'-dihydroxyphenyl)-γ-valerolactone as biomarkers of apple intake, they can be categorized as "phloretin *is_a* 2'-Hydroxy-dihydrochalcone *is_a* Chalcones and dihydrochalcones *is_a* Linear 1,3-diarylpropanoid *is_a* Phenylpropanoids and polyketides *is_a* Biomarker" and "5-(3',4'-dihydroxyphenyl)-γ-valerolactone *is_a* Catechol *is_a* Benzenediol *is_a* Phenol *is_a* Benzenoid *is_a* Biomarker". Because phloretin is a specific marker of apple, this metabolite is exclusively connected via the Food Ontology by the relationships "phloretin *BiomarkerOf* apple" and "apple *HasBiomarker* phloretin" (the properties *BiomarkerOf* and *HasBiomarker* are represented by yellow arrows). On the other hand, 5-(3',4'-dihydroxyphenyl)-γ-valerolactone can be derived from various procyanidin-rich foods (cacao, tea), so it can be connected with them following the same structure described for apple.

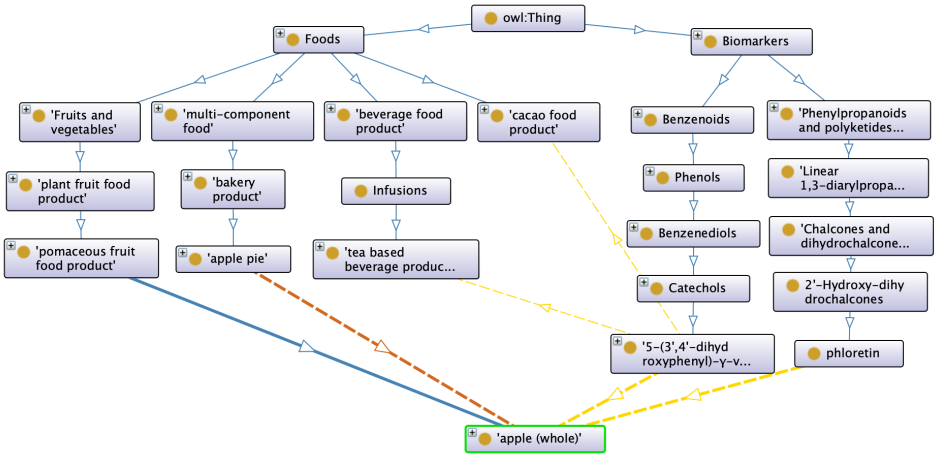


Figure 3.1: The structure of FOBI for the *apple* term (Castellano-Escuder et al., 2020).

FOBI is part of the OBO Foundry project and its identifiers have been indexed into the HMDB (Human Metabolome Database) and FooDB databases to facilitate the interoperability and the exchange of data.

FOBI is freely available in both OWL and OBO formats at the project's Github repository (`https://github.com/pcastellanoescuder/FoodBiomarkerOntology`).

3.1.1.4 Conclusion

FOBI is the first ontology that integrates nutritional and metabolomics data in a comprehensive common language. At the moment, FOBI has a total of 1197 terms (366 from Food sub-ontology and 831 from Biomarker sub-ontology), 11 chemical top-level classes, 13 food top-level classes and 4 different properties that are fully defined and which have clear relationship mappings. FOBI defines the relationships between foods and their metabolites (biomarkers) through a formal ontology.

FOBI allows experts to annotate and analyze nutritional and metabolomics data in a consistent way, making the results comparable

between and across studies in the same field. The development of FOBI will lead to an improvement in the interoperability of nutritional and nutrimetabolomic data thereby making the data sets generated from these studies fully FAIR compliant.

3.1.2 Paper 2: POMAShiny

Pol Castellano-Escuder, Raúl González-Domínguez, Francesc Carmona-Pontaque, Cristina Andrés-Lacueva, Alex Sánchez-Pla (2021). *POMAShiny: a user-friendly web-based workflow for metabolomics and proteomics data analysis*. PLOS Computational Biology, 2021.

- Journal impact factor: **4.7**
- Journal quartile/decile: **Q1/D2** (6 of 59 - Mathematical & Computational Biology)
- Journal quartile/decile: **Q1/D2** (9 of 77 - Biochemical Research Methods)

3.1.2.1 Background

As seen in section "Data analysis in nutrimetabolomics", statistical analysis is one of the critical points in nutrimetabolomics data analysis and it is critical in the subsequent biological interpretation of the results. Due to this fact combined with the computational programming skills needed for this type of analysis, several bioinformatic tools have emerged to simplify metabolomics data analysis. However, sometimes the analysis is still limited to a few hidebound statistical methods and to a low-flexible data sets.

Currently, statistical analysis of metabolomics data is mainly conducted by using several programming tools (Stanstrup et al., 2019) and/or via different web-based tools (Chong et al., 2018; Davidson, Weber, Liu, Sharma-Oates, & Viant, 2016; Giacomoni et al., 2015; Tautenhahn, Patti, Rinehart, & Siuzdak, 2012) according to the aims defined by researchers. Often, web tools are a very popular choice for the community as they provide a fast and easy-to-use GUIs and bring the statistical analysis closer to the community without extensive programming skills. These web-based tools are very useful and consequently, widely used by scientific community. However,

additional statistical approaches that are not implemented in these tools can be really useful in the analysis of these data.

3.1.2.2 Aim

The aim of this work was to develop a web-based tool for metabolomics data analysis to provide alternative statistical methods for improving the biomarker discovery process in the context of nutrimetabolomics studies.

3.1.2.3 Results

POMAShiny is a web-based tool that provides a structured, flexible and user-friendly workflow for the visualization, exploration, and statistical analysis of metabolomics data.

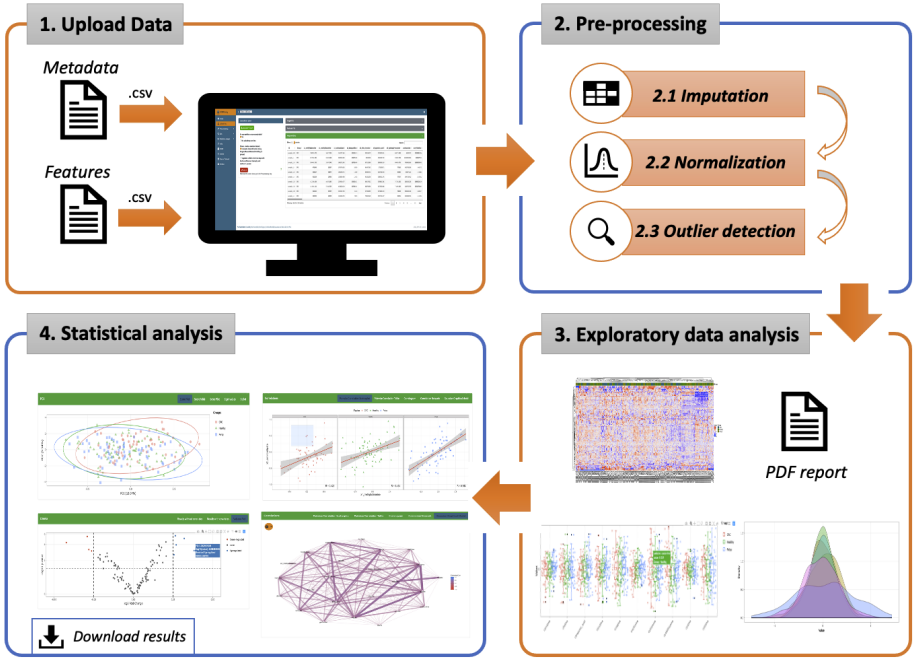


Figure 3.2: POMAShiny's workflow (Castellano-Escuder et al., 2021).

POMAShiny provides an analysis workflow structured in four sequential and well-defined panels: 1) data upload, 2) preprocessing, 3) exploratory data analysis (EDA), and 4) statistical analysis, all of them with their respective sub-panels (Figure 3.2).

This tool integrates several statistical methods (see Table 3.1), some of them widely used in other type of *omics*, and it is based on the POMA R/Bioconductor package, which increases the reproducibility and flexibility of analysis outside the web environment.

POMAShiny and POMA are both freely available at `https://github.com/pcastellanoescuder/POMAShiny` and `https://github.com/pcastellanoescuder/POMA`, respectively.

3.1.2.4 Conclusion

Despite the complexity of metabolomics and nutrimetabolomics data, many of the most widely used web tools for the statistical analysis of these data are not very versatile in the input data structure and limit the analysis to a few statistical methods. POMAShiny is a web-based tool that aims to cover some of these data analysis bottlenecks.

POMAShiny offers an integrated metabolomics data analysis workflow with a wide range of possibilities both for data preprocessing and statistical analysis, including outlier detection methods, flexible exploratory data analysis operations, downloadable reports and several statistical methods from simpler approaches such as univariate statistics to more complex methods such as regularized regression and machine learning algorithms. It requires two files as an input -the target and features file- giving users the possibility to include relevant study covariates (or confounding factors) in the analysis.

This intuitive and powerful web interface allows users to perform an integrated data analysis in an interactive, well documented and extremely user-friendly web environment, making data analysis process more accessible to a wide range of researchers not so familiar with the computing and/or statistical fields.

Table 3.1: Statistical methods provided in POMAShiny. *Methods that allow the use of covariates.

Univariate methods	Parametric	T-test (paired/unpaired)
		ANOVA
		ANCOVA*
		Limma*
	Non-parametric	Mann-Whitney U test (paired/unpaired)
		Kruskal-Wallis
Multivariate methods	Unsupervised	PCA
		k-means
		Multidimensional scaling (MDS)
	Supervised	PLS-DA
		sPLS-DA
Correlation methods	Parametric	Pearson's correlation*
	Non-parametric	Spearman's correlation*
		Kendall's correlation*
	Visualization	Gaussian graphical models (GGMs)
Statistical learning methods	Regularized regression	LASSO regression
		Ridge regression
		Elasticnet regression
	Decision trees	Random forest
Generalized linear models	Logistic regression	Odds ratio calculation*
Permutation tests	Non-parametric	Rank products

3.1.3 Paper 3: The fobitools framework

Pol Castellano-Escuder, Cristina Andrés-Lacueva, Alex Sánchez-Pla. *The fobitools framework: The first steps towards food enrichment analysis. Under review.*

3.1.3.1 Background

As seen in section "Ontologies in nutrimetabolomics", the field of ontologies in life sciences has not stopped growing since the development of the Gene Ontology around 2000 (Ashburner et al., 2000). Currently, this field continues to grow and a wide variety of ontologies have already been developed to cover many different life science domains (Hoehndorf et al., 2015). Despite the growing exploitation of these ontologies in many research areas, some of them are still underexploited due to the programming skills needed in many cases to operate with them.

Among the wide range of applications that ontologies have, probably the most widely known and used is enrichment analysis (see "Biological significance analysis" section). Often, enrichment analysis methodologies explore enriched metabolic pathways or biological processes given a list of genes derived from *omics* experiments. But what if the list to be tested is a list of metabolites from a nutrimetamolomics study where a dietary intervention has been carried out? In this case, a conventional metabolomics enrichment analysis may be useful for exploring altered metabolic pathways if the tested metabolites are endogenous (dietary biomarkers of effect). However, if the study measures exogenous metabolites derived from food and not present in the body (dietary biomarkers of intake), a conventional analysis will not provide any valuable insights. In this work, the food enrichment analysis concept is proposed for the first time. The fobitools framework provides methodologies for performing food enrichment analyses on nutrimetabolomic studies, that is, using the FOBI information to explore enriched foods or food groups (instead of pathways) given a list of exogenous metabolites (e.g., from the food metabolome).

Additionally, other useful features, such as the FOBI network

interactive visualization and the automatic annotation of dietary free-text data using the FOBI information are also provided in this tool.

3.1.3.2 Aim

The main aim of this work was to provide a set of tools for interacting and using FOBI ontology. Thus, facilitating and extending the use of FOBI ontology to the scientific community, not only to researchers with high programming skills but to experts of nutrition and metabolomics fields with limited programming experience.

3.1.3.3 Results

The fobitools framework is composed of the fobitools R/Bioconductor package and the fobitoolsGUI web-based tool. Both fobitools package and fobitoolsGUI web application are freely available at their GitHub repository: `https://github.com/pcastellanoescuder/fobitools` and `https://github.com/pcastellanoescuder/fobitoolsGUI`, respectively.

The fobitools framework is structured in five main functionalities, all of them executable from the R command line (via the fobitools R/Bioconductor package) and from the web interface (via the fobitoolsGUI application).

- **Food enrichment analysis**

ORA and MSEA are the most widely used methodologies for performing enrichment analysis in metabolomics studies (Marco-Ramell et al., 2018; Xia & Wishart, 2010). The fobitools framework provides a couple of functions for performing ORA and MSEA analyses using the metabolite and food sets defined in FOBI, thus the result of these analyses consists of the enriched foods or metabolite sets by a given list of metabolites. On the one hand, to perform the proposed ORA, users need to define a "metabolite universe" (comprised of all metabolites analysed in the study) and a list of selected metabolites from that universe. On the other hand, to perform the MSEA proposed

method, users need to provide a metabolite ranked list containing all metabolites analysed in the study with their metabolite-level statistics sorted in decreasing order (see the fobitools use case). Furthermore, for the two provided methods users can select one of the two FOBI subontologies (Foods or Biomarkers) to be used in the analysis. If the Food subontology is selected, a food enrichment analysis providing enriched food groups will be carried out. Otherwise, if the Biomarker subontology is selected, a conventional enrichment analysis using FOBI's chemical classes will be performed, providing enriched chemical categories.

- **Automatic dietary text annotation**

Often, in nutritional studies, the manual annotation of dietary data collected with self-reporting questionnaires (e.g., 24h DR) is a hard and tedious process. The fobitools framework provides a function for the fast automatic annotation of free dietary text data. This function is composed of five sequential layers that use different text mining strategies as well as regular expressions and semantic similarity techniques. This algorithm allows users to obtain the FOBI names and identifiers of the entities in FOBI's Food sub-ontology that match the free text provided by users. See an example of text annotation with fobitools package in the Table 3.2. The two text strings used in this example were: "*Yesterday I ate a delicious apple pie and a coffee*" and "*pizza without meat*", respectively.

- **Network visualization**

FOBI, like all ontologies, is a knowledge graph, which means, graph-structured information. This feature allows users to visualize specific user-defined parts of FOBI in the form of a graph. The resultant network plots are designed to clearly differentiate the nodes that belong to Food sub-ontology and those that belong to Biomarker sub-ontology, using different colors and shapes in the nodes. In addition, the different properties defined in FOBI, as well as *is_a*, *BiomarkerOf*, and *Contains*, are also indicated with different colors and are easily identifiable. Figure 3.3 shows the resultant FOBI network of plotting the annotated FOBI terms in Table 3.2. Note that property

is_a is colored blue while the property *Contains* is colored yellow.

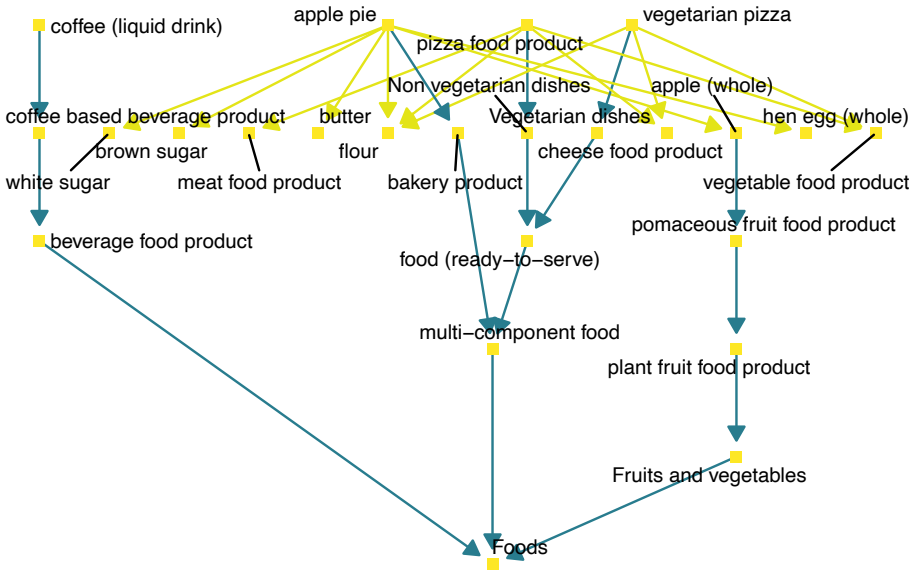


Figure 3.3: FOBI sub-network corresponding to the annotated terms in the Table 3.2. This network has been generated with the fobitools package.

- **FOBI parsing**

FOBI is available in OWL (Web Ontology Language) and OBO (Open Biomedical Ontologies) formats. These formats are common in ontologies and easily interpretable by computers. However, it can be difficult for end-users to query or obtain information from ontologies in these formats. This feature allows users to get information of specific terms of FOBI or obtain all FOBI information in a human-readable table format by parsing FOBI's OBO format into a structured table.

- **Identifier conversion**

The almost 600 FOBI metabolites contain their associated identifiers for different databases, specifically, each FOBI metabolite contains its chemical name, FOBI, HMDB, KEGG, PubChem, InChIKey, InChICode, and ChemSpider identifier (if any). It is common to change the identifiers from one type to another during the analysis, depending on the

Table 3.2: Example of dietary free-text annotation with fobitools package.

FOOD_ID	FOOD_NAME	FOBI_ID	FOBI_NAME
101	Yesterday I ate a delicious apple pie and a coffee	FOODON:00002473	apple (whole)
101	Yesterday I ate a delicious apple pie and a coffee	FOODON:00002475	apple pie
101	Yesterday I ate a delicious apple pie and a coffee	FOODON:03301036	coffee (liquid drink)
101	Yesterday I ate a delicious apple pie and a coffee	FOODON:00001139	coffee based beverage product
102	pizza without meat	FOODON:03310775	pizza food product
102	pizza without meat	FOBI:007956	vegetarian pizza

database or ontology used. This feature allows users to easily switch the identifiers of all compounds contained in FOBI among all the formats mentioned above.

3.1.3.4 Conclusion

The fobitools framework consists of an R/Bioconductor package and a web-based application with the clear aim of facilitating and extending the use of FOBI ontology to the scientific community, focusing mainly on users with limited programming experience. These two user-friendly tools allow, among others, the visual exploration of FOBI using dynamic and static network plots, the automatic annotation of dietary text data through text mining algorithms, and the performance of food enrichment analysis. All features provided in the fobitools framework can be executed via a graphical user interface and command-line R scripts.

3.2 Application of developed tools

3.2.1 Paper 4: Assessing adherence to healthy dietary habits through the urinary food metabolome

Pol Castellano-Escuder, Raúl González-Domínguez, Marie-France Vaillant, Patricia Casas-Agustench, Nicole Hidalgo-Liberona, Núria Estanyol-Torres, Thomas Wilson, Manfred Beckmann, Amanda J Lloyd, Marion Oberli, Christophe Moinard, Christophe Pison, Jean-Christian Borel, Marie Joyeux-Faure, Mariette Sicard, Svetlana Artemova, Hugo Terrisse, Paul Dancer, John Draper, Alex Sánchez-Pla, Cristina Andres-Lacueva. *Assessing adherence to healthy dietary habits through the urinary food metabolome: results from a European two-centre study. Submitted.*

3.2.1.1 Background

Diet is one of the most important modifiable lifestyle factors in human health and in chronic disease prevention. Thus, accurate dietary assessment is essential to reliably evaluate the adherence to healthy habits. In order to identify urinary metabolites that could serve as robust biomarkers of the diet quality, we studied a population-based cohort with repeated urine sampling and dietary assessment at baseline, six and twelve months over a year. Urine samples were subjected to large-scale metabolomics analysis for comprehensive quantitative characterization of the food-related metabolome. Then, regularized regression analysis was applied to identify those metabolites robustly associated with the AHEI-2010, and to investigate the reproducibility of these associations over time.

3.2.1.2 Aim

The aim of this study was to identify urinary metabolites that could serve as robust biomarkers of the diet quality, assessed through the

Alternative Healthy Eating Index 2010.

3.2.1.3 Study design

This work is part of the EIT Health project Cook2Health (Figure 3.4). The Cook2Health project is a randomized iterventional study of 160 participants from two different countries (United Kingdom and France) where the half of these participants were provided with a cooking device and nutritional coaching. Urine metabolites of all participants were measured at three different times of the study (baseline, six and twelve months). At the same time points, FFQs and the AHEI-2010 were calculated for each participant.

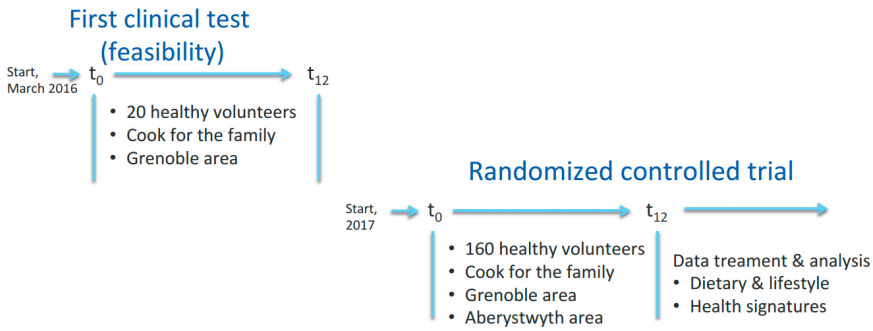


Figure 3.4: Cook2Health project study design.

In this study, data from the Cook2Health project were used to answer the question of the study, which was to explore associations between AHEI-2010 and urine metabolites. However, in this study, intervention was not considered as the main variable of the experiment, although its effect was corrected in the analyses.

3.2.1.4 Results

The most remarkable result was the positive association of numerous polyphenol microbial metabolites with the AHEI-2010 score, urinary enterolactone glucuronide showed reproducible association at the three study time points. Furthermore, strong associations were

found between the AHEI-2010 and various metabolites related to the intake of coffee, red meat and fish, whereas other polyphenol phase II metabolites were associated with higher AHEI-2010 scores at one of the three time points investigated. Therefore, we have demonstrated that urinary metabolites, and particularly microbiota-derived metabolites, could serve as reliable indicators of the adherence to healthy dietary habits.

3.2.1.5 Conclusion

This study has demonstrated that the urinary food-related metabolome is strongly associated with the adherence to healthy dietary habits as assessed through the AHEI-2010. Many of the metabolites identified were microbial-derived compounds, thus supporting a major role of the gut microbiota in the interplay between diet and health. Despite the high variability across the three study time points for these compounds, enterolactone glucuronide showed reproducible association over the one-year follow-up. Furthermore, robust associations were found between the AHEI-2010 score and various metabolites reflecting the intake of coffee, red meat and fish, whereas other food products showed robust association at one of the three time points here investigated.

3.2.2 Paper 5: The food-related serum metabolome associates with later cognitive decline in older subjects

Raúl González-Domínguez, **Pol Castellano-Escuder**, Francisco Carmona, Sophie Lefèvre-Arbogast, Dorrain Y. Low, Andrea Du Preez, Silvie R. Ruigrok, Claudine Manach, Mireia Urpi-Sarda, Aniko Korosi, Paul J. Lucassen, Ludwig Aigner, Mercè Pallàs, Sandrine Thuret, Cécilia Samieri, Alex Sánchez-Pla, Cristina Andres-Lacueva. *The food-related serum metabolome associates with later cognitive decline in older subjects: A twelve-year prospective observational study. Submitted.*

3.2.2.1 Background

Nowadays, diet is considered an important modulator of cognitive decline and dementia, but the available evidence is, however, still fragmented and often inconsistent. To decipher a role for diet in the early onset of cognitive decline, we here studied the long-term prospective Three-City Cohort, that consists of two separate, nested case-control sample sets from different geographic regions (Bordeaux and Dijon). The food-related and microbiota-derived circulating metabolome was studied in participants free of dementia at baseline, by subjecting serum samples to large-scale, quantitative metabolomics analysis.

3.2.2.2 Aim

The main aim of this study was to determine if the early onset of cognitive decline associates with food-related serum metabolome. Then, the secondary aim of this study was to determine which specific metabolites (from the food metabolome) associate with cognitive decline.

3.2.2.3 Study design

This work is part of the D-CogPlast project. The D-CogPlast project is an observational study that involves nested case-control samples built among participants from three French cities (Bordeaux, Dijon and Montpellier). This study is a population-based cohort on dementia that includes older persons (>65 years) (Table 3.3). Sociodemographic and lifestyle characteristics, medical information, neuropsychological testing, blood pressure, anthropometric measurements and fasting serum samples were collected at baseline, and follow-up visits were then scheduled every two-three years for neuropsychological assessment.

Table 3.3: Clinical and demographic characteristics of the discovery and validation case-control samples of the D-CogPlast study.

	Bordeaux samples		**Dijon samples**	
	Cases (N=209)	**Controls (N=209)**	**Cases (N=212)**	**Controls (N=212)**
Age (years)	75.9 ± 4.5	75.7 ± 4.2	76.5 ± 5.2	76.1 ± 4.7
Gender (male/female)	71/138	71/138	78/134	78/134
Education level, ≥ secondary school (%)	28.7	28.7	28.3	28.3
BMI (kg m-2)	26.8 ± 4.4	26.1 ± 3.6	25.8 ± 4.6	25.1 ± 3.6
Number of medications regularly consumed	4.9 ± 2.7	4.1 ± 2.4	5.5 ± 3.0	4.0 ± 3.0
ApoE-ϵ4 (%)	25.8	12.0	26.9	20.8
Diabetes (%)	12.9	5.7	12.7	5.7
History of cardiovascular diseases (%)	33.5	27.8	41.0	30.2

From the entire Bordeaux and Dijon cohorts, eligible participants were selected for the presented study if they were not diagnosed with dementia at baseline, had available serum samples, and had at least one repeated cognitive evaluation over the subsequent 12 years. To build the case-control samples on cognitive decline, a composite score of global cognition was defined at each follow-up visit as the average of five neuropsychological tests, defining cases as the participants with the worst results on these tests. Then, each case was matched to a control with the same age, gender and education level. Serum metabolites were compared between cases and controls in order to find novel associations between serum metabolome and cognitive decline.

3.2.2.4 Results

The results revealed a protective association between metabolites derived from cocoa, coffee, mushrooms, red wine, the microbial metabolism of polyphenol-rich foods, and cognitive decline. A harmful association was found with metabolites related to unhealthy dietary components, such as artificial sweeteners, alcohol and food additives. Furthermore, we found associations indicating that perturbations in the microbiota-related metabolism of aromatic amino acids and of fatty acid β-oxidation might be involved in cognitive decline. Although the specific metabolite signatures were different between the two study sample sets, due to inter-individual variability factors, a substantial part of these findings was consistent across the two samples, suggesting robust support for such an association.

3.2.2.5 Conclusion

This study suggests that food-related and microbiota-derived metabolites may play an important role in the later development of cognitive decline. These results support a protective association between metabolites reflecting the consumption of polyphenol-rich foods (e.g., fruits and vegetables), cocoa, coffee, mushrooms and red wine with cognitive decline, whereas other food components related to unhealthy dietary components (e.g., alcohol, artificial sweeteners) may have deleterious effects on cognition.

3.3 Software

3.3.1 R/Bioconductor packages

Two Bioconductor (Gentleman et al., 2004) packages were developed in the context of this thesis: the POMA and fobitools packages.

All software described here was developed using the R programming language (R Core Team, 2019) and following the best practices for R

package development described in Hadley Wickham's book *R packages* (Wickham, 2015). In addition, all packages described here were written following the tidyverse (Wickham et al., 2019) philosophy, in order to keep all code clean and readable, facilitating the contribution of other users and the software maintenance. These two packages are tested with a well defined testthat architecture (Wickham, 2011) on a continuous integration system using GitHub Actions, covering tests on Linux, Mac and Windows OS with current R versions.

3.3.1.1 POMA

This package introduces a structured, reproducible and easy-to-use workflow for the visualization, preprocessing, EDA, and statistical analysis of metabolomics data, enabling a flexible data cleaning and statistical analysis processes in one comprehensible and user-friendly R/Bioconductor package.

POMA uses *MSnSet* S4 class objects defined in the MSnbase package (Gatto & Lilley, 2012) which inherits from class *eSet* of the Biobase package (Huber et al., 2015) and is fully integrated in the Bioconductor environment. POMA is available at `https://bioconductor.org`.

Available documents:

- POMA manual
- POMA vignette: POMA Workflow (see the POMA use case)
- POMA vignette: POMA Normalization
- POMA vignette: POMA EDA Example
- POMA website: `https://pcastellanoescuder.github.io/POMA/`
- POMA GitHub repository: `https://github.com/pcastellanoescuder/POMA`

3.3.1.2 fobitools

This package provides a set of functions for interacting with FOBI (Castellano-Escuder, González-Domínguez, Wishart, Andrés-Lacueva, & Sánchez-Pla, 2020). This package is focused on the novel concept of food enrichment analysis in nutrimetabolomic studies. However, other

useful features such as the network interactive visualization of FOBI and the automatic annotation of dietary free-text data are also provided The fobitools package is available at `https://bioconductor.org`..

Available documents:

- fobitools manual
- fobitools vignette: Simple food ORA
- fobitools vignette: Use case ST000291 (see the fobitools use case)
- fobitools vignette: Use case ST000629
- fobitools vignette: Dietary text annotation
- fobitools website: `https://pcastellanoescuder.github.io/fobitools/`
- fobitools GitHub repository: `https://github.com/pcastellanoescuder/fobitools`

3.3.2 Graphical User Interfaces

Three graphical user interfaces (GUIs) were developed in the context of this thesis. The first two GUIs POMAShiny and fobitoolsGUI are based on the functions of the two R packages described in the above section, while the third GUI is based on two preexisting R/Bioconductor packages named msmsEDA (Gregori et al., 2020a) and msmsTests (Gregori et al., 2020b).

All these three GUIs have been developed in R and using the Shiny framework (Chang, Cheng, Allaire, Xie, & McPherson, 2020).

3.3.2.1 POMAShiny

POMAShiny is a web-based tool that provides a structured, flexible and user-friendly workflow for preprocessing, exploration, and statistical analysis of metabolomics data. This tool is based on the POMA R/Bioconductor package, which increases the reproducibility and flexibility of the analysis outside the web environment. POMAShiny's workflow is structured in four sequential and well-defined panels: 1) data upload, 2) preprocessing, 3) EDA and 4) statistical analysis panels.

Available documents:

- POMAShiny URL: `https://webapps.nutrimetabolomics.com/POMAShiny`
- POMAShiny GitHub repository: `https://github.com/pcastellanoescuder/POMAShiny`
- POMAShiny Docker image: `https://hub.docker.com/repository/docker/pcastellanoescuder/pomashiny`

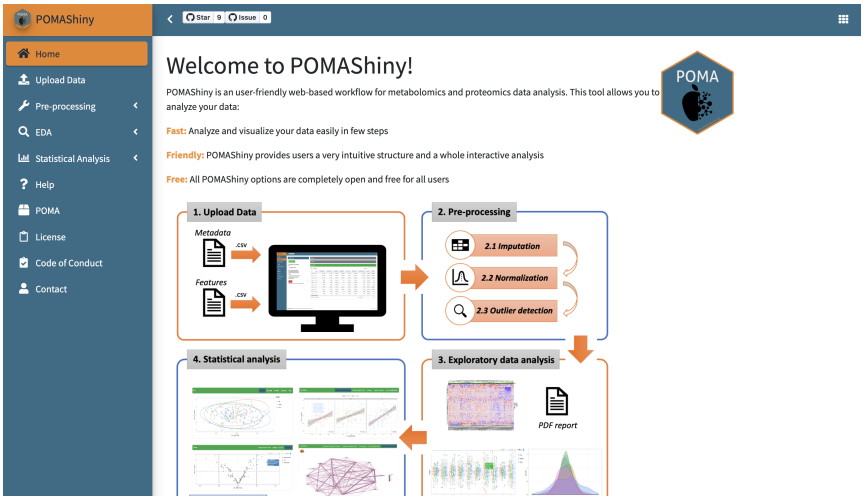


Figure 3.5: Screenshot of the POMAShiny *Home* page.

3.3.2.2 fobitoolsGUI

The fobitoolsGUI is a web-based tool based on the fobitools R package. This user-friendly web interface provides a set of tools for interacting with FOBI. A collection of basic manipulation tools for biological significance analysis, graph visualization and text mining strategies for annotating nutritional data are provided here:

- FOBI graph static visualization
- FOBI graph dynamic visualization
- Extract FOBI information in a downloadable table
- Compound ID conversion (among metabolite names, FOBI, ChemSpider, KEGG, PubChemCID, InChIKey, InChICode and HMDB IDs)
- Biological significance analysis using ORA and MSEA methods:

  - Chemical class enrichment analysis: ORA and MSEA using FOBI chemical classes as metabolite sets
  - Food enrichment analysis: ORA and MSEA using FOBI food groups as metabolite sets
- Text mining algorithm for annotating free-text dietary data

Available documents:

- fobitoolsGUI URL: `https://webapps.nutrimetabolomics.com/fobitoolsGUI`
- fobitoolsGUI GitHub repository: `https://github.com/pcastellanoescuder/fobitoolsGUI`

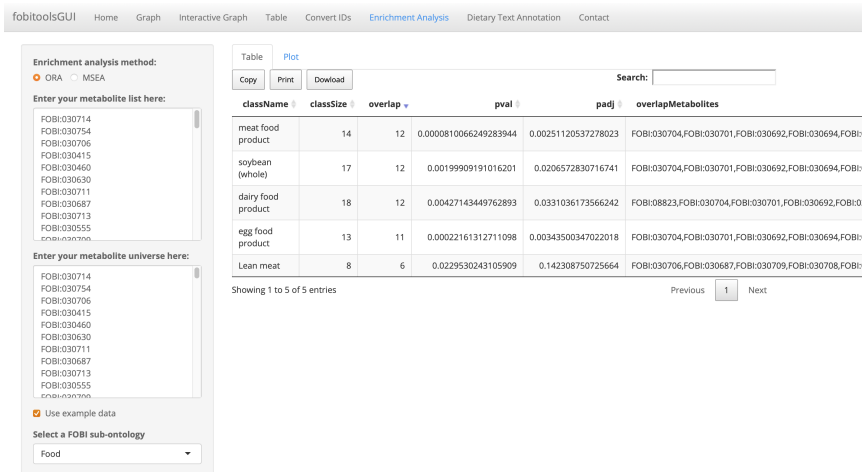


Figure 3.6: Screenshot of the fobitoolsGUI *Enrichment Analysis* page.

3.3.2.3 POMAcounts

POMAcounts is a web-based tool for EDA and statistical analysis of mass spectrometry spectral counts data. This GUI is based on the R/Bioconductor packages msmsEDA (Gregori et al., 2020a) and msmsTests (Gregori et al., 2020b). The name of POMAcounts is given by the large similarity of both the *frontend* and the *backend* that it shares with POMAShiny.

Available documents:

- POMAcounts URL: `http://uebshiny.vhir.org:3838/POMAcounts`
- POMAcounts GitHub repository: `https://github.com/pcastellanoescuder/POMAcounts`

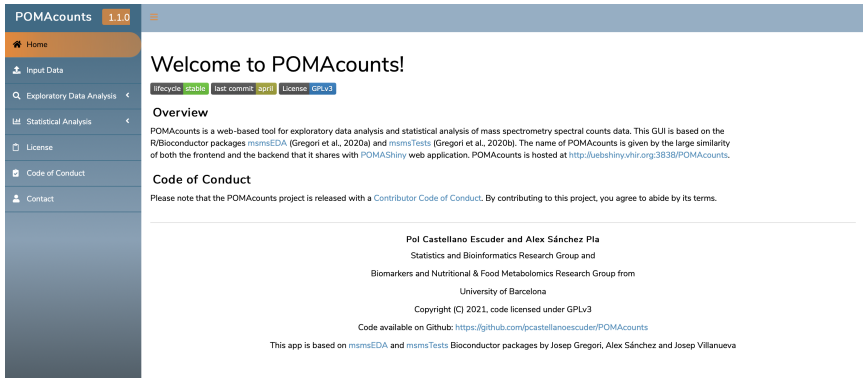


Figure 3.7: Screenshot of the POMAcounts *Home* page.

Chapter 4

Discussion

While a summary of each publication is included in the results section and the whole discussion of each publication is included in the appendices, this chapter aims to provide an overview of this thesis through a general discussion that encompasses all the individual results presented above.

As explained in the introduction, nutrimetabolomics is the integration of the fields of nutrition and metabolomics, becoming one of the most promising ways to improve nutritional assessment and dietary treatments in the future (Ulaszewska et al., 2019).

However, nutrimetabolomics also combine the intrinsic complexity of each of the two fields it integrates, dealing with challenges that make it difficult to analyse and interpret its results. For example, low reproducibility of urine metabolomics analyses using LC-MS, bias in nutritional data obtained through dietary questionnaires, or large interindividual variability of subjects in observational and interventional studies. In addition, many of the relationships between diet-derived metabolites and foods are not fully known, leading to discussion, and making it difficult not only to identify potential dietary biomarkers or to develop predictive models, but to study basic relationship between the fields of nutrition and metabolomics.

For this reason, all the resources and tools developed in the context of this thesis arise from this complexity of nutrimetabolomics data, proposing alternatives focused on improving their integration,

statistical analysis, and biological interpretation.

A detailed knowledge of the complex interrelationships between foods, food components, and food intake biomarkers is critical for understanding nutrition and metabolism. This understanding will allow to obtain much more accurate and objective dietary assessments based on metabolomics in the future. In turn, it will also help develop personalized nutritional strategies to design specific diets based on the patient's phenotype, disease status, microbiome, or metabolome, among others.

In view of this scenario, the first main objective of this thesis was focused on studying, characterizing, and defining the relationships between metabolites and diet in a clear and robust way, providing the scientific community with a consensus starting point when designing nutrimetabolomics studies, interpreting their results or establishing comparisons and meta-analyses between different studies in this field (**specific objective 1**).

Currently, different complete and useful databases provide information on those metabolites associated with certain foods, such as the Exposome-Explorer (Neveu et al., 2016), Phenol-Explorer (Rothwell et al., 2013), PhytoHub (`http://phytohub.eu/`) and Food Database (FooDB) (`http://foodb.ca/`). However, these resources describe this information in a very accurate but also heterogeneous way, making it difficult to compare studies and sometimes providing slightly different information for the same compounds or foods. Furthermore, despite the computational advantages it entails (see the "Ontologies" section), there is currently no resource in the form of an ontology to define these associations.

For this reason, we have developed the FOBI ontology (**specific objective 1**). FOBI is a specific ontology for the field of nutrimetabolomics composed of two subontologies with independent but interrelated hierarchical information; the Food subontology and Biomarker subontology. Food subontology consists of different raw and complex (multi-component) foods grouped according to their nutritional groups, while Biomarker subontology contains dietary metabolites classified according to their chemical classes. In this way,

the common and standardized language used to define and relate the elements of each subontology allows users to interpret the relationships between foods and biomarkers at both user and computational levels. Thus, FOBI can be used for different purposes, from making simple queries to complex computational queries using all the information stored in the ontology.

Ontologies facilitate many practical applications in the field of bioinformatics, such as annotating entities, performing different enrichment analyses, performing semantic similarity analyses, or even discovering new relationships between entities (Hoehndorf et al., 2015). In the case of FOBI ontology, the clearest application is the annotation of foods and dietary biomarkers, facilitating the comparability and interoperability between studies and projects in the field of nutrimetabolomics. Since FOBI provides detailed definitions of the associations between different types of foods and their associated metabolites, it becomes a significant improvement for nutrimetabolomics research.

FOBI information can also serve to facilitate study designs, from hypothesis generation (e.g., expected metabolites that occur after a dietary intervention) to experimental design (e.g., optimization of targeted metabolomics methods focused on metabolites of interest). In addition, one of the main applications of FOBI is the ability to perform biological significance analyses in nutrimetabolomic studies, which until now was not possible due to the lack of specific methodologies.

The FOBI ontology consists of **paper 1**, presented in the results section.

As briefly explained in the introduction, one of the main reasons that hinders the use of ontologies is the required programming knowledge to interact with them, extract information, and use them in general. FOBI, like any ontology, also has this limitation. For this reason, once FOBI was developed, we considered to offer users of the nutrimetabolomics community a resource for using FOBI in an easy and fast way. This led to **specific objective 1a** and **specific objective 1b**.

In order to offer users the ability to carry out some of the FOBI

ontology applications mentioned above, we developed the fobitools framework, specifically focused on the new concept of food enrichment analysis.

The fobitools framework consists of a Bioconductor package (**specific objective 1a**) and a web application (**specific objective 1b**) that aim to facilitate and extend the use of the FOBI ontology to the scientific community. These two tools were developed following the "*user-friendly*" philosophy and allow, among others, the performance of different food enrichment analyses, the exploration of FOBI using static and dynamic network plots and automatic annotation of free dietary text data using text mining algorithms.

This is an open-source project, so the scientific community can easily use and contribute to it. Thus, fobitools framework allows researchers to use the FOBI ontology in a quick and easy way, either from the R command line or from the web application, where users do not need to have programming notions.

This tool introduces the concept of food enrichment analysis for the first time, allowing users to explore enriched foods or food groups based on lists of metabolites obtained in nutrimetabolomics studies.

While this tool can be a substantial improvement for the interpretation of results in nutrimetabolomics studies, some limitations should also be noted. Currently, since the FOBI ontology is in its first release version, the analysis with the fobitools framework may be limited to a small number of foods and metabolites in comparison to other ontologies and databases. Thus, future efforts will be aimed at expanding the FOBI ontology, leading to an increase in the number of metabolites, foods, and metabolite-food relationships. On the other hand, the fobitools framework provides the methodology for interacting and using the FOBI ontology regardless of the amount of information it contains. Therefore, future improvements on FOBI will have a direct impact on the fobitools framework, increasing its usefulness and allowing more accurate, complete, and robust analyses. Regarding the future software enhancements of fobitools, these will be primarily intended at implementing new enrichment analysis methods.

The fobitools framework consists of **paper 3**, presented in the results section. Unfortunately, although this tool is already in use, it has not been applied in publications 4 and 5 because this tool and these papers were carried out simultaneously.

The second major goal of this thesis was the development of methods and tools for contributing to the improvement of the data analysis process in metabolomics (and therefore nutrimetabolomics) studies to help improve the biomarker discovery process, among others (**specific objective 2**).

Often one of the main applications of metabolomics is the characterization of new therapeutic targets in the fields of human health and personalized medicine (Wishart, 2016). For this reason, over the last decade several tools have emerged for contributing to the analysis of such complex data (Stanstrup et al., 2019). However, many of them still limit the analysis to a small number of statistical methods (Gardinassi, Xia, Safo, & Li, 2017), which forces researchers to use an extensive battery of different tools to meet all the needs of the analysis.

In order to contribute to the extension of methods and tools available for the analysis of metabolomics data, the web tool POMAShiny developed in the context of this thesis, provides a complete and structured workflow that covers most of the data analysis processes, including preprocessing, exploration and statistical analysis, with the intention of being a complementary, easy to use and intuitive tool that addresses some of the problems that are not covered by other tools (**specific objective 2a**). This workflow is integrated into an attractive graphical user interface that provides several methods for data analysis, including univariate statistical methods, multivariate and dimension reduction methods, feature selection methods, regularized regression approaches, machine learning classification algorithms, prediction model strategies, and various high-quality interactive visualization options.

This new tool is based on the POMA Bioconductor package (**specific objective 2**) and integrates many of the most widely used methods for metabolomics data analysis (Gardinassi et al., 2017; Stanstrup et al., 2019), as well as incorporating new useful and powerful

alternatives. POMAShiny allows users to perform an integrated data analysis in an interactive, intuitive, and well-documented web environment, making the data analysis process more accessible to a wide range of researchers.

The joint existence of both the POMA package (fully integrated within the Bioconductor environment) and the POMAShiny web interface means a huge increase in the reproducibility of the tool, also contributing to the reusability of existing methods in the R and Bioconductor environments (Gentleman et al., 2004; R Core Team, 2019), in addition to allowing the easy extension, integration, and interoperability with other workflows, such as the RforMassSpectrometry initiative (RforMassSpectrometry.org), which provides the data structures used in the POMA package. Therefore, users can perform spectra data processing and other routine MS workflow operations using the RforMassSpectrometry initiative packages and then easily migrate to POMA/POMAShiny to perform the statistical analyses.

The POMAShiny web aplication consists of **paper 2**, presented in the results section.

With the achievement of specific objectives 1 and 2, this thesis presents a set of tools and resources that allow researchers to conduct statistical analyses of metabolomics studies using a wide variety of methods, as well as providing the results of these analyses with their biological significance in a nutrimetabolomics context. In addition, graphical user interfaces are provided for all the tools presented in this work, facilitating their use regardless of users' programming knowledge.

Regarding the **specific objective 3**, this thesis also presents two nutrimetabolomic studies where several data analysis concepts have been applied as well as different statistical methodologies provided by the tools and resources discussed previously.

In the first of these studies, the goal was to identify metabolites or groups of metabolites associated with the AHEI-2010 score (Chiuve et

al., 2012) (**specific objective 3a**).

Although numerous studies have previously investigated the association between circulating metabolites and consumption of certain food groups, only a few of them have focused on the identification of biomarkers of healthy dietary patterns. In this regard, several studies have recently addressed the identification of potential metabolomic markers of AHEI-2010 in serum and plasma samples from different populations (Akbaraly et al., 2018; Bagheri et al., 2020; McCullough et al., 2019; Walker et al., 2020).

In this work, we aimed to identify urinary metabolites associated with the AHEI-2010, which could serve as biomarkers of adherence to healthy dietary patterns.

In this study, the POMA tool (**paper 2**) was used to perform all preprocessing and exploratory analysis processes, and part of the statistical analysis.

We used both LASSO and *limma* approaches to identify those urinary metabolites associated with healthy and unhealthy dietary habits. Urinary enterolactone glucuronide levels were positively associated with the AHEI-2010 score using both methodologies. This association was consistently replicated at the three study time points investigated.

Enterolactone is the main microbial-derived metabolite of dietary lignans, a subclass of polyphenols widely distributed in plant foods such as fruits, vegetables, wholegrains, legumes and nuts (Senizza et al., 2020). Lignans are known to have different anti-inflammatory and antioxidant properties; several epidemiological studies have shown that high circulating concentrations of enterolactones are associated with a lower risk of cardiovascular disease (Rienks, Barbaresko, & Nöthlings, 2017), different cancers (Micek et al., 2021), and neurodegenerative disorders (Reddy et al., 2020), among others. Therefore, this metabolite could be considered as a reliable and robust biomarker to assess adherence to healthy dietary patterns.

The LASSO regression also identified a reproducible association of

the metabolite 5-(hydroxymethyl-2-furoyl)glycine with high AHEI-2010 scores over time, as well as a positive association with 2-furoylglycine in one of the three time points studied. Furan metabolites have previously been proposed as biomarkers of various heat-processed food products, such as nuts (Prior, Wu, & Gu, 2006) and coffee (Heinzmann, Holmes, Kochhar, Nicholson, & Schmitt-Kopplin, 2015).

Additionally, although it was not corroborated in the three time periods studied, a strong negative association was also found between L-carnitine and AHEI-2010, along with a negative association of carnosine, reflecting a harmful effect of the consumption of red and processed meat on health. Moreover, a negative association of tobacco-derived metabolites with AHEI-2010 was also identified. In contrast, several metabolites reflecting fish and shellfish intake showed positive associations with AHEI-2010. In addition, other consistent associations between the AHEI-2010 score and other candidate biomarkers for food intake defined in the FOBI ontology (Castellano-Escuder et al., 2020) (**paper 1**) were also identified, but only in one of the three time points investigated. A positive association was found between AHEI-2010 and several metabolites of red wine (e.g., resveratrol), citrus, olive oil, and berries.

In conclusion, these results show that several diet-related metabolites are strongly associated with adherence to healthy dietary habits assessed with the AHEI-2010 (**paper 4**).

Finally, the second application study presented in this thesis aimed to identify metabolites or groups of metabolites associated with disease risk or health status (**specific objective 3b**).

The association of modifiable lifestyle factors with the pathogenesis of cognitive decline (CD) and dementia is well accepted today (Peters et al., 2019). Diet has been identified as a key factor in maintaining proper brain function (Flanagan et al., 2020). In fact, many components of the diet can modulate the molecular mechanisms that contribute to CD, including oxidative stress, neuroinflammation, and vascular dysfunction (Vauzour et al., 2017).

The aim of this study was to decipher the role of diet in the

development of cognitive decline through a large-scale targeted metabolomics approach.

In this study, the POMA tool (**paper 2**) was also used to carry out all preprocessing and exploratory analysis processes, including the imputation of missing values, data normalization and treatment of outliers (or atypical samples).

In order to identify serum metabolites associated with CD, we used a conditional logistic LASSO regression combined with the *bootstrap* technique, to stabilize the results and the intrinsic variability of the LASSO method.

Many of the metabolites identified in the two study populations (see the "Paper 5" section), including polyphenol derivatives and aromatic amino acids, suggest a close interaction between diet, microbiota, and CD. The gut microbiota has been recognized as an important factor in health and cognition, as many microbial-derived metabolites have essential metabolic and signalling properties that can modulate brain function (Needham, Kaddurah-Daouk, & Mazmanian, 2020; Parker, Fonseca, & Carding, 2020). Therefore, it has been hypothesized that the gut microbiota and the molecules it produces could be part of a network that links diet to cognitive function through the "gut-brain axis" (Collins, Surette, & Bercik, 2012).

In both study populations, an inverse association was observed between various phenolic acids and other plant-derived metabolites with the risk of CD, providing more evidence on the protective effect of eating polyphenol-rich foods (i.e., fruits and vegetables) against CD (Mottaghi, Amirabdollahian, & Haghighatdoost, 2018).

In line with the previous study of the D-CogPlast project, performed using an untargeted metabolomic approach (Low et al., 2019), we also observed a negative association between 3-methylxanthine (a metabolite derived from theobromine present in cocoa) and CD. The circulating levels of 3-methylxanthine were highly correlated with theobromine, which was also negatively associated with CD in both populations, reinforcing the protective effect of cocoa consumption against CD (Moreira, Diógenes, Mendonca, Lunet, & Barros, 2016). In addition,

the results of this work also suggest that coffee intake is negatively associated with the risk of CD (protective effect). As explained above, 2-furoylglycine is a biomarker of coffee consumption, which was found to be associated with a lower risk of CD in both study populations.

In addition to these potentially protective associations between polyphenol-rich foods, cocoa, coffee, red wine, and CD, these results also point to a harmful association of certain dietary components on cognitive function, including, for example, artificial sweeteners.

In conclusion, these results suggest a protective association mainly of microbiota-derived metabolites, fruits and vegetables, and coffee with cognitive decline, while other metabolites related to unhealthy dietary habits, such as sugar-sweetened drinks, can have detrimental effects on cognition (**paper 5**).

Part III

Conclusions

This thesis has studied the essential aspects of dietary biomarker discovery by metabolomics and the bases of biological significance analyses in nutrimetabolomic studies. The main contributions to these areas are highlighted below. Specifically, it has been shown that:

- FOBI is the first ontology that integrates nutritional and metabolomics data using a standardized common language to define the relationships between foods and their associated metabolites. At the moment, FOBI has a total of 1197 well-defined terms, 11 chemical top-level classes, 13 food top-level classes and 4 different properties with clear relationship mappings.

- FOBI allows experts to annotate and analyze nutritional and metabolomics data in a consistent way, making the results comparable between and across studies of these fields. The development of FOBI will lead to an improvement in the interoperability of nutritional and nutrimetabolomics data, thereby making the data sets generated in these studies fully FAIR compliant.

- POMAShiny is a user-friendly web-based tool that provides an integrated metabolomics data analysis workflow with a wide range of possibilities, both for data preprocessing and statistical analysis, including outlier detection methods, flexible exploratory data analysis operations, downloadable reports and several statistical methods from simpler approaches such as univariate analyses to more complex methods such as regularized regression and machine learning algorithms.

- The fobitools framework consists of an R/Bioconductor package and a web-based application that provide an infrastructure to interact with FOBI ontology in a highly user-friendly way. This framework allows researchers to perform enrichment analyses in nutrimetabolomic studies, among other useful operations, such as the network interactive visualization of FOBI and the automatic annotation of dietary free-text data using the FOBI information.

- The urinary food metabolome is strongly associated with adherence to healthy dietary habits as assessed through the AHEI-2010. Many of the associated metabolites discovered were microbial-derived compounds, including enterolignans, urolithins and phenolic acids, thus supporting a major role of the gut microbiota in the interplay between diet and health.

- Food-related and microbiota-derived metabolites may play an important role in the later development of cognitive decline. Those metabolites reflecting the consumption of polyphenol-rich foods, cocoa, coffee, mushrooms and red wine showed a protective effect on cognitive decline, whereas other dietary metabolites related to unhealthy dietary components showed deleterious effects on cognition.

The developed tools have been implemented in two R/Bioconductor packages -POMA and fobitools- freely available at Bioconductor, and their dissemination has been facilitated by two graphical user interfaces -POMAShiny and fobitoolsGUI- publicly available at GitHub.

Part IV

Resum en català

Agraïments

Estic absolutament convençut que, per molt gran que sigui l'esforç, per moltes hores de dedicació, i fins i tot per molt altes que siguin les capacitats individuals; tots els grans èxits de la vida sempre depenen en gran mesura de les condicions, del context i sobretot de les persones que ens envolten.

Aquesta tesi és un exemple on aquesta idea es fa més que evident i només puc que reforçar-la. És evident que hi ha moltes hores de treball i sacrifici darrere d'aquesta tesi, però, tot i així, crec que res d'això hauria estat possible sense totes les persones que m'han fet costat, no només científicament, sinó de totes les formes possibles. Per aquest motiu, és un honor per mi adreçar aquestes paraules d'agraïment a tothom que ha contribuït d'alguna manera a aquesta fita personal.

En primer lloc, vull expressar el meu agraïment als meus directors de tesi, l'**Alex Sánchez Pla** i la **Cristina Andrés Lacueva**. Els hi estic molt agraït per haver cregut en mi i per haver-me donat l'oportunitat de realitzar una tesi doctoral, fent realitat el meu somni. Gràcies per assessorar-me, guiar-me, corregir el meu treball científic i fer d'aquesta tesi un enorme procés d'aprenentatge que ha superat amb escreix les meves expectatives.

Voldria extendre aquest agraïment als professors, companys i amics del grup d'Estadística i Bioinformàtica, per tot el que m'han ensenyat, que és bona part dels coneixements estadístics que tinc a dia d'avui. Gràcies **Francesc Carmona**, **Esteban Vegas**, **Ferran Reverter**, **Antonio Miñarro** i **Marta Cubedo**.

També vull expressar el meu agraïment als meus companys i amics

del grup de Biomarcadors i Metabolòmica Nutricional i Alimentària, on he passat la major part del temps i amb qui hem compartit infinitat de moments, des de reunions de laboratori fins a viatges, sopars i cerveses. Sense ells hauria estat impossible realitzar aquest treball, i estic agraït de poder dir que he acabat aquesta tesi amb molt més que companys de laboratori. Vull extendre aquest agraïment especialment a la **Magalí Palau**, per haver-me guiat en els primers passos de la tesi; a la **Maria Cristina Cadena**, pel seu somriure i la seva alegria que sempre ajudaven a crear un ambient immillorable al laboratori; i a la **Nicole Hidalgo**, que després de més de tres anys s'ha convertit en gairabé una germana i en un suport indispensable.

També és el meu desig personal donar les gràcies a tots els membres de GRBIO, dels qui he après molt i amb els qui he compartit moments fantàstics, des de seminaris de recerca fins als moments més pintorescos. Vull expressar el meu agraïment a la **Marta Bofill** i al **Guillermo Villacampa**, per ser-hi sempre, recolzant-nos mútuament, discutint infinitat de temes estadístics i compartint moments increïbles, convertir-se en grans amics, tant a dins com a fora de l'entorn de treball.

Gràcies a totes les persones que vaig conèixer a Aberystwyth. Gràcies al **Prof. John Draper** per haver-me acollit al seu laboratori i també gràcies al **Tom Wilson** i a la **Mandy Lloyd** per guiar-me durant la meva estada a l'estranger.

També voldria dedicar un agraïment especial a totes les persones que en el seu moment em van transmetre la seva passió per la ciència, em van assessorar i em van guiar en les decisions que han fet possible, en part, que estigui escrivint aquestes línies avui. El meu més sincer agraïment, **Josep Jiménez**, **Silvia Ribó** i **Judith Cebrià**.

M'agradaria transmetre aquest agraïment a tots els meus amics, tant de dins com de fora de la ciència, que m'han donat un suport indiscutible en la realització d'aquesta fita personal. Gràcies al **Marçal Yll** i a la **Núria Catasús**, amb qui he compartit ciència des que vam començar els nostres estudis de biologia i bioquímica als 18 anys. També agraeixo als meus amics **Enric Martí** i **Adrià Hernández**, que sempre han estat allà per donar-me suport i escoltar-me, especialment a la recta final d'aquest treball. També vull expressar el meu agraïment

al **Damià**, per tota la seva ajuda i per fer més fàcil el meu camí, especialment durant l'últim any de la tesi, que ha coincidit amb la pandèmia de la COVID-19.

Finalment, voldria donar el meu més profund agraïment a la meva família, als meus grans referents, als que sempre han estat al meu costat. Per a la meva mare, un exemple de perseverança, perfeccionisme i superació personal. Al meu pare, del qual no paro d'aprendre i que m'ha ensenyat a afrontar tots els reptes de la vida. I la meva germana, la **Carlota**, que sempre m'ha donat un suport incondicional en tots els meus reptes personals i que estic segur, aconseguirà tot el que es proposi a la vida. I finalment, però no menys important, el meu agraïment i amor a la **Montse**, la meva parella, que ha aguantat pacientment molts dies festius, caps de setmana i fins i tot vacances dedicades a aquest treball.

Gràcies també a tots els que no he esmentat, que han contribuït d'alguna manera a aquesta fita. Moltes gràcies.

Chapter 5

Introducció

El descobriment de les relacions entre nutrició i salut és un dels principals objectius de la nutrició moderna. Gràcies als grans avenços en el camp de la metabolòmica en els darrers anys, aquesta tècnica d'alt rendiment s'ha convertit en un aliat indispensable per a la recerca nutricional, essent la metabolòmica nutricional (o nutrimetabolòmica) una eina clau per explorar les relacions entre dieta i salut i per predir la ingesta d'aliments mitjançant perfils metabolòmics, entre d'altres.

Tanmateix, moltes de les relacions entre els metabòlits i els aliments encara no estan del tot clares i són objecte de discussió, cosa que requereix estudis més profunds en aquesta àrea. Aquesta tesi se centra en l'estudi exhaustiu d'aquestes relacions entre metabòlits i dieta per tal d'entendre millor la seva complexitat i contribuir a la millora i simplificació de l'anàlisi de dades de nutrimetabolòmica, així com a la millora de la interpretació biològica dels seus resultats.

Per assolir aquest objectiu, aquest treball proposa diferents eines bioinformàtiques dissenyades per als problemes derivats de l'anàlisi i integració d'aquestes dades. Diferents eines i recursos, com ara una ontologia que defineix les relacions entre metabòlits i aliments, una eina d'anàlisi estadística per a dades de metabolòmica i eines per a l'anàlisi de la significació biològica en estudis de nutrimetabolòmica es presenten a continuació.

5.1 Metabolòmica

En ciències de la vida, el sufix "*-omic*" fa referència a la caracterització i quantificació conjunta de grups de molècules biològiques que es tradueixen en l'estructura, la funció i la dinàmica d'un organisme. Per tant, les diferents disciplines d'estudi d'alt rendiment es classifiquen com a "*omiques*" diferents dins el món de la biologia, sent la genòmica, la transcriptòmica, la proteòmica i la metabolòmica les principals i més estudiades, respectivament.

5.1.1 Metaboloma

El metaboloma és el conjunt de totes les molècules de baix pes molecular (metabòlits) presents en un sistema biològic. És el resultat de les reaccions bioquímiques catalitzades per les proteïnes del proteoma i determina el fenotip final de l'organisme. El nombre dels diferents compostos del metaboloma varia segons l'organisme, però canvia constantment a causa de totes les reaccions químiques que es produeixen a l'organisme (Færgestad et al., 2009).

5.1.1.1 Metaboloma humà

Concretament, aquesta tesi se centra en el metaboloma humà, compost per tots els metabòlits que es troben en el cos humà. Al seu torn, tots aquests metabòlits que constitueixen el metaboloma humà poden provenir de quatre fonts ben definides descrites a continuació:

- **Metaboloma alimentari**

Durant els darrers 15 anys, s'han proposat diferents definicions per aquest tipus de metaboloma (Cevallos-Cevallos et al., 2009; Fardet et al., 2008; Wishart, 2008), però, el més acceptat avui és el següent:

"*El metaboloma alimentari es defineix com la part del metaboloma humà derivada directament de la digestió i la biotransformació dels aliments i els seus components*" (Scalbert et al., 2014).

Segons aquesta definició, el metaboloma alimentari no considera aquells compostos presents en la naturalesa dels aliments, sinó els compostos derivats de l'absorció, la digestió i les transformacions bioquímiques que experimenten els aliments després de la ingestió. Aquest tipus de metaboloma humà es compon de més de 25.000 compostos i és extremadament complex i variable segons la dieta. Aquesta tesi se centra principalment en aquest tipus de metaboloma.

- **Metaboloma endogen**

El metaboloma endogen és el conjunt de compostos presents en el metaboloma humà que són produïts de manera natural per l'organisme. De la mateixa manera que el metaboloma alimentari, aquest metaboloma també es pot veure influït per la dieta, ja que aquesta pot alterar els metabòlits endògens i les vies metabòliques en que participen.

- **Metaboloma de farmacològic i de contaminants**

El metaboloma farmacològic és el conjunt de compostos xenobiòtics derivats de medicaments i/o nutracèutics presents en el metaboloma humà, mentre que el metaboloma de contaminants està compost per derivats metabòlics de la contaminació ambiental.

5.1.2 Tècniques d'obtenció de perfils de metabolòmics

Les tècniques més utilitzades per l'obtenció de perfils de metabolòmics són l'espectrometria de masses i l'espectroscòpia de ressonància magnètica nuclear (NMR).

5.1.2.1 Espectrometria de masses

L'espectrometria de masses (MS) és una tècnica analítica que mesura la relació massa-càrrega dels ions. Els resultats es presenten normalment com un espectre de masses, un gràfic d'intensitats en funció de la relació massa-càrrega. L'espectrometria de masses s'utilitza en molts camps diferents i es pot utilitzar en mostres sòlides, líquides o gasoses,

tant pures com complexes.

Una millora notable per a la tècnica d'espectrometria de masses és utilitzar-la conjuntament amb diferents tècniques cromatogràfiques, com per exemple, la cromatografia de gasos o la cromatografia de líquids.

Totes les dades utilitzades en aquesta tesi han estat obtingudes a partir de la combinació de la cromatografia de líquids d'alt rendiment amb l'espectrometria de masses.

5.1.2.2 Ressonància magnètica nuclear

La ressonància magnètica nuclear (NMR) és una tècnica espectroscòpica per observar els camps magnètics locals al voltant dels nuclis atòmics. Aquesta tècnica s'utilitza per identificar proteïnes, metabòlits i altres molècules complexes. A més de la identificació, l'espectroscòpia de NMR proporciona informació detallada sobre l'estructura, la dinàmica, l'estat de reacció i l'entorn químic de les molècules. Els espectres de NMR són únics, reproduïbles i sovint molt predicibles per a molècules petites.

5.1.3 Estratègies d'obtenció de perfils de metabolòmics

Les estratègies d'obtenció de perfils de metabolòmics es poden dividir en dos grups; la metabolòmica no dirigida i la metabolòmica dirigida (Roberts et al., 2012).

La metabolòmica no dirigida se centra en la detecció global i la quantificació relativa de totes les petites molècules en una mostra, inclosos els compostos químics desconeguts.

Aquest tipus d'estratègia s'ha de combinar amb tècniques avançades per reduïr els extensos conjunts de dades generats a un conjunt més petit de senyals gestionables (Schrimpe-Rutledge et al., 2016). Aquesta aproximació ofereix l'oportunitat de descobrir noves dianes, ja que la cobertura del metaboloma només està restringida per les metodologies

de preparació de mostres i la sensibilitat i especificitat inherents de la tècnica analítica emprada. Tanmateix, els principals reptes de l'anàlisi no dirigida radiquen en els protocols i el temps necessari per processar les grans quantitats de dades brutes generades, les dificultats per identificar i caracteritzar petites molècules desconegudes i el biaix cap a la detecció de molècules amb una elevada abundància (Schrimpe-Rutledge et al., 2016).

En canvi, la metabolòmica dirigida se centra en la detecció de grups definits de metabòlits químicament caracteritzats i anotats bioquímicament, amb l'oportunitat d'una quantificació absoluta.

La metabolòmica dirigida té com a objectiu mesurar grups predefinits de metabòlits caracteritzats i interpretats bioquímicament (un subconjunt del metaboloma). Aquesta reducció de la cobertura del metaboloma significa que la metabolòmica dirigida depèn d'un coneixement previ dels metabòlits i de les seves vies bioquímiques, cosa que dificulta el descobriment de dianes metabòliques noves (Roberts et al., 2012). En aquesta aproximació, el processament i l'anàlisi de dades solen ser menys intensos en "mà d'obra" en comparació amb la metabolòmica no dirigida, ja que no és necessari identificar compostos desconeguts.

Totes les dades utilitzades en aquesta tesi s'han obtingut mitjançant metabolòmica dirigida.

5.2 Nutrimetabolòmica

Els recents avenços en el camp de la metabolòmica han permès una millor comprensió de les rutes metabòliques, les funcions gèniques o la regulació d'enzims importants. Al mateix temps, la integració de la metabolòmica amb la nutrició (metabolòmica nutricional o nutrimetabolòmica) milloren les pràctiques clíniques i de recerca actuals proporcionant una visió més profunda de les relacions entre diversos metabòlits i l'estat de salut (Ulaszewska et al., 2019).

L'objectiu principal de la nutrimetabolòmica és estudiar les

pertorbacions del metaboloma humà provocades per dietes específiques, aliments, nutrients, microorganismes o compostos bioactius.

La nutrimetabolòmica proporciona biomarcadors més individualitzats que altres tècniques i s'espera que proporcioni millors indicadors d'efectes dietètics. En última instància, la nutrimetabolòmica té com a objectiu aconseguir una nutrició pronòstica i diagnòstica personalitzada, convertint la nutrimetabolòmica en una de les vies més prometedores per millorar l'atenció nutricional i el tractament dietètic dels pacients en el futur (Ulaszewska et al., 2019).

5.2.1 Estudis nutricionals

Els estudis nutricionals es poden dividir en dos grups: els estudis d'intervenció i els estudis observacionals (o de cohort).

En els estudis d'intervenció, els participants de l'estudi consumeixen els aliments d'interès en una sola ingesta (estudi agut) o en menjars repetits durant un període de temps (estudis a mig o llarg termini) (Scalbert et al., 2014). En estudis aguts, els biofluids es recullen durant un període de fins a 24 hores després del consum dels aliments d'interès i es compara amb els participants que consumeixen un aliment control, identificant així biomarcadors potencials per aquells aliments d'interès (Scalbert et al., 2014). L'orina és el biofluid de referència en aquest tipus d'estudis (Tebani & Bekri, 2019).

Els estudis observacionals també poden jugar un paper important en el descobriment de biomarcadors. Els baixos consumidors (o no consumidors) i alts consumidors es seleccionen a partir de les dades sobre ingesta d'aliments recollides mitjançant qüestionaris de freqüència de consum, recordatoris dietètics, o altres tècniques d'avaluació dietètica. En aquest cas, es comparen perfils metabolòmics entre aquests subgrups per donar a conèixer possibles biomarcadors dietètics que reflecteixen la ingesta d'aliments habituals, sempre que aquests biomarcadors tinguin una vida mitjana suficient a l'organisme i que els aliments es consumeixin regularment (Scalbert et al., 2014).

5.2.2 Mètodes per a l'assessorament dietètic

Durant dècades, la recerca nutricional ha estat un pilar crucial per revelar les relacions entre dieta i salut tant a escala individual com poblacional. No obstant, la consistència, la validació i la reproductibilitat de l'assessorament dietètic han estat els grans punts limitants (Tebani & Bekri, 2019). Malgrat els inconvenients coneguts d'aquests mètodes, els recordatoris dietètics de 24 hores (DR) i els qüestionaris de freqüència de consum (FFQ) han estat els mètodes més utilitzats per a l'assessorament dietètic durant els darrers anys (Park et al., 2018).

Els recordatoris dietètics inclouen una col · lecció estructurada d'informació detallada sobre la ingesta d'aliments durant les 24 hores anteriors a la realització del qüestionari. Així doncs, aquesta metodologia consisteix en un formulari obert on els participants poden informar de tot tipus d'aliments i receptes que han menjat durant el dia anterior.

D'altra banda, els qüestionaris de freqüència de consum són eines d'avaluació dietètica en forma de qüestionari per estimar la freqüència i, en alguns casos, la mida de la porció del consum d'aliments i begudes durant un període de temps especificat, normalment el darrer mes, tres mesos o any. A diferència d'un DR, un FFQ és un qüestionari tancat, sovint de 80 a 120 ítems (inclosos aliments i begudes).

Els FFQs són l'opció preferida en estudis a gran escala de dieta i salut, ja que, malgrat que tant els DRs com els FFQs requereixen una preparació notable abans de la implementació, la gestió i el processament d'un FFQ validat és menys complex per estudis de grans dimensions (Tebani & Bekri, 2019). Els FFQs també es poden utilitzar per calcular diferents índexs nutricionals i de salut, per exemple, l'AHEI-2010 (Chiuve et al., 2012).

Aquest índex es va desenvolupar com a mesura alternativa de la qualitat de la dieta per identificar el risc de patir malalties cròniques relacionades amb la dieta en el futur (Leung et al., 2012, 2014; Wang et al., 2014).

El gran nombre d'inconvenients que presenten els mètodes

d'assesorament dietètic esmentats, com per exemple la variació nutricional entre dies, les estimacions imprecises de la mida de les porcions o la manca d'objectivitat, fan palesa la necessitat d'una aproximació més robusta i objectiva per a l'assesorament dietètic. La utilització de les *omiques* (com la metabolòmica) permet estudiar la nutrició de forma integral mitjançant la identificació de biomarcadors específics de la dieta per avaluar objectivament la ingesta. Aquests biomarcadors poden proporcionar informació útil per omplir els buits dels mètodes d'assesorament dietètic utilitzats actualment (Bekri, 2016).

5.3 Biomarcadors

El terme "*biomarcador*" fa referència a una molecula indicativa d'un estat biològic i que es pot mesurar de manera precisa i reproduïble. No obstant, hi ha altres definicions de biomarcador a la literatura actualment, que afortunadament, coincideixen en gran mesura.

En nutrimetabolòmica, s'entén com a biomarcador un metabòlit que informa sobre la ingesta d'un aliment específic o d'un patró d'alimentació. Així doncs, es necessiten biomarcadors robustos i informatius per comprendre la interacció entre la dieta i la salut (Trepanowski & Ioannidis, 2018; Zeevi et al., 2015).

Tot i això, els biomarcadors dietètics o nutricionals tenen dos grans limitacions; 1) encara no hi ha un consens clar sobre la definició de biomarcador nutricional, la seva avaluació i ús, i 2) fins i tot els biomarcadors ben validats no tenen la consistència suficient per donar suport a recomanacions nutricionals clares (Tebani & Bekri, 2019).

Malgrat aquestes limitacions, la definició més acceptada de biomarcador dietètic va ser proposada per la iniciativa FoodBAll (`https://foodmetabolome.org`), classificant els biomarcadors dietètics com:

"*Mesures específiques dins del cos que reflecteixen amb precisió la ingesta d'un component alimentari o aliment*".

Aquests biomarcadors dietètics es poden dividir en dos grans grups: els biomarcadors d'ingesta (o biomarcadors exògens), que provenen directament de la dieta, i els biomarcadors d'efecte (o biomarcadors endògens), que no es troben al metaboloma alimentari però que es poden veure alterats per la dieta.

5.4 Anàlisi de dades nutrimetabolòmiques

Sovint, l'anàlisi de dades és un dels punts crítics en els estudis de nutrimetabolòmica. En aquest context, l'anàlisi de dades s'entén com el procés d'aplicació sistemàtica de tècniques estadístiques per extraure interpretacions biològiques de les dades. Aquest procés es compon de diferents operacions de preprocessament (com la imputació de valors faltants i la normalització), diferents mètodes estadístics i d'anàlisi de la significació biològica. No obstant, hi ha una gran varietat de mètodes per dur a terme processos d'anàlisi de dades, cadascun amb les seves diferents aplicacions, avantatges i desavantatges.

L'aplicació web POMAShiny desenvolupada en el context d'aquesta tesi se centra en aquesta secció, proporcionant diferents mètodes per l'anàlisi de dades de metabolòmica (vegeu Taula 3.1).

5.4.1 Modelització estadística

L'estadística clàssica proporciona als investigadors un conjunt d'eines ben consolidades per abordar qüestions de recerca com per exemple comparar la resposta a diferents tractaments o modelar els efectes d'un conjunt de variables sobre la concentració d'un metabòlit. Els models estadístics clàssics acostumen a ser molt flexibles, permetent incloure en els models coneixement extern i variables de confusió (Ulaszewska et al., 2019). Aquest tipus de models solen ser força interpretables ja que degut a les múltiples colinealitats entre variables i de la manca d'informació rellevant en algunes d'elles, només algunes variables d'estudi s'inclouen al model final, facilitant la seva interpretació.

Els mètodes de modelització estadística més utilitzats en el camp

de la nutrimetabolòmica són els models lineals (LM), els models lineals generalitzats (GLM) i els models additius generalitzats (GAM).

5.4.2 Mineria de dades

Els mètodes de mineria de dades se centren en l'anàlisi de diverses variables al mateix temps, tenint en compte les diferents relacions entre elles. Aquests mètodes poden proporcionar informació sobre l'estructura de les dades i les diferents relacions internes que no s'observarien amb els models estadístics clàssics. No obstant, la interpretació d'aquests mètodes pot ser molt més complexa.

Els mètodes de mineria de dades més utilitzats en el camp de la nutrimetabolòmica són l'anàlisi de components principals (PCA) i els mínims quadrats parcials (PLS), amb algunes de les seves variants.

L'anàlisi de components principals és un mètode no supervisat per a la reducció de la dimensió d'una matriu de dades. Aquest mètode consisteix en el càlcul de la matriu de covariança de les dades i la descomposició dels valors propis en aquesta matriu de covariància sense considerar cap variable resposta (com ara el grup d'intervenció, tractament, etc.).

D'altra banda, el mètode dels mínims quadrats parcials (PLS) és una alternativa supervisada a l'anàlisi de components principals. El mètode PLS també és un mètode de reducció de la dimensió, que tracta d'identificar un nou conjunt de variables que són combinacions lineals de les variables originals i, a continuació, ajusta un LM amb aquestes noves variables (James et al., 2013). A diferència del PCA, el PLS identifica aquestes noves variables de manera supervisada, és a dir, a partir de la variable resposta.

No obstant, aquests mètodes multivariants poden donar resultats molt atractius, que malauradament no es poden generalitzar per a tots els estudis de nutrimetabolòmica. Per tant, és necessària una validació exhaustiva d'aquests mètodes (Ulaszewska et al., 2019).

5.4.3 Aprenentatge estadístic

En els últims anys, els mètodes d'aprenentatge estadístic han rebut una atenció creixent en el camp de la nutrimetabolòmica. Aquesta adopció generalitzada es deu a la necessitat d'identificar biomarcadors dietètics o panells de biomarcadors dietètics amb capacitat predictiva per assolir l'ambiciós objectiu de predir la ingesta d'aliments a partir dels metabòlits en orina, sang o sèrum, i no només identificar aquells metabòlits associats a determinats patrons dietètics o aliments. Avui en dia, aquest objectiu s'ha convertit en un dels objectius més importants en el camp de la nutrimetabolòmica.

Tanmateix, aquests mètodes també poden presentar alguns inconvenients, com l'elevat nombre de mostres necessàries per dur-los a terme amb una certa robustesa, o la baixa interpretabilitat d'aquests mètodes, més difícils d'interpretar que els models estadístics clàssics o els mètodes de mineria de dades, respectivament.

Els mètodes d'aprenentatge estadístic *LASSO*, regressió de *Ridge*, *elasticnet* i boscos aleatoris, s'han implementat com a recurs per a problemes de classificació a les eines POMA i POMAShiny, presentades més endavant a la secció de resultats.

5.5 Ontologies

La creixent aparició de tècniques analítiques d'alt rendiment en les ciències de la vida durant les darreres tres dècades ha creat desafiaments significatius en la gestió de les dades generades. Actualment, un dels principals problemes als quals s'enfronten els investigadors rau en la pregunta: *on són aquestes dades i com les podem utilitzar?* Malauradament, l'heterogeneïtat de les plataformes d'emmagatzematge i els formats de les dades sovint dificulten el seu accés i ús generalitzat.

En aquest sentit, la creació d'ontologies, definida com a "*l'especificació d'un vocabulari de representació per a un domini compartit - definicions de classes, relacions, funcions i altres objectes*" (Kramer & Beißbarth, 2017), és de vital importància per ajudar a

analitzar, anotar i homogeneïtzar aquests grans i complexos conjunts de dades (Hoehndorf et al., 2015; Schlegel et al., 2015).

Per tant, una ontologia adequada hauria de proporcionar una representació formal del coneixement en un domini determinat (per exemple, la nutrició o la metabolòmica), de manera que tant els humans com els ordinadors puguin entendre els conceptes que conté i aprendre sobre el domini que s'està representant (Rubin et al., 2008). Normalment, les ontologies s'estructuren dins d'una jerarquia del coneixement on els conceptes estan connectats mitjançant relacions semàntiques estandarditzades (per exemple, "és part de" o "és un ingredient de") especificant formalment relacions de coneixement, com ara generalitzacions d'especificacions del domini d'interès (Vitali et al., 2018).

Mentre existeixen diferents ontologies per a definir conceptes específics dels camps de la nutrició i la metabolòmica, com per exemple FoodOn (Dooley et al., 2018), ONS (Vitali et al., 2018) o ChEBI (Degtyarenko et al., 2007), no existeix actualment cap ontologia específica per al camp de la nutrimetabolòmica que defineixi les relacions entre els aliments i els biomarcadors alimentaris.

Actualment, diferents bases de dades proporcionen informació sobre metabòlits i aliments, incloent Exposome-Explorer (Neveu et al., 2016), Phenol-Explorer (Rothwell et al., 2013), PhytoHub (`http://phytohub.eu/`) i Food Database (FooDB) (`http://foodb.ca/`). Malgrat que totes aquestes bases de dades contenen informació sobre aliments i els seus metabòlits associats, la complexitat d'aquestes relacions fa que es descriguin de forma molt diferent entre elles.

Aquesta manca d'unificació de la informació i la manca d'una estructura jeràrquica dificulten la comparació i cerca de dades. Per tant, és necessàri el desenvolupament d'una ontologia per definir de forma clara les relacions entre les dades nutricionals i metabolòmiques (Maruvada et al., 2020). Aquesta ontologia podria tenir múltiples aplicacions pràctiques en estudis de nutrimetabolòmica, sent l'anotació de termes la més evident, però incloent també altres aplicacions com la realització de diferents anàlisis de significació biològica i la realització

d'anàlisis de similitud semàntica.

En el context d'aquest treball, s'ha desenvolupat una ontologia anomenada FOBI (Ontologia d'Aliments i Biomarcadors) amb l'objectiu de proporcionar un llenguatge comú estandaritzat per descriure les complexes relacions entre la dieta i els seus metabòlits associats en estudis de nutrimetabolòmica. L'ontologia FOBI es presenta més endavant a la secció de resultats.

5.6 Anàlisi de la significació biològica

A diferència del concepte de "*significació estadística*", el concepte de "*significació biològica*" es pot tractar des de diferents perspectives i definir-lo de diferents maneres. Per exemple, aquest concepte es pot tractar des d'un punt de vista clínic, fent referència a un efecte que té un impacte notable sobre la salut, o des d'un punt de vista *òmic*, fent referència a la interpretació biològica o a la rellevància biològica de les diferències estadístiques en experiments d'*òmiques*. En aquesta tesi, aquest concepte es tracta exclusivament des d'una perspectiva *òmica*.

L'anàlisi de la significació biològica (BSA), també conegut com a anàlisi d'enriquiment, anàlisi d'enriquiment de rutes o anàlisi d'enriquiment funcional, denota qualsevol mètode que es beneficiï de la informació coneguda de rutes o xarxes biològiques per extraure coneixement d'un sistema biològic (Creixell et al., 2015; Reimand et al., 2019). En altres paraules, aquest tipus d'anàlisi integra el coneixement biològic existent (de diferents fonts com bases de dades i ontologies) i els resultats estadístics dels estudis *òmics*, obtenint una comprensió més profunda dels sistemes biològics. Atès que els BSA utilitzen els resultats derivats de les anàlisis estadístiques, aquests consisteixen en el darrer pas d'un procés d'anàlisi de dades *òmiques* (Figura 1.7).

En la majoria d'estudis d'*òmiques*, el resultat de les anàlisis estadístiques sol ser una llista de variables seleccionades segons

criteris estadístics predefinits. Els mètodes de BSA utilitzen aquestes variables seleccionades per explorar rutes, malalties, etc. biològicament rellevants/significants, segons la naturalesa de la llista de variables (com gens o metabòlits) i la font utilitzada per extraure el coneixement biològic previ (ontologies i bases de dades). Per tant, la informació d'entrada dels BSA sol ser una llista de variables biològiques i el resultat sol ser una llista de rutes biològiques associades a la llista d'entrada, amb la seva significació estadística corresponent (Figura 1.19).

Per tant, aquests mètodes permeten als investigadors passar de llistes de gens o metabòlits a rutes metabòliques, malalties i altres característiques associades a aquestes llistes.

5.6.1 Mètodes d'anàlisi de la significació biològica

Degut al gran nombre d'aplicacions d'aquests mètodes i del seu gran ús en el camp de les *òmiques*, en els darrers anys s'han proposat diferents enfocaments per a dur a terme BSA. Actualment, les aproximacions més populars utilitzades pels BSA són l'anàlisi de sobre-representació (ORA) i l'anàlisi d'enriquiment de conjunts de gens (GSEA), amb les seves variants per a altres camps com l'anàlisi d'enriquiment de conjunts de metabòlits (MSEA) (Xia & Wishart, 2010).

La figura 1.19 mostra de forma molt sintètica aquests dos enfocaments més comuns per realitzar BSA, amb els seus diferents requeriments d'entrada. Tenint en compte aquest esquema, l'ORA requereix la llista de gens (o metabòlits) seleccionats sense cap ordre. D'altra banda, el mètode GSEA requereix una llista ordenada de gens, juntament amb la mètrica que s'ha utilitzat per ordenar-los, com ara el p-valor o la mida de l'efecte. En ambdós casos, ORA i GSEA, el resultat de l'anàlisi d'enriquiment és una taula amb les vies metabòliques enriquides juntament amb la seva significació estadística associada.

5.6.2 Anàlisi de la significació biològica en nutrimetabolòmica

Actualment, els BSA en estudis de nutrimetabolòmica es realitzen mitjançant eines no específiques (per exemple, mètodes d'ORA i MSEA concebuts per a estudis de metabolòmica general).

Sovint, aquests mètodes d'anàlisi d'enriquiment utilitzen bases de dades com KEGG (`https://www.genome.jp/kegg/`) o REACTOME (`https://reactome.org`) per obtenir la informació biològica de diferents conjunts de metabòlits i rutes metabòliques. Tanmateix, l'ús d'aquestes bases de dades genèriques, enfocades a les vies metabòliques de diferents organismes, pot excloure de l'anàlisi aquells compostos exògens derivats de la dieta i altres substàncies que no participen directament en les vies metabòliques, per exemple, els metabòlits que formen part del metaboloma alimentari, farmacològic i de la contaminació.

Aquesta limitació requereix el desenvolupament de noves eines i mètodes que permetin als investigadors realitzar anàlisis d'enriquiment robustos a partir de llistes de metabòlits derivats d'estudis de nutrimetabolòmica, considerant els metabòlits exògens i les seves relacions amb els aliments.

En el context d'aquesta tesi, s'han desenvolupat diferents mètodes per fer possible el nou concepte d'anàlisi d'enriquiment alimentari. Aquestes noves metodologies permetran als investigadors obtenir grups d'aliments enriquits a partir de llistes de metabòlits.

Aquest bloc fa referència a un dels principals objectius d'aquesta tesi i tots els mètodes desenvolupats amb aquesta finalitat s'han integrat dins del marc de treball fobitools, presentat més endavant a la secció de resultats.

Chapter 6

Objectius

Aquesta tesi se centra en el desenvolupament de mètodes i eines per avançar en el descobriment de les complexes relacions entre la dieta i els metabòlits derivats dels aliments, així com en facilitar la interpretació dels resultats en estudis de nutrimetabolòmica.

6.1 Objectiu principal

L'objectiu principal d'aquesta tesi és l'estudi en profunditat de les relacions entre el metaboloma alimentari i la ingesta d'aliments en el context de diferents estudis de nutrimetabolòmica. Aquest és un objectiu transversal i en depenen tots els objectius específics.

6.2 Objectius específics

Complementàriament a l'objectiu principal, es considera la següent llista d'objectius específics:

1. El desenvolupament d'una ontologia que defineixi clarament les relacions entre metabòlits i aliments mitjançant un llenguatge comú i estandarditzat.

    (a) El desenvolupament d'una eina de codi obert que permeti la fàcil consulta i utilització de l'ontologia creada en l'objectiu específic 1. Aquesta eina permetrà funcions com:

   i. L'anotació automàtica de dades dietètiques de text lliure.
   ii. L'anàlisi de la significació biològica en estudis de nutrimetabolòmica.

   (b) Proporcionar una interfíce gràfica de codi obert per l'eina desenvolupada en l'objectiu específic anterior (a) per tal de fer-la més accessible a la comunitat científica.

2. El desenvolupament d'una eina per l'anàlisi estadística de dades de metabolòmica que inclogui mètodes alternatius/complementaris per ajudar a millorar el procés de descobriment de biomarcadors en el context dels estudis de nutrimetabolòmica.

   (a) Proporcionar una interfíce gràfica de codi obert per l'eina desenvolupada en l'objectiu específic 2 per tal de fer-la més accessible a la comunitat científica.

3. Aplicar les eines desenvolupades en els objectius específics 1 i 2 a estudis de nutrimetabolòmica reals.

   (a) Identificar metabòlits o grups de metabòlits associats a l'índex de dieta saludable AHEI-2010.

   (b) Identificar metabòlits o grups de metabòlits associats al risc de malaltia o a l'estat de salut.

Chapter 7

Resultats

Totes les publicacions presentades en aquesta tesi s'han enviat a revistes científiques internacionals. En aquest apartat, cada publicació es presenta com un breu resum del contingut del treball, juntament amb el factor d'impacte de la revista, el quartil i el decil (si el manuscrit ja ha estat acceptat).

Les publicacions es divideixen en desenvolupaments metodològics i aplicacions dels mètodes desenvolupats. En els treballs metodològics, les contribucions realitzades compleixen els objectius de la tesi, mentre que en els treballs d'aplicació s'aplica la metodologia i les eines desenvolupades per descobrir i identificar metabòlits associats a la dieta (biomarcadors dietètics), a l'AHEI-2010 i a l'estat de salut o malaltia.

L'última secció d'aquesta secció està dedicada a tot el programari, tant els paquets d'R com les interfícies gràfiques (GUIs), desenvolupat en el context d'aquesta tesi.

7.1 Desenvolupaments metodològics i de programari

7.1.1 Article 1: L'ontologia de biomarcadors i aliments

Pol Castellano-Escuder, Raúl González-Domínguez, David S. Wishart, Cristina Andrés-Lacueva, Alex Sánchez-Pla (2020). *FOBI: una ontologia per representar dades d'ingesta d'aliments i associar-les a dades metabolòmiques*. Database, 2020.

- Factor d'impacte de la revista: **2.593**
- Quartil/decil de la revista: **Q2/D3** (15 de 59 - Biologia matemàtica i computacional)

RESUM

La recerca nutricional es pot dur a terme mitjançant dos enfocaments complementaris: (i) mètodes tradicionals de recopilació de la informació de la ingesta o (ii) mitjançant tècniques de metabolòmica per analitzar biomarcadors d'ingesta d'aliments en els biofluids corporals. No obstant, la complexitat i l'heterogeneïtat d'aquests dos tipus de dades tan diferents sovint dificulten la seva anàlisi i integració. Per combatre aquest repte, hem desenvolupat una nova ontologia que descriu els aliments i els seus metabòlits associats de manera jeràrquica. Aquesta ontologia utilitza un sistema formal de denominació, definicions de categories, propietats i relacions entre ambdós tipus de dades. L'ontologia presentada s'anomena FOBI (Food-Biomarker Ontology) i està formada per dues subontologies interconnectades. La primera és l'ontologia dels aliments, que consisteix en diferents aliments simples i aliments "multi-component", mentre que la segona és l'ontologia de biomarcadors, que conté biomarcadors d'ingesta d'aliments classificats en les seves classes químiques. Aquestes dues subontologies són conceptualment independents, però interconnectades per diferents propietats. Això permet visualitzar dades i informació sobre aliments i biomarcadors d'aliments de manera bidireccional, passant de la metabolòmica a les dades nutricionals o viceversa. Les possibles aplicacions d'aquesta ontologia inclouen l'anotació d'aliments i biomarcadors mitjançant una nomenclatura

ben definida i estandaritzada, l'estandardització de fluxos de treball de nutrimetabolòmica o l'aplicació de diferents anàlisis d'enriquiment en estudis de nutrimetabolòmica.

7.1.2 Article 2: POMAShiny

Pol Castellano-Escuder, Raúl González-Domínguez, Francesc Carmona-Pontaque, Cristina Andrés-Lacueva, Alex Sánchez-Pla (2021). *POMAShiny: un flux de treball web fàcil d'utilitzar per l'anàlisi de dades de metabolòmica i proteòmica.* PLOS Computational Biology, 2021.

- Factor d'impacte de la revista: **4.7**
- Quartil/decil de la revista: **Q1/D2** (6 of 59 - Biologia matemàtica i computacional)
- Quartil/decil de la revista: **Q1/D2** (9 of 77 - Mètodes de recerca bioquímica)

RESUM

La metabolòmica i la proteòmica, com altres *òmiques*, solen afrontar un complex repte de mineria de dades per tal de proporcionar un resultat comprensible i interpretable. Sovint, l'anàlisi estadística és un dels reptes més complexos i és fonamental en la interpretació biològica dels resultats. Per aquest motiu, combinat amb les habilitats de programació necessàries per a dur a terme aquests tipus d'anàlisi, s'han proposat diferents eines bioinformàtiques dirigides a simplificar l'anàlisi de dades de metabolòmica i proteòmica durant els últims anys. No obstant això, a vegades l'anàlisi encara està limitat a un nombre reduït de mètodes estadístics amb una flexibilitat reduïda. POMAShiny és una eina web que proporciona un flux de treball estructurat, flexible i fàcil d'utilitzar per a la visualització, exploració i anàlisi estadística de dades de metabolòmica i proteòmica. Aquesta eina integra diversos mètodes estadístics, alguns d'ells àmpliament utilitzats en altres tipus d'*òmiques*, i es basa en el paquet de Bioconductor POMA, fet que suposa un increment en la reproductibilitat i la flexibilitat de les anàlisis fora de l'entorn web. POMAShiny i POMA estan disponibles a `https://github.com/nutrimetabolomics/POMAShiny`

i `https://github.com/nutrimetabolomics/POMA`, respectivament.

7.1.3 Article 3: L'entorn de treball fobitools

Pol Castellano-Escuder, Cristina Andrés-Lacueva, Alex Sánchez-Pla. *L'entorn de treball fobitools: els primers passos cap a l'anàlisi d'enriquiment alimentari. En revisió.*

RESUM

L'ontologia FOBI pot ser de gran ajuda en estudis nutrimetabolòmics a causa de la seva gran varietat d'aplicacions, inclosa la possibilitat de realitzar diferents anàlisis d'enriquiment. Tot i això, els coneixements de programació necessàris per utilitzar-la poden limitar-ne l'ús per part de la comunitat científica. Aquí presentem l'entorn de treball fobitools, format per un paquet de Bioconductor i la seva interfície gràfica complementària. Aquestes dues eines permeten als investigadors interactuar i explorar l'ontologia FOBI d'una manera senzilla. L'entorn de treball fobitools està centrat en el nou concepte d'anàlisi d'enriquiment alimentari en estudis de nutrimetabolòmica. Adicionalment, també es presenten altres funcions útils, com la visualització interactiva en xarxa de FOBI o l'anotació automàtica de dades dietètiques de text lliure. Tant el paquet fobitools com l'aplicació web fobitoolsGUI, estan disponibles a `https://github.com/nutrimetabolomics/fobitools` i `https://github.com/nutrimetabolomics/fobitoolsGUI`, respectivament.

7.2 Aplicació de les eines desenvolupades

7.2.1 Article 4: Avaluació de l'adherència a hàbits dietètics saludables mitjançant el metaboloma alimentari en orina

Pol Castellano-Escuder, Raúl González-Domínguez, Marie-France Vaillant, Patricia Casas-Agustench, Nicole Hidalgo-Liberona, Núria

Estanyol-Torres, Thomas Wilson, Manfred Beckmann, Amanda J Lloyd, Marion Oberli, Christophe Moinard, Christophe Pison, Jean-Christian Borel, Marie Joyeux-Faure, Mariette Sicard, Svetlana Artemova, Hugo Terrisse, Paul Dancer, John Draper, Alex Sánchez-Pla, Cristina Andres-Lacueva. *Avaluació de l'adherència a hàbits dietètics saludables mitjançant el metaboloma alimentari en orina. Enviat.*

RESUM

La dieta és un dels factors d'estil de vida modificables més importants en la salut humana i en la prevenció de malalties cròniques. Per tant, l'avaluació dietètica precisa és essencial per avaluar de forma fiable l'adherència a hàbits saludables. L'objectiu d'aquest estudi era identificar metabòlits en orina que poguessin servir com a biomarcadors robusts de qualitat de la dieta, tal com s'avalua a través de l'AHEI-2010.

En aquest estudi vam recollir mostres de dos centres de 160 voluntaris sans, d'entre 25 i 50 anys, que vivien en parella o en família, amb mostres d'orina i una avaluació dietètica repetides al principi de l'estudi, als sis mesos i als dotze mesos al llarg de l'any. Les mostres d'orina es van sotmetre a un anàlisi de metabolòmica a gran escala per obtenir una caracterització quantitativa completa del metaboloma alimentari. A continuació, es van dur a terme diferents anàlisis de regressió regularitzada i anàlisis *limma* per identificar aquells metabòlits associats a l'AHEI-2010 i investigar la reproductibilitat d'aquestes associacions al llarg del temps.

El resultat més rellevant va ser l'associació positiva de nombrosos metabòlits microbians de polifenols amb la puntuació AHEI-2010. A més, es van identificar fortes associacions entre l'AHEI-2010 i els metabòlits relacionats amb la ingesta de cafè, carn vermella i peix. Així doncs, en aquest estudi es demostra que els metabòlits en orina, i en particular els derivats de la microbiota, podrien servir com a indicadors fiables de l'adherència a hàbits dietètics saludables.

7.2.2 Article 5: El metaboloma serològic relacionat amb els aliments s'associa amb un detreioriament cognitiu tardà en individus d'edat avançada

Raúl González-Domínguez, **Pol Castellano-Escuder**, Francisco Carmona, Sophie Lefèvre-Arbogast, Dorrain Y. Low, Andrea Du Preez, Silvie R. Ruigrok, Claudine Manach, Mireia Urpi-Sarda, Aniko Korosi, Paul J. Lucassen, Ludwig Aigner, Mercè Pallàs, Sandrine Thuret, Cécilia Samieri, Alex Sánchez-Pla, Cristina Andres-Lacueva. *El metaboloma serològic relacionat amb els aliments s'associa amb un detreioriament cognitiu tardà en individus d'edat avançada. Enviat.*

RESUM

Actualment, es reconeix que la nutrició i els compostos bioactius de la dieta són crucials en l'aparició del deteriorament cognitiu i de la demència, però la majoria de les proves existents al respecte són transversals i, sovint, són inconsistents i fragmentades. Per donar a conèixer el rol de la dieta en la patogènesi primerenca del deteriorament cognitiu, vam estudiar dues cohorts independents i prospectives a llarg termini de casos i controls niats a partir de participants sense demència al principi de l'estudi "Three-City".

Es va realitzar una anàlisi metabolòmica de mostres de sèrum per obtenir una caracterització completa i quantitativa del metaboloma relacionat amb els aliments. Després, l'anàlisi de regressió LASSO va revelar una associació protectora de diversos metabòlits derivats del cacau, el cafè, els bolets, el vi negre i el metabolisme microbià d'aliments rics en polifenols amb la funció cognitiva, així com l'efecte nociu del consum de productes alimentaris associats a dietes poc saludables, com ara com a aliments edulcorats artificialment, que contenen alcohol i processats. A més, també vam observar importants pertorbacions metabòliques que afecten el metabolisme relacionat amb la microbiota dels aminoàcids aromàtics i la β-oxidació dels àcids grassos. Curiosament, tot i que les signatures específiques dels metabòlits eren diferents entre les dues cohorts d'estudi a causa de factors de variabilitat interindividuals, la majoria d'aquestes

associacions de deteriorament cognitiu amb la ingesta d'aliments concrets i vies metabòliques alterades es van corroborar en la fase de validació.

Així doncs, aquests resultats representen un pas endavant en el descobriment dels esdeveniments patològics darrere del deteriorament cognitiu precoç relacionats amb la dieta, la microbiota intestinal i el metabolisme endogen, que al seu torn podrien proporcionar dianes potencials per desenvolupar estratègies dietètiques preventives i terapèutiques per protegir la salut cognitiva.

7.3 Programari

7.3.1 Paquets de Bioconductor

En el context d'aquesta tesi s'han desenvolupat dos paquets de Bioconductor (Gentleman et al., 2004): els paquets POMA i fobitools.

Tot el programari descrit aquí es s'ha desenvolupat utilitzant el llenguatge de programació R (R Core Team, 2019) i seguint les bones pràctiques per al desenvolupament de paquets d'R descrites al llibre *R packages* de Hadley Wickham (Wickham, 2015). A més, tots els paquets descrits aquí s'han escrit seguint la filosofia tidyverse (Wickham et al., 2019), per tal de mantenir tot el codi net i llegible, facilitant la contribució d'altres usuaris i el manteniment del programari.

7.3.1.1 POMA

Aquest paquet introdueix un flux de treball estructurat, reproduïble i fàcil d'utilitzar per a la visualització, preprocessament, exploració i anàlisi estadística de dades de metabolòmica.

POMA utilitza els objectes *MSnSet* de la classe S4 definits al paquet MSnbase (Gatto & Lilley, 2012) i està totalment integrat a l'entorn Bioconductor. POMA està disponible a `https://bioconductor.org`.

Documents disponibles:

- Manual de POMA
- Vinyeta de POMA: Flux de treball POMA (vegeu el cas d'ús de POMA)
- Vinyeta de POMA: Normalització amb POMA
- Vinyeta de POMA: Exemple de POMA EDA
- Lloc web de POMA: `https://pcastellanoescuder.github.io/POMA/`
- Repositori GitHub de POMA: `https://github.com/pcastellanoescuder/POMA`

7.3.1.2 fobitools

Aquest paquet proporciona un conjunt de funcions per interactuar amb FOBI (Castellano-Escuder et al., 2020). Aquest paquet se centra en el nou concepte d'anàlisi d'enriquiment alimentari en estudis nutrimetabolòmics. Tanmateix, també s'ofereixen altres funcions útils, com ara la visualització interactiva de la xarxa de FOBI i l'anotació automàtica de dades dietètiques de text lliure.

Documents disponibles:

- Manual de fobitools
- Vinyeta de fobitools: Anàlisi de sobrerepresentació alimentari simple
- Vinyeta de fobitools: Cas d'ús ST000291 (vegeu el cas d'ús de fobitools)
- Vinyeta de fobitools: Cas d'ús ST000629
- Vinyeta de fobitools: Anotació de text dietètic
- Lloc web de fobitools: `https://pcastellanoescuder.github.io/fobitools/`
- Repositori GitHub de fobitools: `https://github.com/pcastellanoescuder/fobitools`

7.3.2 Interfícies gràfiques

En el context d'aquesta tesi s'han desenvolupat tres interfícies gràfiques (GUI). Les dues primeres GUIs POMAShiny i fobitoolsGUI es basen en les funcions dels dos paquets d'R descrits a la secció anterior, mentre que la tercera GUI es basa en dos paquets de Bioconductor preexistents anomenats msmsEDA (Gregori et al., 2020a) i msmsTests (Gregori et

al., 2020b).

Aquestes tres interfícies gràfiques han estat desenvolupades en R i utilitzant l'entorn de treball Shiny (Chang et al., 2020).

7.3.2.1 POMAShiny

POMAShiny és una eina web que proporciona un flux de treball estructurat, flexible i fàcil d'utilitzar per al preprocessament, l'exploració i l'anàlisi estadística de dades de metabolòmica. Aquesta eina es basa en el paquet de Bioconductor POMA, que incrementa la reproductibilitat i flexibilitat de l'anàlisi fora de l'entorn web. El flux de treball de POMAShiny s'estructura en quatre panells seqüencials i ben definits: 1) càrrega de dades, 2) preprocessament, 3) EDA i 4) anàlisi estadística.

Documents disponibles:

- URL POMAShiny: `https://webapps.nutrimetabolomics.com/POMAShiny`
- Repositori GitHub de POMAShiny: `https://github.com/pcastellanoescuder/POMAShiny`
- Imatge Docker de POMAShiny: `https://hub.docker.com/repository/docker/pcastellanoescuder/pomashiny`

7.3.2.2 fobitoolsGUI

fobitoolsGUI és una eina web basada en el paquet fobitools. Aquesta interfície web fàcil d'utilitzar proporciona un conjunt d'eines per interactuar amb FOBI. Aquesta aplicació proporciona una col · lecció d'eines bàsiques de manipulació per a l'anàlisi de la significació biològica, la visualització de xarxes i les estratègies de mineria de text per anotar dades nutricionals.

Documents disponibles:

- URL fobitoolsGUI: `https://webapps.nutrimetabolomics.com/fobitoolsGUI`
- Repositori GitHub de fobitoolsGUI: `https://github.com/pcastellanoescuder/fobitoolsGUI`

7.3.2.3 POMAcounts

POMAcounts és una eina web per a l'anàlisi exploratòria i l'anàlisi estadística de dades de comptatges d'espectrometria de masses. Aquesta interfície gràfica es basa en els paquets de Bioconductor msmsEDA (Gregori et al., 2020a) i msmsTests (Gregori et al., 2020b). El nom de POMAcounts ve donat per la gran semblança tant del *frontend* com del *backend* que comparteix amb POMAShiny.

Documents disponibles:

- URL de POMAcounts: `http://uebshiny.vhir.org:3838/POMAcounts`
- Repositori GitHub de POMAcounts: `https://github.com/pcastellanoescuder/POMAcounts`

Chapter 8

Discussió

Mentre que s'inclou un resum de cada publicació a la secció de resultats i les discussions complertes de cada article en els annexos d'aquest treball, aquest capítol te com a objectiu proporcionar una visió general d'aquesta tesi mitjançant una discussió que engloba tots els resultats individuals presentats anteriorment.

Com s'ha explicat a la introducció, la nutrimetabolòmica consisteix en la integració dels camps de la nutrició i la metabolòmica, esdevenint una de les vies més prometedores per millorar l'assessorament nutricional i els tractaments dietètics en el futur (Ulaszewska et al., 2019).

No obstant, les dades de nutrimetabolòmica també combinen la complexitat intrínseca de cadascun dels dos camps que integra, havent de fer front a reptes que dificulten l'anàlisi i interpretació dels seus resultats. Per exemple, la compromesa reproductibilitat de les anàlisis de metabolòmica en orina mitjançant LC-MS, el biaix de les dades nutricionals obtingudes mitjançant qüestionaris dietètics o la gran variabilitat interindividual dels individus en estudis poblacionals i d'intervenció. A més, moltes de les relacions entre els metabòlits derivats de la dieta i els aliments no es coneixen del tot, donant lloc a discussió i dificultant ja no només la identificació de potencials biomarcadors de dieta o el desenvolupament de models predictius, sinó l'estudi bàsic de la relació entre els camps de la nutrició i de la metabolòmica, del que falta molt per descobrir.

Per aquest motiu, tots els recursos i eines desenvolupats en el context d'aquesta tesi emergeixen d'aquesta complexitat de les dades de nutrimetabolòmica, proposant alternatives centrades en la millora de la seva integració, anàlisi estadística i interpretació biològica.

El coneixement de les complexes interrelacions entre tipus d'aliments, components alimentaris, ingredients alimentaris i biomarcadors d'ingesta d'aliments és fonamental per facilitar la comprensió de la nutrició i el metabolisme. Aquesta comprensió permetrà en un futur obtenir avaluacions molt més precises i objectives de la ingesta d'aliments basades en la metabolòmica. Al teu torn, també ajudarà al desenvolupament d'estratègies nutricionals personalitzades per dissenyar dietes específiques segons el fenotip, estat de la malaltia, microbioma o metaboloma del pacient, entre d'altres.

En vista d'aquest escenari, el primer gran objectiu d'aquesta tesi es va centrar en estudiar, caracteritzar i definir les relacions entre els metabòlits i la dieta d'una forma clara i robusta, que permetés a la comunitat científica partir d'un punt de consens a l'hora de dissenyar estudis de nutrimetabolòmica, interpretar els seus resultats o bé establir comparacions i metanàlisis entre diferents estudis d'aquest camp (**objectiu específic 1**).

Existeixen bases de dades molt complertes actualment amb l'objectiu de proporcionar informació sobre aquells metabòlits associats a determinats aliments, com per exemple les bases de dades Exposome-Explorer (Neveu et al., 2016), Phenol-Explorer (Rothwell et al., 2013), PhytoHub (`http://phytohub.eu/`) i Food Database (FooDB) (`http://foodb.ca/`). No obstant, aquests recursos descriuen aquesta informació d'una forma molt acurada però també heterogènia entre ells, dificultant la comparació d'estudis i a vegades proporcionant informació lleugerament diferent per als mateixos compostos o aliments. A més, tot i els avantatges computacionals que això suposa (veure introducció), no existeix actualment cap recurs en forma d'ontologia per definir aquestes associacions.

Per aquest motiu, hem desenvolupat l'ontologia FOBI (**objectiu específic 1**). FOBI és una ontologia específica del camp de la nutrimetabolòmica composta per dues subontologies amb jerarquies

independents però interrelacionades; la subontologia d'aliments i la subontologia de biomarcadors. La subontologia d'aliments consisteix en diferents aliments simples i complexos (multi-component) agrupats segons els seus principals grups nutricionals, mentre que la subontologia de biomarcadors conté els metabòlits derivats de la dieta classificats segons les seves classes químiques. D'aquesta manera, el llenguatge comú i estandarditzat utilitzat per definir i relacionar els elements de cada subontologia permeten interpretar fàcilment les relacions entre aliments i biomarcadors tant a nivell d'usuari com computacional. De la mateixa manera, FOBI es pot utilitzar per a diferents propòsits, des de fer consultes simples fins a consultes computacionals complexes simultàniament mitjançant tota la informació emmagatzemada a l'ontologia.

Les ontologies faciliten moltes aplicacions pràctiques en el camp de la bioinformàtica, com anotar entitats dels camps de la metabolòmica i la nutrició, dur a terme diferents anàlisis d'enriquiment, realitzar anàlisis de similitud semàntica o fins i tot descobrir noves relacions entre entitats (Hoehndorf et al., 2015). En el cas de l'ontologia FOBI, l'aplicació més clara consisteix en l'anotació d'aliments i biomarcadors dietètics. Això facilita enormement la comparabilitat i la interoperabilitat entre estudis i projectes en el camp de la nutrimetabolòmica. FOBI suposa una millora per la investigació en el camp de la nutrimetabolòmica gràcies a la definició detallada de les associacions entre diferents tipus d'aliments i els seus metabòlits associats.

La informació de FOBI també pot servir per facilitar els dissenys d'estudi, des de la generació d'hipòtesis (per exemple, els metabòlits esperats que es produeixen després d'una intervenció dietètica) fins al disseny experimental (per exemple, l'optimització de mètodes de metabolòmica dirigits focalitzats en metabòlits d'interès). A més, una de les principals aplicacions que tindrà FOBI és que proporcionarà la capacitat de dur a terme anàlisis de significació biològica en estudis de nutrimetabolòmica, cosa que fins ara no era possible degut a la manca d'ontologies específiques per aquest camp i a la manca d'eines per fer-ho possible.

L'ontologia FOBI consisteix en l'**article 1**, presentat a l'apartat de resultats.

Com s'ha comentat breument en la introducció, un dels principals motius que limita l'ús de les ontologies són els coneixements de programació necessaris per poder interactuar amb elles, extraure informació i utilitzar-les en general. FOBI com tota ontologia, també es troba amb aquesta limitació. Per aquest motiu, un cop desenvolupada l'ontologia, ens vam plantejar com oferir als usuaris de la comunitat nutrimetabolòmica la possibilitat d'utilitzar FOBI d'una manera fàcil i ràpida. Això porta a l'**objectiu específic 1a** i a l'**objectiu específic 1b**.

Per a oferir als usuaris la possibilitat de dur a terme algunes de les aplicacions de l'ontologia FOBI mencionades anteriorment, vam desenvolupar el marc de treball fobitools, enfocat específicament al nou concepte d'anàlisi d'enriquiment alimentari i als usuaris sense coneixements extensos de programació.

El marc de treball fobitools consisteix en un paquet de Bioconductor (**objectiu específic 1a**) i una aplicació web (**objectiu específic 1b**) que tenen com a objectiu facilitar i estendre l'ús de l'ontologia FOBI a la comunitat científica. Aquestes dues eines segueixen la filosofia "*user-friendly*" (fàcil d'utilitzar) i permeten, entre altres, la realització d'anàlisis d'enriquiment alimentari, l'exploració de FOBI mitjançant gràfics de xarxa estàtics i dinàmics i l'anotació automàtica de dades dietètiques de text lliure mitjançant algorismes de mineria de text.

Aquest és un projecte de codi obert, pel que la comunitat científica pot utilitzar-lo i contribuir-hi fàcilment. Així doncs, marc de treball fobitools permet als investigadors utilitzar l'ontologia FOBI d'una manera ràpida i senzilla, ja sigui des de la línia de comandes d'R (usuaris amb coneixements de programació) o des de l'aplicació web (on els usuaris no necessiten tenir nocions de programació).

Aquesta eina introdueix per primera vegada el concepte d'anàlisi d'enriquiment alimentari, fent possible que els usuaris puguin explorar aliments o grups d'aliments enriquits a partir de llistes de metabòlits obtingudes en estudis de nutrimetabolòmica.

Si bé és cert que aquesta eina suposa una millora substancial per a

la interpretació dels resultats en estudis de nutrimetabolòmica, també caben a destacar algunes limitacions. Actualment, degut al fet que l'ontologia FOBI es troba en la seva primera versió, l'anàlisi amb l'eina fobitools pot estar limitat a un nombre no molt elevat d'aliments i metabòlits en comparació amb altres ontologies i bases de dades. Així doncs, els esforços futurs aniran dirigits a expandir l'ontologia FOBI, donant lloc a un augment del nombre de metabòlits, aliments i relacions metabòlit-aliment. D'altra banda, el marc de treball fobitools proporciona la metodologia per interactuar amb l'ontologia FOBI independentment de la quantitat d'informació que contingui aquesta. Per tant, les futures millores a FOBI tindran un impacte directe en l'eina fobitools, augmentant la seva utilitat i permetent realitzar anàlisis més precises, completes i robustes. Pel que fa a les millores de programari futures de l'eina fobitools, aquestes aniran dirigides principalment a la implementació de nous mètodes d'anàlisi d'enriquiment.

El marc de treball fobitools consisteix en l'**article 3**, presentat a l'apartat de resultats. Malauradament, tot i que aquesta eina ja està en funcionament i està sent utilitzada per diferents usuaris, no s'ha utilitzat en els articles 4 i 5, ja que l'eina fobitools i aquests articles es van desenvolupar simultàniament.

El segon gran objectiu d'aquesta tesi va ser el desenvolupament de mètodes i eines que contribuïssin a la millora del procés d'anàlisi de dades en estudis de metabolòmica (i per tant, de nutrimetabolòmica) per ajudar a millorar el procés de descobriment de biomarcadors, entre d'altres (**objectiu específic 2**).

Sovint, una de les principals aplicacions de la metabolòmica és la caracterització de noves dianes terapèutiques en els camps de la salut humana i la medicina personalitzada (Wishart, 2016). Per aquest motiu, durant l'última dècada han aparegut diverses eines que contribueixen a l'anàlisi d'aquestes dades tan complexes (Stanstrup et al., 2019). No obstant, moltes limiten l'anàlisi a un nombre de mètodes estadístics reduït (Gardinassi et al., 2017), fet que obliga els investigadors a utilitzar una extensa bateria d'eines diferents per satisfer totes les necessitats de l'anàlisi.

Amb l'objectiu de contribuir a l'extensió dels mètodes i eines disponibles per a l'anàlisi de dades de metabolòmica, l'eina web POMAShiny desenvolupada en el context d'aquesta tesi, proporciona un flux de treball complet i estructurat que cobreix el els processos de preprocessament, exploració i l'anàlisi estadística d'aquestes dades, amb la intenció de ser una eina complementària, fàcil d'utilitzar i intuïtiva que aborda alguns dels problemes que no estan coberts per altres eines (**objectiu específic 2a**). Aquest flux de treball s'integra en una atractiva interfície gràfica que proporciona diversos mètodes per a l'anàlisi de dades, inclosos mètodes estadístics univariants, mètodes multivariants i de reducció de la dimensió, mètodes de selecció de variables, aproximacions d'anàlisi de regressió regularitzada, algorismes de classificació d'aprenentatge automàtic, estratègies de models de predicció i diverses opcions de visualització interactiva d'alta qualitat.

Aquesta nova eina es basa en el paquet de Bioconductor POMA (**objectiu específic 2**) i integra molts dels mètodes més utilitzats per a l'anàlisi de dades de metabolòmica (Gardinassi et al., 2017; Stanstrup et al., 2019), a més d'incorporar noves alternatives útils i potents. POMAShiny permet als usuaris dur a terme anàlisis de dades integrats en un entorn web interactiu, intuïtiu i ben documentat, fent que el procés d'anàlisi de dades sigui més accessible per a un ampli ventall d'investigadors no molt familiaritzats amb els camps de la programació i l'estadística.

L'existència conjunta tant del paquet de POMA (completament integrat dins l'entorn de Bioconductor) com de la interfície web POMAShiny significa un enorme increment de la reproductibilitat de l'eina, contribuint també a la reutilització de mètodes existents als entorns R i Bioconductor (Gentleman et al., 2004; R Core Team, 2019), a més de permetre una fàcil extensió, integració i interoperabilitat amb altres fluxos de treball, com la iniciativa RforMassSpectrometry (RforMassSpectrometry.org), que proporciona les estructures de dades utilitzades pel paquet POMA. Per tant, els usuaris poden realitzar el processament de dades d'espectrometria i altres operacions rutinàries de flux de treball de MS utilitzant els paquets de la iniciativa RforMassSpectrometry i després migrar fàcilment a POMA/POMAShiny per realitzar l'anàlisi estadística.

L'eina web POMAShiny consisteix en l'**article 2**, presentat a l'apartat de resultats.

Amb l'assoliment dels objectius específics 1 i 2, aquesta tesi presenta un conjunt d'eines i recursos que permeten als investigadors dur a terme les anàlisis estadístiques dels estudis utilitzant una gran varietat de mètodes, així com dotar els resultats d'aquestes anàlisis de la seva significació biològica en un context de nutrimetabolòmica. A més, es proporcionen interfícies gràfiques per totes les eines presentades en aquest treball, facilitant la seva utilització independentment dels coneixements de programació dels usuaris.

Pel que fa a l'**objectiu específic 3**, en aquesta tesi també es presenten dos estudis de nutrimetabolòmica on s'han aplicat gran varietat de conceptes d'anàlisi de dades així com diferents metodologies estadístiques proporcionades per les eines i recursos discutits anteriorment.

En el primer d'aquests estudis, l'objectiu era la identificació de metabòlits o grups de metabòlits associats a l'índex de dieta saludable AHEI-2010 (Chiuve et al., 2012) (**objectiu específic 3a**).

En aquest context, tot i que nombrosos estudis han investigat prèviament l'associació entre els metabòlits circulants i el consum de determinats grups d'aliments, només alguns d'ells s'han centrat en la identificació de biomarcadors de patrons dietètics saludables. En aquest sentit, diversos estudis han abordat recentment la identificació de potencials marcadors metabolòmics de l'AHEI-2010 en mostres de sèrum i plasma de diferents poblacions (Akbaraly et al., 2018; Bagheri et al., 2020; McCullough et al., 2019; Walker et al., 2020).

En aquest estudi, pretenïem identificar metabòlits en orina associats a l'índex AHEI-2010, que podrien servir com a biomarcadors d'adherència a patrons dietètics saludables.

En aquest estudi es va utilitzar l'eina POMA (**article 2**) per dur a terme tota la part de preprocessament, anàlisi exploratòria i una part

de l'anàlisi estadística.

Mitjançant una anàlisi de regressió regularitzada LASSO i una anàlisi univariant *limma*, es va poder identificar la robusta associació positiva entre l'índex AHEI-2010 i els nivells d'enterolactona glucurònid en orina al llarg del temps. L'enterolactona és el principal metabòlit derivat dels la microbiota dels lignans dietètics, una subclasse de polifenols àmpliament distribuïda en aliments vegetals com fruites, verdures, cereals integrals, llegums i fruits secs (Senizza et al., 2020). Es coneix que els lignans tenen diferents propietats antiinflamatòries i antioxidants; diversos estudis epidemiològics han demostrat que les altes concentracions circulants d'enterolactones s'associen a un menor risc de malalties cardiovasculars (Rienks et al., 2017), diversos càncers (Micek et al., 2021), trastorns neurodegeneratius (Reddy et al., 2020), entre d'altres. Per tant, es podria considerar aquest metabòlit com un biomarcador fiable i robust per avaluar com de saludable és una dieta.

La regressió LASSO també va permetre identificar una associació reproduïble del metabòlit 5-(hidroximetil-2-furoil) glicina amb puntuacions altes de l'AHEI-2010 al llarg del temps, així com una associació positiva amb el metabòlit 2-furoilglicina en un dels tres temps estudiats. Els furans s'han proposat prèviament com a biomarcadors de diferents productes alimentaris processats amb calor, com ara fruits secs (Prior et al., 2006) i cafè (Heinzmann et al., 2015).

Addicionalment, malgrat que no es corrobora en els tres períodes de temps estudiats, també es va trobar una forta associació negativa entre la L-carnitina i l'AHEI-2010, juntament amb una associació negativa també de la carnosina, reflectint un efecte perjudicial del consum de carn vermella i processada sobre la salut, i una associació negativa dels metabòlits derivats del tabac. Per contra, diversos metabòlits que reflecteixen la ingesta de peix i marisc van mostrar associacions positives amb l'AHEI-2010. A més, també van trobar associacions consistents entre l'índex AHEI-2010 i altres biomarcadors candidats a la ingesta d'aliments definits a l'ontologia FOBI (Castellano-Escuder et al., 2020) (**article 1**), però només en un dels tres punts de temps investigats. En particular, es va trobar una associació positiva amb diversos metabòlits del vi negre (e.g., resveratrols) i metabòlits relacionats amb la ingesta de cítrics, oli d'oliva i fruits vermells.

En conclusió, aquests resultats mostren que una sèrie de metabòlits relacionats amb la ingesta d'aliments estan fortament associats a l'adhesió a hàbits dietètics saludables avaluats amb l'índex AHEI-2010 (**article 4**).

Finalment, el segon estudi d'aplicació presentat en aquesta tesi tenia l'objectiu d'identificar metabòlits o grups de metabòlits associats al risc de malaltia o a l'estat de salut (**objectiu específic 3b**).

La participació de factors d'estil de vida modificables en la patogènesi del deteriorament cognitiu (CD) i de la demència està ben acceptada avui en dia (Peters et al., 2019). En particular, s'ha identificat la dieta com un factor fonamental en el manteniment d'una funció cerebral adequada (Flanagan et al., 2020). De fet, molts components de la dieta poden modular els mecanismes moleculars que contribueixen al CD, inclosos l'estrès oxidatiu, la neuroinflamació i la disfunció vascular (Vauzour et al., 2017).

L'objectiu d'aquest estudi era desxifrar el paper de la dieta en el desenvolupament del deteriorament cognitiu a través d'un enfocament de metabolòmica dirigida a gran escala.

En aquest estudi també es va utilitzar l'eina POMA (**article 2**) per dur a terme tota la part de preprocessament i anàlisi exploratòria, incloent la imputació de valors faltants, la normalització de les dades i l'eliminació d'*outliers* (o mostres atípiques).

Per identificar metabòlits en sèrum associats al CD es va utilitzar una regressió logística condicionada de LASSO combinada amb la tècnica *bootstrap*, per tal d'estabilitzar els resultats i la variabilitat intrínseca del mètode LASSO.

Molts dels metabòlits identificats en les dues poblacions d'estudi (veure article 5), inclosos derivats de polifenols i aminoàcids aromàtics, suggereixen una estreta interacció entre dieta, microbiota intestinal i CD. La microbiota intestinal s'ha reconegut com un factor important en la salut i la funció cognitiva, ja que molts metabòlits derivats de la microbiota tenen propietats metabòliques i de senyalització essencials

que poden modular la funció cerebral (Needham et al., 2020; Parker et al., 2020). Per tant, s'ha plantejat la hipòtesi que la microbiota intestinal i les molècules que produeixen podrien formar part d'una xarxa que uneix la dieta amb la funció cognitiva a través de l'"eix intestí-cervell" (Collins et al., 2012).

En les dues poblacions d'estudi es va observar una associació inversa entre diversos àcids fenòlics i altres metabòlits derivats de plantes amb les probabilitats de patir CD, proporcionant més evidència sobre l'efecte protector del consum d'aliments rics en polifenols (és a dir, fruites i verdures) contra el CD (Mottaghi et al., 2018).

En línia amb l'estudi anterior del projecte D-CogPlast, realitzat utilitzant una aproximació de metabolòmica no dirigida (Low et al., 2019), també vam observar una associació negativa entre la 3-metilxantina (un metabòlit derivat de la teobromina present al cacau) i el CD. Els nivells de 3-metilxantina estaven altament correlacionats amb la teobromina, que també es va associar negativament amb CD en ambdues poblacions, reforçant l'efecte protector del consum de cacau contra el CD (Moreira et al., 2016). A més, els resultats d'aquest treball també suggereixen que la ingesta de cafè està associada negativament amb el risc de CD (efecte protector). Com s'ha explicat anteriorment, la 2-furoilglicina és un biomarcador del consum de cafè, que va resultar estar associada a probabilitats més baixes de CD en les dues poblacions d'estudi.

A més d'aquestes associacions potencialment protectores entre aliments rics en polifenols, cacau, cafè, vi negre i CD, aquests resultats també apunten a una associació nociva de certs components dietètics sobre la funció cognitiva, incloent per exemple els edulcorants artificials.

En conclusió, aquests resultats apunten a una associació protectora principalment dels metabòlits derivats de la microbiota, fruites i verdures i cafè amb el deteriorament cognitiu, mentre que altres metabòlits relacionats amb aliments poc saludables, com les begudes ensucrades, poden tenir efectes nocius sobre la cognició (**article 5**).

Chapter 9

Conclusions

En aquesta tesi s'han estudiat els aspectes bàsics del descobriment de biomarcadors dietètics mitjançant la metabolòmica, així com les bases de l'anàlisi de significació biològica en estudis de nutrimetabolòmica. A continuació, es destaquen les principals contribucions a aquestes àrees. En concret, s'ha demostrat que:

- FOBI és la primera ontologia que integra dades nutricionals i metabolòmiques mitjançant un llenguatge comú estandarditzat per definir les relacions entre els aliments i els seus metabolits associats. Actualment, FOBI té un total de 1197 termes, 11 classes químiques de primer nivell, 13 classes alimentàries de primer nivell i 4 tipus de relacions diferents entre nodes.

- FOBI permet als experts anotar i analitzar dades nutricionals i metabolòmiques d'una manera robusta, fent que els resultats siguin comparables entre estudis d'aquests camps. El desenvolupament de FOBI conduirà a una millora de la interoperabilitat de les dades nutricionals i nutrimetabolòmiques, fent així que els conjunts de dades generats en aquests estudis siguin plenament compatibles amb la filosofia FAIR.

- POMAShiny és una eina web fàcil d'utilitzar que proporciona un flux de treball integrat per l'anàlisi de dades de metabolòmica amb un ampli ventall de possibilitats, tant per al preprocessament

de dades com per l'anàlisi estadística, inclosos mètodes de detecció d'*outliers*, operacions per l'anàlisi exploratòria, informes descarregables i deferents metodologies estadístiques, des d'enfocaments més senzills com ara l'analisi univariant fins a mètodes més complexos com ara la regressió regularitzada i algorismes d'aprenentatge automàtic.

- L'entorn de treball fobitools consisteix en un paquet de Bioconductor i una aplicació web que proporcionen una infraestructura per interactuar amb l'ontologia FOBI d'una manera senzilla. Aquest entorn de treball permet als investigadors realitzar anàlisis d'enriquiment en estudis de nutrimetabolòmica, entre altres operacions útils, com ara la visualització interactiva en xarxa de FOBI i l'anotació automàtica de dades dietètiques de text lliure mitjançant la informació de FOBI.

- El metaboloma alimentari en orina està fortament associat a l'adherència a hàbits dietètics saludables avaluats a través de l'AHEI-2010. Molts dels metabòlits associats descoberts van ser compostos derivats de la microbiota, inclosos els enterolignans, les urolitines i els àcids fenòlics, donant suport a un rol important de la microbiota intestinal en la interacció entre la dieta i la salut.

- Els metabòlits derivats de la dieta i de la microbiota poden tenir un paper important en el desenvolupament del deteriorament cognitiu tardà. Els metabòlits que reflecteixen el consum d'aliments rics en polifenols, cacau, cafè, bolets i vi negre van mostrar un efecte protector sobre el deteriorament cognitiu, mentre que altres metabòlits de la dieta relacionats amb components dietètics poc saludables van mostrar efectes nocius sobre la cognició.

Les eines desenvolupades en aquesta tesi s'han implementat en dos paquets d'R/Bioconductor -POMA i fobitools- d'accés obert, disponibles a Bioconductor, i la seva difusió ha estat facilitada per dues interfícies gràfiques -POMAShiny i fobitoolsGUI- disponibles a

GitHub.

Part V

References

Akbaraly, T., Würtz, P., Singh-Manoux, A., Shipley, M. J., Haapakoski, R., Lehto, M., ... others. (2018). Association of circulating metabolites with healthy diet and risk of cardiovascular disease: Analysis of two cohort studies. *Scientific Reports*, *8*(1), 1–14.

Andersen, M.-B. S., Kristensen, M., Manach, C., Pujos-Guillot, E., Poulsen, S. K., Larsen, T. M., ... Dragsted, L. (2014). Discovery and validation of urinary exposure markers for different plant foods by untargeted metabolomics. *Analytical and Bioanalytical Chemistry*, *406*(7), 1829–1844.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... others. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29.

Bagheri, M., Willett, W., Townsend, M. K., Kraft, P., Ivey, K. L., Rimm, E. B., ... others. (2020). A lipid-related metabolomic pattern of diet quality. *The American Journal of Clinical Nutrition*, *112*(6), 1613–1630.

Bekri, S. (2016). The role of metabolomics in precision medicine. *Expert Review of Precision Medicine and Drug Development*, *1*(6), 517–532.

Boccard, J., & Rutledge, D. N. (2013). A consensus orthogonal partial least squares discriminant analysis (opls-da) strategy for multiblock omics data fusion. *Analytica Chimica Acta*, *769*, 30–39.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Castellano-Escuder, P., González-Domínguez, R., Wishart, D. S., Andrés-Lacueva, C., & Sánchez-Pla, A. (2020). FOBI: An ontology to represent food intake data and associate it with metabolomic data. *Database*, *2020*.

Cevallos-Cevallos, J. M., Reyes-De-Corcuera, J. I., Etxeberria, E., Danyluk, M. D., & Rodrick, G. E. (2009). Metabolomic analysis in food science: A review. *Trends in Food Science & Technology*, *20*(11-12), 557–566.

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2020). *Shiny: Web application framework for r*. Retrieved from `https://CRAN.R-project.org/package=shiny`

Chiuve, S. E., Fung, T. T., Rimm, E. B., Hu, F. B., McCullough, M. L., Wang, M., ... Willett, W. C. (2012). Alternative dietary indices both strongly predict risk of chronic disease. *The Journal of Nutrition*, *142*(6), 1009–1018.

Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., ... Xia, J. (2018). MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. *Nucleic Acids Research*, *46*(W1), W486–W494.

Collins, S. M., Surette, M., & Bercik, P. (2012). The interplay between the intestinal microbiota and the brain. *Nature Reviews Microbiology*, *10*(11), 735–742.

Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., ... others. (2015). Pathway and network analysis of cancer genomes. *Nature Methods*, *12*(7), 615.

Davidson, R. L., Weber, R. J., Liu, H., Sharma-Oates, A., & Viant, M. R. (2016). Galaxy-m: A galaxy workflow for processing and analyzing direct infusion and liquid chromatography mass spectrometry-based metabolomics data. *GigaScience*, *5*(1), s13742–016.

Degtyarenko, K., De Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., ... Ashburner, M. (2007). ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, *36*(suppl_1), D344–D350.

Dooley, D. M., Griffiths, E. J., Gosal, G. S., Buttigieg, P. L., Hoehndorf, R., Lange, M. C., ... Hsiao, W. W. (2018). FoodOn: A harmonized food ontology to increase global food traceability, quality control and data integration. *Npj Science of Food*, *2*(1), 1–10.

Emwas, A.-H. M. (2015). The strengths and weaknesses of nmr spectroscopy and mass spectrometry with particular focus on metabolomics research. In *Metabonomics* (pp. 161–193). Springer.

Erban, A., Fehrle, I., Martinez-Seidel, F., Brigante, F., Más, A. L., Baroni, V., ... Kopka, J. (2019). Discovery of food identity markers by metabolomics and machine learning technology. *Scientific Reports*, *9*(1), 1–19.

Fardet, A., Llorach, R., Orsoni, A., Martin, J.-F., Pujos-Guillot, E.,

Lapierre, C., & Scalbert, A. (2008). Metabolomics provide new insight on the metabolism of dietary phytochemicals in rats. *The Journal of Nutrition*, *138*(7), 1282–1287.

Flanagan, E., Lamport, D., Brennan, L., Burnet, P., Calabrese, V., Cunnane, S. C., ... others. (2020). Nutrition and the ageing brain: Moving towards clinical applications. *Ageing Research Reviews*, 101079.

Freedman, L. S., Schatzkin, A., Midthune, D., & Kipnis, V. (2011). Dealing with dietary measurement error in nutritional cohort studies. *Journal of the National Cancer Institute*, *103*(14), 1086–1092.

Friedman, J., Hastie, T., Tibshirani, R., & others. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics New York.

Færgestad, E. M., Langsrud, Ø., Høy, M., Hollung, K., Sæbø, S., Liland, K. H., ... Martens, H. (2009). 4.08 - analysis of megavariate data in functional genomics. In S. D. Brown, R. Tauler, & B. Walczak (Eds.), *Comprehensive chemometrics* (pp. 221–278). Oxford: Elsevier. `http://doi.org/https://doi.org/10.1016/B978-044452701-1.00011-9`

Garden, L., Clark, H., Whybrow, S., & Stubbs, R. J. (2018). Is misreporting of dietary intake by weighed food records or 24-hour recalls food specific? *European Journal of Clinical Nutrition*, *72*(7), 1026–1034.

Gardinassi, L. G., Xia, J., Safo, S. E., & Li, S. (2017). Bioinformatics tools for the interpretation of metabolomics data. *Current Pharmacology Reports*, *3*(6), 374–383.

Gatto, L., & Lilley, K. (2012). MSnbase - an r/bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics*, *28*, 288–289.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., ... others. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, *5*(10), R80.

Giacomoni, F., Le Corguillé, G., Monsoor, M., Landi, M., Pericard, P., Pétéra, M., ... others. (2015). Workflow4Metabolomics: A

collaborative research infrastructure for computational metabolomics. *Bioinformatics*, *31*(9), 1493–1495.

Goeman, J. J., & Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: Methodological issues. *Bioinformatics*, *23*(8), 980–987.

Gregori, J., Sanchez, A., & Villanueva, J. (2020a). *MsmsEDA: Exploratory data analysis of lc-ms/ms data by spectral counts*.

Gregori, J., Sanchez, A., & Villanueva, J. (2020b). *MsmsTests: LC-ms/ms differential expression tests*.

Guglielmetti, S., Bernardi, S., Del Bo, C., Cherubini, A., Porrini, M., Gargari, G., ... others. (2020). Effect of a polyphenol-rich dietary pattern on intestinal permeability and gut and blood microbiomics in older subjects: Study protocol of the maple randomised controlled trial. *BMC Geriatrics*, *20*(1), 1–10.

Heinzmann, S. S., Holmes, E., Kochhar, S., Nicholson, J. K., & Schmitt-Kopplin, P. (2015). 2-furoylglycine as a candidate biomarker of coffee consumption. *Journal of Agricultural and Food Chemistry*, *63*(38), 8615–8621.

Hoehndorf, R., Schofield, P. N., & Gkoutos, G. V. (2015). The role of ontologies in biological and biomedical research: A functional perspective. *Briefings in Bioinformatics*, *16*(6), 1069–1080.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., ... Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, *12*(2), 115–121. Retrieved from `http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html`

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Comput Biol*, *8*(2), e1002375.

Kramer, F., & Beißbarth, T. (2017). Working with ontologies. In *Bioinformatics* (pp. 123–135). Springer.

Leung, C. W., Ding, E. L., Catalano, P. J., Villamor, E., Rimm, E. B., & Willett, W. C. (2012). Dietary intake and dietary quality of low-income adults in the supplemental nutrition assistance program. *The American Journal of Clinical Nutrition*, *96*(5), 977–988.

Leung, C. W., Epel, E. S., Ritchie, L. D., Crawford, P. B., & Laraia, B. A. (2014). Food insecurity is inversely associated with diet quality of lower-income adults. *Journal of the Academy of Nutrition and Dietetics*, *114*(12), 1943–1953.

Lê Cao, K.-A., Boitard, S., & Besse, P. (2011). Sparse pls discriminant analysis: Biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, *12*(1), 253.

Low, D. Y., Lefévre-Arbogast, S., González-Domínguez, R., Urpi-Sarda, M., Micheau, P., Petera, M., ... others. (2019). Diet-related metabolites associated with cognitive decline revealed by untargeted metabolomics in a prospective cohort. *Molecular Nutrition & Food Research*, *63*(18), 1900177.

Marco-Ramell, A., Palau-Rodriguez, M., Alay, A., Tulipani, S., Urpi-Sarda, M., Sanchez-Pla, A., & Andres-Lacueva, C. (2018). Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics data. *BMC Bioinformatics*, *19*(1), 1.

Maruvada, P., Lampe, J. W., Wishart, D. S., Barupal, D., Chester, D. N., Dodd, D., ... others. (2020). Perspective: Dietary biomarkers of intake and exposure—exploration with omics approaches. *Advances in Nutrition*, *11*(2), 200–215.

McCullough, M. L., Maliniak, M. L., Stevens, V. L., Carter, B. D., Hodge, R. A., & Wang, Y. (2019). Metabolomic markers of healthy dietary patterns in us postmenopausal women. *The American Journal of Clinical Nutrition*, *109*(5), 1439–1451.

Micek, A., Godos, J., Brzostek, T., Gniadek, A., Favari, C., Mena, P., ... Grosso, G. (2021). Dietary phytoestrogens and biomarkers of their intake in relation to cancer survival and recurrence: A comprehensive systematic review with meta-analysis. *Nutrition Reviews*, *79*(1), 42–65.

Moreira, A., Diógenes, M. J., Mendonca, A. de, Lunet, N., & Barros,

H. (2016). Chocolate consumption is associated with a lower risk of cognitive decline. *Journal of Alzheimer's Disease*, *53*(1), 85–93.

Mottaghi, T., Amirabdollahian, F., & Haghighatdoost, F. (2018). Fruit and vegetable intake and cognitive impairment: A systematic review and meta-analysis of observational studies. *European Journal of Clinical Nutrition*, *72*(10), 1336–1344.

Mulligan, A. A., Luben, R. N., Bhaniani, A., Parry-Smith, D. J., O'Connor, L., Khawaja, A. P., ... Khaw, K.-T. (2014). A new tool for converting food frequency questionnaire data into nutrient and food group values: FETA research methods and availability. *BMJ Open*, *4*(3). http://doi.org/10.1136/bmjopen-2013-004503

Needham, B. D., Kaddurah-Daouk, R., & Mazmanian, S. K. (2020). Gut microbial molecules in behavioural and neurodegenerative conditions. *Nature Reviews Neuroscience*, *21*(12), 717–731.

Neveu, V., Moussy, A., Rouaix, H., Wedekind, R., Pon, A., Knox, C., ... Scalbert, A. (2016). Exposome-explorer: A manually-curated database on biomarkers of exposure to dietary and environmental factors. *Nucleic Acids Research*, gkw980.

Noy, N. F., McGuinness, D. L., & others. (2001). Ontology development 101: A guide to creating your first ontology. Stanford knowledge systems laboratory technical report KSL-01-05 and ....

Park, Y., Dodd, K. W., Kipnis, V., Thompson, F. E., Potischman, N., Schoeller, D. A., ... others. (2018). Comparison of self-reported dietary intakes from the automated self-administered 24-h recall, 4-d food records, and food-frequency questionnaires against recovery biomarkers. *The American Journal of Clinical Nutrition*, *107*(1), 80–93.

Parker, A., Fonseca, S., & Carding, S. R. (2020). Gut microbes and metabolites as modulators of blood-brain barrier integrity and brain health. *Gut Microbes*, *11*(2), 135–157.

Peters, R., Booth, A., Rockwood, K., Peters, J., D'Este, C., & Anstey, K. J. (2019). Combining modifiable risk factors and risk of dementia: A systematic review and meta-analysis. *BMJ Open*, *9*(1), e022846.

Praticò, G., Gao, Q., Scalbert, A., Vergères, G., Kolehmainen, M.,

Manach, C., ... others. (2018). Guidelines for biomarker of food intake reviews (bfirev): How to conduct an extensive literature search for biomarker of food intake discovery. *Genes & Nutrition*, *13*(1), 3.

Prior, R. L., Wu, X., & Gu, L. (2006). Identification and urinary excretion of metabolites of 5-(hydroxymethyl)-2-furfural in human subjects following consumption of dried plums or dried plum juice. *Journal of Agricultural and Food Chemistry*, *54*(10), 3744–3749.

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from `https://www.R-project.org/`

Reddy, V. P., Aryal, P., Robinson, S., Rafiu, R., Obrenovich, M., & Perry, G. (2020). Polyphenols in alzheimer's disease and in the gut–brain axis. *Microorganisms*, *8*(2), 199.

Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., ... others. (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler, gsea, cytoscape and enrichmentmap. *Nature Protocols*, *14*(2), 482–517.

Rienks, J., Barbaresko, J., & Nöthlings, U. (2017). Association of polyphenol biomarkers with cardiovascular disease and mortality risk: A systematic review and meta-analysis of observational studies. *Nutrients*, *9*(4), 415.

Roberts, L. D., Souza, A. L., Gerszten, R. E., & Clish, C. B. (2012). Targeted metabolomics. *Current Protocols in Molecular Biology*, *98*(1), 30–2.

Rothwell, J. A., Perez-Jimenez, J., Neveu, V., Medina-Remon, A., M'Hiri, N., García-Lobato, P., ... others. (2013). Phenol-explorer 3.0: A major update of the phenol-explorer database to incorporate data on the effects of food processing on polyphenol content. *Database*, *2013*.

Rubin, D. L., Shah, N. H., & Noy, N. F. (2008). Biomedical ontologies: A functional perspective. *Briefings in Bioinformatics*, *9*(1), 75–90.

Sansone, S.-A., Fan, T., Goodacre, R., Griffin, J. L., Hardy, N. W., Kaddurah-Daouk, R., ... others. (2007). The metabolomics standards initiative. *Nature Biotechnology*, *25*(8), 846–849.

Sansone, S.-A., Schober, D., Atherton, H. J., Fiehn, O., Jenkins, H., Rocca-Serra, P., ... others. (2007). Metabolomics standards initiative: Ontology working group work in progress. *Metabolomics*, *3*(3), 249–256.

Scalbert, A., Brennan, L., Manach, C., Andres-Lacueva, C., Dragsted, L. O., Draper, J., ... Wishart, D. S. (2014). The food metabolome: A window over dietary exposure. *The American Journal of Clinical Nutrition*, *99*(6), 1286–1308.

Schlegel, D., Ruttenberg, A., & Elkin, P. (2015). Ontologies in metabolomics. *Metabolomics*, *5*, e137.

Schrimpe-Rutledge, A. C., Codreanu, S. G., Sherrod, S. D., & McLean, J. A. (2016). Untargeted metabolomics strategies—challenges and emerging directions. *Journal of the American Society for Mass Spectrometry*, *27*(12), 1897–1905.

Senizza, A., Rocchetti, G., Mosele, J. I., Patrone, V., Callegari, M. L., Morelli, L., & Lucini, L. (2020). Lignans and gut microbiota: An interplay revealing potential health implications. *Molecules*, *25*(23), 5709.

Stanstrup, J., Broeckling, C. D., Helmus, R., Hoffmann, N., Mathé, E., Naake, T., ... others. (2019). The metaRbolomics toolbox in bioconductor and beyond. *Metabolites*, *9*(10), 200.

Strimbu, K., & Tavel, J. A. (2010). What are biomarkers? *Current Opinion in HIV and AIDS*, *5*(6), 463.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ... others. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545–15550.

Tautenhahn, R., Patti, G. J., Rinehart, D., & Siuzdak, G. (2012). XCMS online: A web-based platform to process untargeted metabolomic data. *Analytical Chemistry*, *84*(11), 5035–5039.

Tebani, A., & Bekri, S. (2019). Paving the way to precision nutrition through metabolomics. *Frontiers in Nutrition*, *6*, 41.

Thompson, F. E., & Subar, A. F. (2013). Chapter 1 - dietary assessment methodology. In A. M. Coulston, C. J. Boushey, & M. G. Ferruzzi (Eds.), *Nutrition in the prevention and treatment of disease (third edition)* (Third Edition, pp. 5–46). Academic Press. `http://doi.org/https://doi.org/10.1016/B978-0-12-391884-0.00001-9`

Trepanowski, J. F., & Ioannidis, J. P. (2018). Perspective: Limiting dependence on nonrandomized studies and improving randomized trials in human nutrition research: Why and how. *Advances in Nutrition*, *9*(4), 367–377.

Ulaszewska, M. M., Weinert, C. H., Trimigno, A., Portmann, R., Andres Lacueva, C., Badertscher, R., ... others. (2019). Nutrimetabolomics: An integrative action for metabolomic analyses in human nutritional studies. *Molecular Nutrition & Food Research*, *63*(1), 1800384.

Vauzour, D., Camprubi-Robles, M., Miquel-Kergoat, S., Andres-Lacueva, C., Bánáti, D., Barberger-Gateau, P., ... others. (2017). Nutrition for the ageing brain: Towards evidence for an optimal diet. *Ageing Research Reviews*, *35*, 222–240.

Vitali, F., Lombardo, R., Rivero, D., Mattivi, F., Franceschi, P., Bordoni, A., ... others. (2018). ONS: An ontology for a standardized description of interventions and observational studies in nutrition. *Genes & Nutrition*, *13*(1), 12.

Walker, M. E., Song, R. J., Xu, X., Gerszten, R. E., Ngo, D., Clish, C. B., ... others. (2020). Proteomic and metabolomic correlates of healthy dietary patterns: The framingham heart study. *Nutrients*, *12*(5), 1476.

Wang, D. D., Leung, C. W., Li, Y., Ding, E. L., Chiuve, S. E., Hu, F. B., & Willett, W. C. (2014). Trends in dietary quality among adults in the united states, 1999 through 2010. *JAMA Internal Medicine*, *174*(10), 1587–1595.

Wang, H., Lei, M., Chen, Y., Li, M., & Zou, L. (2019). Intelligent identification of maceral components of coal based on image segmentation and classification. *Applied Sciences*, *9*(16), 3245.

Wickham, H. (2011). Testthat: Get started with testing. *The R*

Journal, *3*, 5–10. Retrieved from https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf

Wickham, H. (2015). *R packages* (1st ed.). O'Reilly Media, Inc.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. http://doi.org/10.21105/joss.01686

Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., & Mons, B. (2016). The fair guiding principles for scientific data management and stewardship. Sci data. 2016; 3: 160018. Epub 2016/03/16. doi: 10.1038/sdata. 2016.18. PubMed PMID: 26978244.

Wishart, D. S. (2008). Metabolomics: Applications to food science and nutrition research. *Trends in Food Science & Technology*, *19*(9), 482–493.

Wishart, D. S. (2016). Emerging applications of metabolomics in drug discovery and precision medicine. *Nature Reviews Drug Discovery*, *15*(7), 473.

Xia, J., & Wishart, D. S. (2010). MSEA: A web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Research*, *38*(suppl_2), W71–W77.

Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., ... others. (2015). Personalized nutrition by prediction of glycemic responses. *Cell*, *163*(5), 1079–1094.

Part VI

Appendices

Appendix A

A.1 Thesis publications

This section includes those scientific papers done in the context of this thesis that have already been accepted and published. Those articles submitted to scientific journals but not yet accepted cannot be attached here for reasons of confidentiality and plagiarism.

A.1.1 Paper 1: Food-Biomarker Ontology

Database, 2020, 1–8
doi: 10.1093/databa/baaa033
Original article

Original article

FOBI: an ontology to represent food intake data and associate it with metabolomic data

Pol Castellano-Escuder[1,2,3], Raúl González-Domínguez[1,3], David S. Wishart[4], Cristina Andrés-Lacueva[1,3] and Alex Sánchez-Pla[2,3,*]

[1]Biomarkers and Nutritional & Food Metabolomics Research Group, Department of Nutrition, Food Science and Gastronomy, University of Barcelona, Barcelona, Spain, [2]Statistics and Bioinformatics Research Group, Department of Genetics, Microbiology and Statistics, University of Barcelona, Barcelona, Spain, [3]CIBERFES, Instituto de Salud Carlos III, Madrid, Spain and [4]Department of Biological Sciences, University of Alberta, Edmonton, AB, T6G 2E8, Canada
*Corresponding author: asanchez@ub.edu

Citation details: Castellano-Escuder,P., González-Domínguez,R., Wishart,D.S. *et al*. FOBI: an ontology to represent food intake data and associate it with metabolomic data. *Database* (2020) Vol. 2020: article ID baaa033; doi:10.1093/database/baaa033

Received 01 November 2019; Revised 28 February 2020; Accepted 22 April 2020

Abstract

Nutrition research can be conducted by using two complementary approaches: (i) traditional self-reporting methods or (ii) via metabolomics techniques to analyze food intake biomarkers in biofluids. However, the complexity and heterogeneity of these two very different types of data often hinder their analysis and integration. To manage this challenge, we have developed a novel ontology that describes food and their associated metabolite entities in a hierarchical way. This ontology uses a formal naming system, category definitions, properties and relations between both types of data. The ontology presented is called FOBI (Food-Biomarker Ontology) and it is composed of two interconnected sub-ontologies. One is a 'Food Ontology' consisting of raw foods and 'multi-component foods' while the second is a 'Biomarker Ontology' containing food intake biomarkers classified by their chemical classes. These two sub-ontologies are conceptually independent but interconnected by different properties. This allows data and information regarding foods and food biomarkers to be visualized in a bidirectional way, going from metabolomics to nutritional data or vice versa. Potential applications of this ontology include the annotation of foods and biomarkers using a well-defined and consistent nomenclature, the standardized reporting of metabolomics workflows (e.g. metabolite identification, experimental design) or the application of different enrichment analysis approaches to analyze nutrimetabolomic data. **Availability**: FOBI is freely available in both OWL (Web Ontology Language) and OBO (Open Biomedical Ontologies) formats at the project's Github repository (https://github.com/pcastellanoescuder/FoodBiomarkerOntology) and FOBI visualization tool is available in https://polcastellano.shinyapps.io/FOBI_Visualization_Tool/.

Introduction

The growing emergence of high-throughput analytical techniques in the life sciences over the past three decades, such as next-generation DNA sequencing, proteomics, metabolomics and other high-throughput omics approaches, has created significant challenges in data management. Currently, one of the main problems that researchers face lies in the question: *where are these data sets and how can I use them?* Unfortunately, the heterogeneity of storage platforms, data formats and privacy requirements of some of them often hinders their widespread access and use. In this vein, the creation of ontologies, defined as the 'specification of a representational vocabulary for a shared domain of discourse—definitions of classes, relations, functions and other objects' [1], is of vital importance to help analyze, annotate and homogenize these large and complex data sets [2, 3]. This is a major issue within the 'FAIR Guiding Principles for scientific data management and stewardship' [4], which aim to improve the findability, accessibility, interoperability and reusability of data. In particular, ontologies play a central role in the 'Interoperability' concept [1, 5], which establishes that '(meta)data has to use a formal, accessible, shared and broadly applicable language for knowledge representation' [4].

Nutritional research largely relies on accurate dietary assessment, which is of great relevance to evaluate food intake and dietary habits. Dietary assessments also help in understanding the association between nutrition and health status. Nutritional research is often conducted by using two complementary approaches: (i) self-reporting methods (e.g. food frequency questionnaires, dietary recalls) [6] and (ii) the measurement of dietary biomarkers using a variety of analytical chemistry techniques, including metabolomics [7, 8]. With regard to traditional dietary assessment tools, it should be noted that subjective self-reports generate very complex textual data, containing types and quantities of foods and recipes in very diverse and heterogeneous formats that depend on the country/region, socio-demographic factors, etc.

To properly annotate this nutritional data using a common language, the most relevant ontology in nutrition research is FoodOn [9]. FoodOn is a comprehensive ontology composed of 'term hierarchy facets' that cover basic raw food source ingredients, packaging methods, cooking methods and preservation methods. It also includes an upper-level consisting of a variety of product type schemes under which food products can be categorized. On the other hand, the metabolomics standards initiative has also highlighted the importance of ontologies in metabolomics [10]. As Schlegel *et al*. reported, 'the application of ontologies to metabolomics can improve the consistency of study data and can help link data using relationships that extend the computational capacity of the study data and enrich that knowledge source with a myriad of nationally available data to help fuel hypothesis driven laboratory based research' [3].

In response to this, ChEBI (Chemical Entities of Biological Interest, https://www.ebi.ac.uk/chebi/) has developed a reference ontology for describing chemical compounds of biological interest in terms of their chemical structures, chemical categories and roles [11]. The ChEBI ontology is manually maintained and annotated. More recently, an automatic method for describing and classifying chemicals, called ClassyFire [12], has been developed and widely adopted by databases such as ChEBI, PubChem [13] and the Human Metabolome Database (HMDB) [14]. ClassyFire uses the ChemOnt ontology, consisting of more than 4800 different categories (with definitions) hierarchically structured into 11 different levels (Kingdom, SuperClass, Class, SubClass, etc.). Additionally, the HMDB has developed the ChemFOnt (chemical functional ontology) to describe the biological and industrial functions of all the compounds and metabolites found in this database. ChemFOnt consists of four major categories (physiological effect, disposition, process and role), 152 sub-categories and more than 4100 defined terms.

Although most existing ontologies have been specifically designed for a single theme, there are also some others composed of interconnected sub-ontologies, thus enabling users to establish relationships among different variables. For instance, ChEBI is organized in two sub-ontologies: (i) 'Molecular Structure', in which molecular entities are classified according to structure and (ii) 'Subatomic Particle', which classifies particles smaller than atoms. On the other hand, the Gene Ontology, includes three independent sub-ontologies: (i) 'Biological process', referred to a biological objective to which the gene or gene product contributes; (ii) 'Molecular function', defined as the biochemical activity of a gene product; and (iii) 'Cellular component', which refers to the place in the cell where a gene product is active [15]. In this regard, we would argue that nutritional research also generates large amounts of complex and inter-related data coming from self-reporting methods and metabolomics experiments. Therefore, an interconnected set of sub-ontologies would be particularly useful for defining relationships between both metabolomics data and self-reported dietary questionnaires. To facilitate the construction of such an ontology that describes both foods and their associated metabolite biomarkers, we will draw from several open-access databases. These include Exposome-Explorer [16], Phenol-Explorer [17], PhytoHub (http://phytohub.eu/) and Food Database (FooDB) (http://

foodb.ca/)—all of which contain rich information about food constituents and food metabolites.

However, relationships between foods and their metabolites are extremely complex and the way they are described varies tremendously across these databases. This lack of commonality and the lack of a common, hierarchical structure makes data comparison and data searching quite difficult. Therefore, the development of a comprehensive ontology to clearly define the relationships between nutritional (food composition) and metabolomics (food metabolite or biomarker) data is needed. This ontology could have multiple practical applications in nutrimetabolomics, being the annotation of terms using a consistent and standardized nomenclature the most basic one, but of great importance in this research field due to the inherent complexity and heterogeneity of the data managed (i.e. multiple names/synonyms to define the same food/metabolite). Additionally, other potential applications of the ontology could be the ability to perform different enrichment analysis (e.g. to investigate patterns of food consumption on the basis of metabolomics data sets) or to conduct semantic similarity analysis (e.g. to establish novel associations between foods and metabolites). In this work, we describe FOBI (the Food-Biomarker Ontology), an ontology created with the aim of providing a common language to describe the many complex relationships in nutrimetabolomics research. This new ontology will allow users (and online databases) to integrate dictionaries and analyze these two kinds of data independently or together in a consistent and homogeneous way.

Results

FOBI is a freely available comprehensive ontology composed of two interconnected sub-ontologies including the 'Food Ontology' and the 'Biomarker Ontology'. This ontology has been built using Protégé [18] and is available in OWL (Web Ontology Language) and OBO (Open Biomedical Ontologies) formats at the project's Github repository (https://github.com/pcastellanoescuder/FoodBiomarkerOntology). FOBI consists of 1197 terms, 4 different properties, 13 food top-level classes, 11 biomarker top-level classes and more than 4500 relationships. Furthermore, FOBI is part of OBOFoundry project and FOBI IDs have been indexed into the HMDB and FooDB databases to facilitate the interoperability and the exchange of data.

Food Ontology

The Food Ontology was created on the basis of dietary data obtained from self-reported surveys for dietary assessment, including food frequency questionnaires (FFQ) and dietary recalls (DRs) [6]. The FFQ is a closed-ended survey that provides information on long-term dietary habits regarding a pre-defined list of 100–150 food items. On the other hand, DRs collect detailed information about foods consumed over a specific period (e.g. 24 hours, 3 days). To expand our Food Ontology as much as possible, we used the knowhow of our research group in working with FFQs and DRs collected from previous and ongoing projects. These projects involved cohorts from various European countries (e.g. Spain, France, United Kingdom). This allowed us to cover common foods for various dietary patterns, thus potentiating the applicability of FOBI in diverse research projects.

Accordingly, the Food Ontology is composed of more than 350 entities classified in different food classes. For this purpose, we considered both 'raw foods' and 'multi-component foods', with a multi-component food defined as any food item composed by two or more raw foods. In turn, the Food Ontology also describes the major ingredients forming part of each multi-component food according to the literature [19, 20]. These entities were annotated using a common nomenclature to reduce the complexity and heterogeneity of dietary data collected from free text questionnaires. This is because the same food/multi-component food can be named in many different ways (e.g. hamburger, burger, beef burger, etc.). Furthermore, FOBI also includes the FoodOn IDs for those food items common for both ontologies.

Major food classes in the Food Ontology were created considering both the nature of the food and the availability of food intake biomarkers for each class. A total of 13 food top-level classes were generated: beverage food product, cacao food product, dairy food product, egg food product, flavouring additive, fruits and vegetables, grain plant, lipid food product, meat food product, multi-component food, nuts and legumes, spice or herb and sugar. In turn, each of these 13 top-level classes have different subclass structures depending on its nature.

Biomarker Ontology

Food intake biomarkers (FIBs) are compounds derived directly from foods or the metabolism of food compounds that are characteristic or particular to a specific food item (e.g. phloretin for apple) or food category (e.g. glucosinolates for cruciferous vegetables) [7]. An important aspect to highlight on this regard is that, although the concentration of these metabolites in the food product may vary as a response to different factors (e.g. variety, agronomic practices, breeding, food processing), FIBs can always be associated with the consumption of the corresponding food (i.e. apple always contains phloretin, regardless the variety

or cultivation conditions). FIBs potentially consist of a vast number of chemicals with very different physico-chemical properties, including polyphenols and carotenoids, coming from plant-derived foods; derivatives of amino acids and fatty acids (mainly found in animal products); methylxanthines from coffee, tea and cocoa; alkaloids, organic acids and many others. Food constituents can undergo multiple biotransformation steps after ingestion, thus significantly expanding their metabolic complexity. Typically, xenobiotic food constituents are first subjected to phase I and phase II transformations, principally in the liver, kidneys and intestine, for detoxification purposes and to facilitate their excretion. Phase I metabolism normally involves cytochrome P450-mediated oxidation and hydrolysis transformations, while phase II reactions consist of chemical conjugations, such as methylation, acetylation, sulfation, glucuronidation and amino acid conjugation [21]. The gut microbiota also plays a major role in the metabolism of poorly bioavailable food derived metabolites, usually involving ring cleavage reactions and a variety of fermentative pathways to produce smaller, more easily absorbed derivatives [22]. Rather than trying to handle all possible compounds (possibly numbering in the tens of thousands), we chose to gather currently reported food derived metabolites and to define their relationships with foods and dietary patterns.

To create the Biomarker Ontology, we considered almost 600 known food metabolites, including dietary compounds and their host and microbiota-derived metabolites. These compounds were compiled from extensive literature reviews and the information contained in open access databases such as Phenol-Explorer, PhytoHUB and the FooDB. Of particular help was the material produced by the EU-funded FoodBAll project (http://foodmetabolome.org/), which worked on discovering and validating FIBs for a range of foods. The FoodBAll consortium has produced a collection of review articles published over the past 2 years focused on the most frequently consumed food groups [23–30]. Supplementary Table S1 summarizes the major classes of FIBs included in our first draft of FOBI and their associations with foods. It should be noted that this sub-ontology is only composed by food derived metabolites, while biomarkers of effect (i.e. endogenous metabolites altered after food intake) have been discarded. This is not intended to be a final, definitive ontology of food intake biomarkers, since it will be updated with novel FIBs as new studies are reported.

The FIBs in the Biomarker Ontology were classified according to their chemical classes using ClassyFire [12] and ChemOnt (version 2.1).

A key challenge in creating this sub-ontology was the complexity and diversity of the chemical nomenclature of food derived metabolites. For instance, caffeic acid, a relatively simple phenolic acid found in numerous foods such as coffee, can also be named as (E)-3-(3,4-dihydroxyphenyl)prop-2-enoic acid (IUPAC name), trans-3,4-dihydroxycinnamic acid, trans-3,4-dihydroxycinnamate or 3-(3,4-dihydroxyphenyl)acrylic acid, among other names. This disparity is even greater for more complex metabolites or phase II derivatives (e.g. caffeic acid 3-glucuronide, 3,4-dihydroxycinnamic acid 3-glucuronide, 4-hydroxycinnamic acid 3-O-glucuronide, (2S,3S,4S,5R,6S)-6-5-[(1E)-2-carboxyeth-1-en-1-yl]-2-hydroxyphenoxy-3,4,5-trihydroxyoxane-2-carboxylate). To facilitate the use of FOBI, metabolites are named according to the nomenclature commonly employed by nutrimetabolomic researchers, which easily enables users to differentiate isomers and similar metabolites within the same chemical class. Besides the FOBI ID, this ontology also lists the code numbers for HMDB, KEGG, ChEBI, PubChem, InChIKey, InChI and ChemSpider for all these compounds, if available, which further facilitates the interoperability of FOBI and the exchange of data. In addition, the Biomarker Ontology also contains some putative FIBs previously identified via targeted metabolomics by our research group [31, 32]. It should be noted that, for most of these biomarkers, only InChIKey and InChI codes are available. This is because only a few of them are listed in HMBD, KEGG, ChEBI, PubChem or ChemSpider so there is limited information about their biological roles and potential food sources.

In addition, we have created a synonym file with all these annotations for all food intake biomarkers or food metabolites included in the Biomarker Ontology, which can be freely download as a.csv file (Supplementary Table S2).

Ontology architecture

The architecture of FOBI is composed by classes corresponding to the items from the two sub-ontologies previously described (Food and Biomarker Ontologies), based on ChEBI (for metabolites) and FoodOn (for foods), respectively, and edges representing their relationships. Within the Food Ontology, raw foods are connected with the corresponding food class by the property *is_a*. On the other hand, multi-component foods are related to raw foods by the property *Contains*, in the same way that raw foods are connected with multi-component foods in the form of *IsIngredientOf*. For the Biomarker Ontology, the relationship between individual metabolites and the chemical class (defined by ClassyFire) is also defined by the property *is_a*. Finally, nodes from the Food and Biomarker Ontologies are interconnected by the inverse properties *BiomarkerOf* and *HasBiomarker*.

Figure 1 illustrates the FOBI architecture considering apple as an example. According to this, apple can be

Database, Vol. 2020, Article ID baaa033

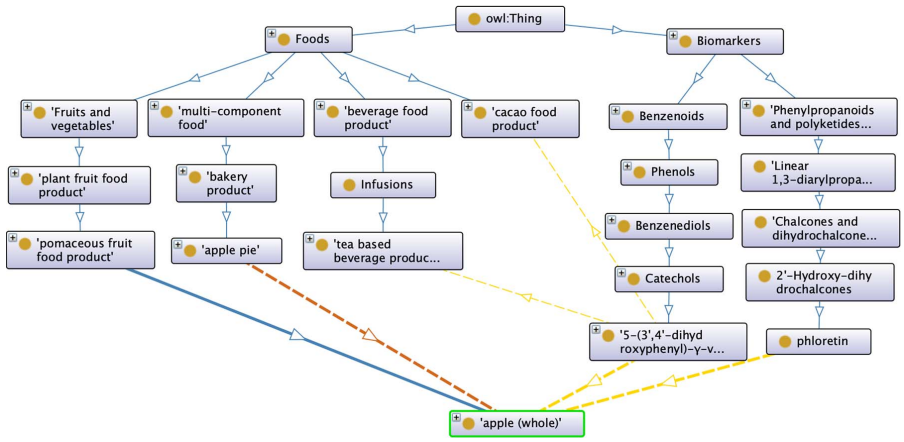


Figure 1. FOBI architecture considering apple as an example.

a raw food with the following relationships 'apple *is_a* pomaceous fruit food product *is_a* plant fruit food product *is_a* Fruits and vegetables *is_a* Food' (the property *is_a* is represented by blue arrows). In addition, apple can also be an ingredient in multi-component foods such as apple pie, so that 'apple *IsIngredientOf* apple pie *is_a* bakery product *is_a* multi-component food *is_a* Food' as well 'apple pie *Contains* apple' (the properties *IsIngredientOf* and *Contains* are represented by orange arrows). Considering phloretin and 5-(3',4'-dihydroxyphenyl)-γ-valerolactone as biomarkers of apple intake, they can be categorized as 'phloretin *is_a* 2'-Hydroxy-dihydrochalcone *is_a* Chalcones and dihydrochalcones *is_a* Linear 1,3-diarylpropanoid *is_a* Phenylpropanoids and polyketides *is_a* Biomarker' and '5-(3',4'-dihydroxyphenyl)-γ-valerolactone *is_a* Catechol *is_a* Benzenediol *is_a* Phenol *is_a* Benzenoid *is_a* Biomarker'. Because phloretin is a specific marker of apple, this metabolite is exclusively connected via the Food Ontology by the relationships 'phloretin *BiomarkerOf* apple' and 'apple *HasBiomarker* phloretin' (the properties *BiomarkerOf* and *HasBiomarker* are represented by yellow arrows). On the other hand, 5-(3',4'-dihydroxyphenyl)-γ-valerolactone can be derived from various procyanidin-rich foods (cacao, tea), so it can be connected with them following the same structure described for apple.

FOBI network analysis

To evaluate the information content of FOBI and its efficiency, we conducted network analysis to compute the average path length (APL) among FOBI's network nodes. The APL is defined as the average number of steps along the shortest paths for all possible pairs of network nodes. The APL can be used for enrichment analysis [33] and is considered a robust measure of a network's topology and its efficiency of information transport [34]. From a more pragmatic point view, the APL can be thought of as a measure to demonstrate whether the entities (or nodes) within an ontology are functionally cohesive. Thus, nodes with high cohesive functionality tend to have lower APL values compared to randomly selected nodes [33].

If we consider an unweighted directed graph G with the set of vertices V. Let $d(v_1, v_2)$, where $v_1, v_2 \in V$ denote the shortest distance between v_1 and v_2. Assume that $d(v_1, v_2) = 0$ if v_2 cannot be reached from v_1. Then, the APL l_G is:

$$l_G = \frac{1}{n \cdot (n-1)} \cdot \sum_{i \neq j} d(v_i, v_j), \qquad (1)$$

where n is the number of vertices in G.

To evaluate the FOBI network, we first calculated its APL and then created 10 000 random graphs using the Erdös–Rényi algorithm [35] and calculated the mean of their APLs. The computed FOBI APL value was 2.33, which is 114.26 standard deviations below the random mean APL (5.30) (Figure 2), thus demonstrating the very high information transport efficiency of FOBI compared to a random network.

Implementation of the FOBI web application

The FOBI's web application (https://polcastellano.shinyapps.io/FOBI_Visualization_Tool/) is powered by Shiny (https://shiny.rstudio.com). This Shiny app imports all FOBI

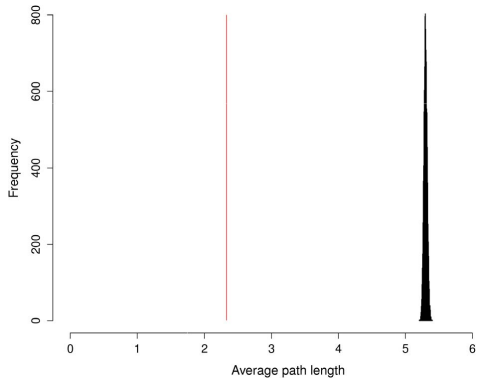


Figure 2. APL of FOBI versus random graphs APLs.

relationships in R and organize them in a table or in a graph according to the user input.

The FOBI application settings panel is shown in Figure 3A. This user-friendly application accepts both food and biomarker entries as FOBI entity, which can be displayed using multiple layouts and using different properties (*is_a*, *BiomarkerOf* and *Contains*). Results can be downloaded in either a table or a graph format.

As summarized in Figure 3, this FOBI application considerably simplifies inspection of the interrelationships between foods and biomarkers. On one hand, food items can be interrelated by their properties *is_a* and *Contains* to show their classification according to food classes and the presence of ingredients in complex multi-component foods (Figure 3B). Similarly, food intake biomarkers can also be categorized according to their chemical class and additionally related to foods by their properties *is_a* and *BiomarkerOf* (Figure 3C).

Discussion

Knowledge of the complex inter-relationships between food types, food components, food ingredients and food intake biomarkers is critical to facilitate our understanding of nutrition and metabolism. Such an understanding will enable accurate metabolomics-based food intake assessment and assist with the development of personalized nutrition strategies to select specific diets according to the subject's phenotype, disease state, microbiome, metabolome, etc.

To this end, we have developed the FOBI. This is a nutrition-specific ontology composed of two sub-ontologies with an independent hierarchy but clear relationships between them. The Food Ontology consists of known foods grouped according to their major (13) nutritional classes, while the Biomarker Ontology contains food derived metabolites categorized according to their chemical classes. The edges linking these two sub-ontologies define, using a common language, the hierarchy of each food and food biomarker entity, as well as the properties that relate these two kinds of data. This architecture can be easily interpreted at both the user and computational level. Likewise, FOBI can be used for different purposes, from making simple queries to complex computational queries simultaneously using all the information stored in the ontology.

Ontologies facilitate many practical applications, such as annotating entities or items, performing enrichment analysis (e.g. over representation analysis), conducting semantic similarity analysis [2] or even to find unexpected patterns. Some of the potential applications of FOBI are described below. The most basic application of FOBI is in the annotation of foods and related food biomarkers using a consistent, well-defined, fully standardized nomenclature. This will facilitate the comparability and interoperability between nutrition studies, projects and research groups. FOBI will also facilitate nutrimetabolomics research thanks to the comprehensive description of associations between food types and food derived metabolites, as summarized in Supplementary Table S1. For instance, interrelationships defined in this ontology, together with the accurate nomenclature defined in the synonym file, will be particularly useful in untargeted metabolomics studies (e.g. acute intervention studies) for discriminant feature identification.

Furthermore, this information can also serve to facilitate study designs, from hypothesis generation (e.g. expected metabolites occurring after a dietary intervention) to experimental design (e.g. optimization of targeted metabolomics methods focused on analytes of interest). Additionally, the availability of FOBI will give nutrimetabolomic researchers the ability to perform enrichment analysis. Given a set of metabolites (e.g. discriminant metabolites identified in a metabolomics study), the hierarchical structure of FOBI enables one to evaluate possible over-representation of specific chemical classes, which could reflect the consumption of particular foods or food groups. This could be a first step towards 'food enrichment analysis'.

Limitations

FOBI has two main limitations. The first one concerns the relationship between foods and their metabolites. The relationships that FOBI contain are limited to the best known and most frequent. However, there can be relationships between foods and their metabolites that FOBI does not contain due to the fact that not every food compound (or its

Database, Vol. 2020, Article ID baaa033 **Page 7 of 8**

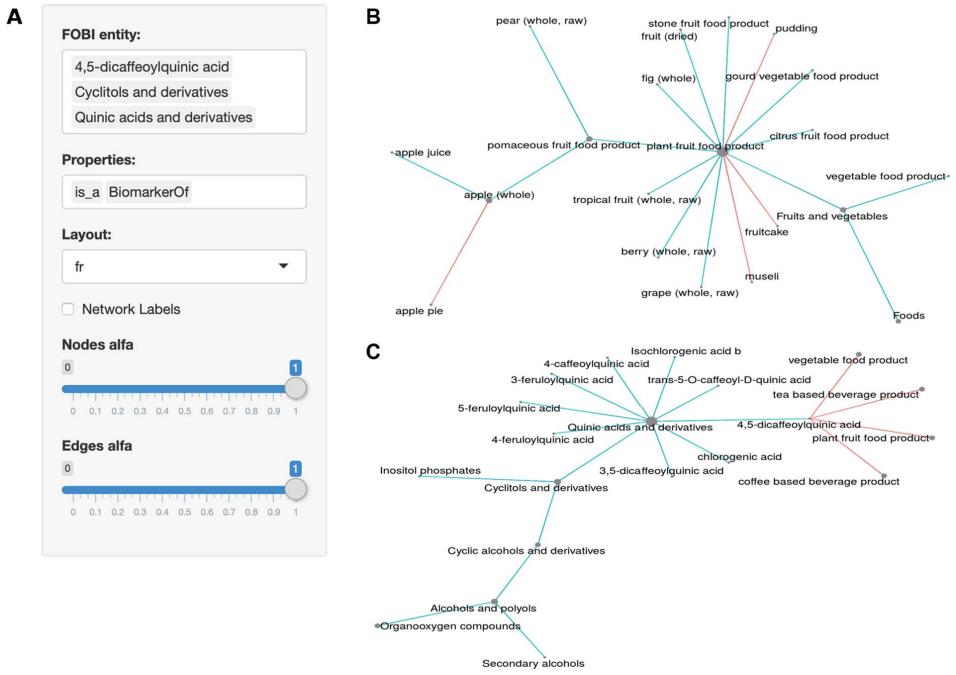


Figure 3. The FOBI web application. (**A**) Settings panel of the web interface; (**B**) interrelationships of food items in FOBI; (**C**) relationships between food items and metabolites in FOBI.

metabolite) has been tested for its presence in certain foods or within certain human biofluids.

The second limitation concerns the limited number of foods and food metabolites/biomarkers in FOBI. Currently, FOBI has more than 350 food nodes (in total) and 590 food biomarkers (only metabolites) corresponding to more than 4500 relationships (among foods, among biomarkers and between foods and biomarkers). As with most ontologies, FOBI is undergoing constant evolution and development. As a result, the number of entities and the quality of relationships described in this ontology will be continuously increasing and improving.

Future work

FOBI is an open-source project that can be readily used and enhanced by anyone in the nutritional and nutrimetabolomic community. Further expansion of the ontology to cover more food types, more food biomarkers and more relationships will certainly increase its utility.

Future efforts will be directed at expanding this ontology and extending it so that it is more widely used in other curated databases such as Exposome-Explorer, Phenol-Explorer, HMDB and FooDB.

Conclusion

FOBI is the first ontology that integrates nutritional and metabolomic data in a comprehensive common language. At the moment, FOBI has a total of 1197 terms (366 from Food Ontology and 831 from Biomarker Ontology), 11 chemical top-level classes, 13 food top-level classes and 4 different properties that are fully defined and which have clear relationship mappings. FOBI defines the relationships between foods and their metabolites (biomarkers) through a formal ontology.

FOBI allows experts to annotate and analyze nutritional and metabolomic data in a consistent way, making the results comparable between and across studies in the same field. The development of FOBI will lead to an improvement in the interoperability of nutritional and nutrimetabolomic data thereby making the data sets generated from these studies fully FAIR compliant.

Funding

Spanish Ministry of Economy and Competitiveness (MINECO) together with the Joint Programming Initiative 'A Healthy Diet for a Healthy Life' (PCIN-2014-133; 2015-238 & PCIN2017-076); the CIBERfes and ISCIII project (AC19/00096) (co-funded

by the FEDER Program from the European Union, 'A way to make Europe'); the Generalitat de Catalunya's Agency AGAUR (2017 SGR 1546); ICREA Academia Award and the EIT Health Innovation by Design project COOK2HEALTH; EIT Health is supported by the European Institute of Innovation and Technology, a body of the European Union; the 'Juan de la Cierva' program from MINECO (FJCI2015-26590 to R.G.-D.).

Supplementary Data

Supplementary data are available at *Database* Online.

Conflict of interest. None declared.

References

1. Kramer,F. and Beißbarth,T. (2017) Working with ontologies. *Methods Mol. Biol.*, **1525**, 123–135.
2. Hoehndorf,R., Schofield,P.N. and Gkoutos,G.V. (2015) The role of ontologies in biological and biomedical research: a functional perspective. *Brief. Bioinform.*, **16**, 1069–1080.
3. Schlegel,D.R., Ruttenberg,A. and Elkin,P.L. (2015) Ontologies in Metabolomics. *Metabolomics*, **5**, e137.
4. Wilkinson,M.D., Dumontier,M., Aalbersberg,I.J. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.
5. Noy,N.F. and McGuinness,D.L. (2001) Ontology development 101: a guide to creating your first ontology. *Stanford Knowledge Systems Laboratory Technical Report, KSL-01-05 and Stanford Medical Informatics Technical Report, SMI-2001-0880*.
6. Shim,J.S., Oh,K. and Kim,H.C. (2014) Dietary assessment methods in epidemiologic studies. *Epidemiol. Health*, **36**.
7. Scalbert,A., Brennan,L., Manach,C. *et al.* (2014) The food metabolome: a window over dietary exposure. *Am. J. Clin. Nutr.*, **99**, 1286–1308.
8. Ulaszewska,M.M., Weinert,C.H., Trimigno,A. *et al.* (2019) Nutrimetabolomics: an integrative action for metabolomic analyses in human nutritional studies. *Mol. Nutr. Food Res.*, **63**, 1800384.
9. Dooley,D.M., Griffiths,E.J., Gosal,G.S. *et al.* (2018) FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. *NPJ Sci. Food*, **2**, 23.
10. Sansone,S.A., Schober,D., Atherton,H.J. *et al.* (2007) Metabolomics standards initiative: ontology working group work in progress. *Metabolomics*, **3**, 249–256.
11. Degtyarenko,K., De Matos,P., Ennis,M. *et al.* (2007) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, **36**, D344–D350.
12. Feunang,Y.D., Eisner,R., Knox,C. *et al.* (2016) ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminform.*, **8**, 61.
13. Kim,S., Chen,J., Cheng,T. *et al.* (2019) PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.*, **47**, D1102–D1109.
14. Wishart,D.S., Feunang,Y.D., Marcu,A. *et al.* (2017) HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.*, **46**, D608–D617.
15. Ashburner,M., Ball,C.A., Blake,J.A. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
16. Neveu,V., Moussy,A., Rouaix,H. *et al.* (2016) Exposome-Explorer: a manually-curated database on biomarkers of exposure to dietary and environmental factors. *Nucleic Acids Res.*, **45**., D979–D984.
17. Rothwell,J.A., Perez-Jimenez,J., Neveu,V. *et al.* (2013) Phenol-Explorer 3.0: a major update of the Phenol-Explorer database to incorporate data on the effects of food processing on polyphenol content. *Database*, **2013**:bat070.
18. Musen,M.A. (2015) The protégé project: a look back and a look forward. *AI Matters*, **1**, 4–12.
19. McCance,R.A. and Widdowson,E.M. (2014) *The Composition of Foods*. Royal Society of Chemistry, Cambridge, UK.
20. Reinivuo,H., Bell,S. and Ovaskainen,M.L. (2009) Harmonisation of recipe calculation procedures in European food composition databases. *J. Food Compos. Anal.*, **22**, 410–413.
21. Manach,C. and Donovan,J.L. (2004) Pharmacokinetics and metabolism of dietary flavonoids in humans. *Free Radic. Res.*, **38**, 771–785.
22. Rowland,I., Gibson,G., Heinken,A. *et al.* (2018) Gut microbiota functions: metabolism of nutrients and other food components. *Eur. J. Nutr.*, **57**, 1–24.
23. Rothwell,J.A., Madrid-Gambin,F., Garcia-Aloy,M. *et al.* (2018) Biomarkers of intake for coffee, tea, and sweetened beverages. *Genes Nutr.*, **13**, 15.
24. Michielsen,C.C., Almanza-Aguilera,E., Brouwer-Brolsma,E.M. *et al.* (2018) Biomarkers of food intake for cocoa and liquorice (products): a systematic review. *Genes Nutr.*, **13**, 22.
25. Praticò,G., Gao,Q., Manach,C. *et al.* (2018) Biomarkers of food intake for Allium vegetables. *Genes Nutr.*, **13**, 34.
26. Ulaszewska,M., Vázquez-Manjarrez,N., Garcia-Aloy,M. *et al.* (2018) Food intake biomarkers for apple, pear, and stone fruit. *Genes Nutr.*, **13**, 29.
27. Münger,L.H., Garcia-Aloy,M., Vázquez-Fresno,R. *et al.* (2018) Biomarker of food intake for assessing the consumption of dairy and egg products. *Genes Nutr.*, **13**, 26.
28. Zhou,X., Gao,Q., Praticò,G. *et al.* (2019) Biomarkers of tuber intake. *Genes Nutr.*, **14**, 9.
29. Garcia-Aloy,M., Hulshof,P.J., Estruel-Amades,S. *et al.* (2019) Biomarkers of food intake for nuts and vegetable oils: an extensive literature search. *Genes Nutr.*, **14**, 7.
30. Harsha,P.S.S., Wahab,R.A., Garcia-Aloy,M. *et al.* (2018) Biomarkers of legume intake in human intervention and observational studies: a systematic review. *Genes Nutr.*, **13**, 25.
31. González-Domínguez,R., Urpi-Sarda,M., Jáuregui,O. *et al.* (2020) Quantitative dietary fingerprinting (QDF)—a novel tool for comprehensive dietary assessment based on urinary nutrimetabolomics. *J. Agric. Food Chem.*, **68**, 1851–1861.
32. González-Domínguez,R., Jáuregui,O., Mena,P. *et al.* (2020) Quantifying the human diet in the crosstalk between nutrition and health by multi-targeted metabolomics of food and microbiota-derived metabolites. *Int. J. Obes.*, In press,
33. Embar,V., Handen,A. and Ganapathiraju,M.K. (2016) Is the average shortest path length of gene set a reflection of their biological relatedness? *J. Bioinform. Comput. Biol.*, **14**, 1660002.
34. Albert,R. and Barabási,A.L. (2002) Statistical mechanics of complex networks. *Rev. Mod. Phys.*, **74**, 47.
35. Erdös,P. and Rényi,A. (1959) On random graphs. *Publ. Math. Debrecen*, **6**, 290–297.

A.1.2 Paper 2: POMAShiny

The manuscript attached below consists of the draft accepted by the journal and is not the final published version, as this does not yet exist.

POMAShiny: a user-friendly web-based workflow for metabolomics and proteomics data analysis

Pol Castellano-Escuder[1,2,3,*], Raúl González-Domínguez[1,3], Francesc Carmona-Pontaque[2,3], Cristina Andrés-Lacueva[1,3], Alex Sánchez-Pla[2,3,*]

1 Biomarkers and Nutritional & Food Metabolomics Research Group, Department of Nutrition, Food Science and Gastronomy, Food Innovation Network (XIA), University of Barcelona, Barcelona, Spain.
2 Statistics and Bioinformatics Research Group, Department of Genetics, Microbiology and Statistics, University of Barcelona, Barcelona, Spain.
3 CIBERFES, Instituto de Salud Carlos III, Madrid, Spain.

*pcastellano@ub.edu and asanchez@ub.edu

Abstract

Metabolomics and proteomics, like other omics domains, usually face a data mining challenge in providing an understandable output to advance in biomarker discovery and precision medicine. Often, statistical analysis is one of the most difficult challenges and it is critical in the subsequent biological interpretation of the results. Because of this, combined with the computational programming skills needed for this type of analysis, several bioinformatic tools aimed at simplifying metabolomics and proteomics data analysis have emerged. However, sometimes the analysis is still limited to a few hidebound statistical methods and to data sets with limited flexibility. POMAShiny is a web-based tool that provides a structured, flexible and user-friendly workflow for the visualization, exploration and statistical analysis of metabolomics and proteomics data. This tool integrates several statistical methods, some of them widely used in other types of omics, and it is based on the POMA R/Bioconductor package, which increases the reproducibility and flexibility of analyses outside the web environment. POMAShiny and POMA are both freely available at `https://github.com/nutrimetabolomics/POMAShiny` and `https://github.com/nutrimetabolomics/POMA`, respectively.

Author Summary

Metabolomics and proteomics are two growing areas in human health and personalized medicine fields. Often, one of the main applications of metabolomics and proteomics is the discovery of novel biomarkers and new therapeutic targets in these areas. However, these data are extremely complex and hard to analyse, since they have a large number of features, several missing values, and often important clinical variables to consider in the analyses. Therefore, powerful and versatile tools are needed to provide efficient methods for data visualization and exploration, as well as a wide range of robust statistical methods to meet all data and users requirements. Although powerful tools do exist for the analysis of these data, many of them are still limiting the analyses in terms of visualization and statistical analysis. To address this limitation and complement the existing tools, we have developed a web-based application, named POMAShiny, for the

data analysis of metabolomics and proteomics. This novel and versatile tool offers a wholly interactive and easy-to-use environment for the analysis of these data, including numerous methods for preprocessing, data visualization and statistical analysis. The POMAShiny open-source tool is extremely flexible and portable, as it can be installed locally and freely accessed online at `https://webapps.nutrimetabolomics.com/POMAShiny`.

Introduction

Metabolomics and proteomics are two rapidly growing areas of omics science that employ analytical techniques such as liquid chromatography (LC), gas chromatography (GC), capillary electrophoresis (CE), mass spectrometry (MS) and nuclear magnetic resonance (NMR) [1]. In turn, the results derived from these analytical techniques can be analysed in many different ways. Often, one of the main applications of metabolomics and proteomics is the characterization of novel therapeutic targets and patterns in human health and precision medicine fields [2,3].

During the last decade, many open-source tools have emerged that contribute to the analysis of these complex data [4]. However, most of these tools remain very specific and may limit the analysis to a few statistical methods [5], forcing researchers to use an extensive battery of different tools to meet all the needs of the analysis.

Currently, statistical analysis of metabolomics and proteomics data is mainly conducted by using several programming tools [4] and/or via different web-based tools [6–9] according to the aims of the researchers. Web-based tools are often a very popular choice for researchers, as they provide fast and easy-to-use graphical interfaces that bring statistical analysis closer to the scientific community without the need for extensive programming skills. However, additional statistical approaches that are not implemented in these tools can be really useful in the data analysis process.

In an effort to contribute to the extension of available methods and options for metabolomics and proteomics data analysis, POMAShiny provides a comprehensive and structured workflow that covers the preprocessing, exploratory data analysis and statistical analysis of these data. This workflow is integrated into a user-friendly, attractive web-based user interface, mainly focused on statistical analysis. This new tool provides several powerful methods, including univariate statistical methods, multivariate and dimension reduction methods, feature selection methods, regularized regression analysis approaches, machine learning classification algorithms, prediction model strategies and several high-quality interactive visualization options.

This new tool is based on the POMA R/Bioconductor package (`http://www.bioconductor.org`). POMAShiny integrates many of the most widely used methods for metabolomics and proteomics data analysis [4,5] and incorporates new useful and powerful alternatives.

The joint existence of both the POMA R/Bioconductor package and the POMAShiny web interface means a huge increase in the reproducibility of the tool, contributing to the reusability of previous existing methods in the R and Bioconductor environments [10,11], as well as allowing easy extension, integration and interoperability with other workflows, such as the RforMassSpectrometry initiative (`RforMassSpectrometry.org`), which provides the data structures used by the POMA package. Therefore, users can perform the spectral data processing and other routine MS workflow operations using the RforMassSpectrometry complementary packages, and then easily migrate to POMA to perform the statistical analysis without changing the data structure.

Design and Implementation

POMAShiny is a web-based tool wholly written in the open-source R programming language [10] and available under GPL-3.0 license. POMAShiny is powered by the Shiny framework [12] and all source code is available at the project's GitHub repository (`https://github.com/nutrimetabolomics/POMAShiny`).

On the one hand, POMAShiny's back-end structure uses the POMA [13], MSnbase [14] and tidyverse family [15] R packages to keep all code clean and as readable as possible, thereby facilitating the software maintenance. On the other hand, POMAShiny's front-end is based on the bs4Dash [16] R package, providing a highly easy-to-use dashboard design with most JavaScript features that makes the web interface very attractive for users. According to this design, the main menu with all panels and options is on the left side of the page while the main display screen is in the centre right of the page.

All functions provided in POMAShiny are tested with the testthat R package [17] on a continuous integration system using Travis, AppVeyor and GitHub Actions, covering tests on Linux, Mac and Windows with the current R versions and achieving more than 95% of code coverage [18] (`https://github.com/nutrimetabolomics/POMA`).

Users can download and launch POMAShiny locally (on Linux, Mac and Windows) or they can access the app online version hosted at `https://webapps.nutrimetabolomics.com/POMAShiny`. For a better experience, the authors recommend Safari or Chrome web browsers.

POMAShiny has been containerized using Docker. The Docker image is freely available at DockerHub, meaning a huge increase in the reproducibility, portability and scalability of the tool. Both local and Docker launch instructions are available at `https://github.com/nutrimetabolomics/POMAShiny`.

Results

POMAShiny provides an analysis workflow structured in the four sequential and well-defined panels described below, namely 1) data upload, 2) preprocessing, 3) exploratory data analysis (EDA) and 4) statistical analysis, all of them with their respective subpanels (see Fig 1).

Fig 1. POMAShiny's workflow. The workflow is divided into four well-separated panels. Both target and features files are required as an input. Once these files are uploaded, the data are preprocessed and prepared for display in the exploratory data analysis panel (EDA). Finally, after the preprocessing and EDA panels, several statistical methods and options are provided in the statistical analysis panel, where users can analyse the data and download the results in both plot and table formats (*icons made by Freepik from www.flaticon.com*).

Data upload

In order to keep "raw data" as "raw" as possible and create the ability to include covariates in the analysis, POMAShiny requires two comma-separated value (CSV) files as an input: the target (or metadata) file and the features file. The target file must provide the sample names in the first column and the group labels (e.g., control and case) in the second column. Optionally, from the third column (included), users can

upload relevant covariates to be used in subsequent statistical analysis. The features file contains all quantified features in the experiment, one in each column starting from the first one. The order of rows in both files must be the same. Once these files are uploaded, POMAShiny converts them internally into an *MSnSet* object, as defined in the MSnbase R/Bioconductor package [14].

Once the target and features files are uploaded, users can select specific rows (samples) on the target file tab to create a data subset for those selected samples. If any selection is made, only the selected samples will be analysed.

Furthermore, POMAShiny provides a function that allows users to combine different features that are part of the same entity. This optional operation can be very useful when the data contain different peptides that are part of the same protein or different ions representing the same compound. If users enable this option, a "grouping file" (CSV) indicating which features should be combined will be required. Several methods for performing this task are provided in POMAShiny, as well as the ability to download a table with the coefficients of variation of those combined features.

Preprocessing

Missing value imputation. Often, for biological and technical reasons (e.g., inaccurate peak detection, values under the limit of quantification, etc.), some features cannot be identified or quantified in some metabolomics and proteomics samples [19, 20]. To address this problem, several methods have been developed and compared to identify the best approaches for data imputation in this context [19–21]. POMAShiny workflow offers a missing value imputation panel composed of different operations divided into three sequential steps: 1) distinction between zeros and missing values, 2) removal of features with a high percentage of missing values, and 3) imputation of remaining missing values.

First, if the data contain zeros or a combination of zeros and missing values, users can distinguish between them by using the algorithm provided by POMAShiny. For example, if the data contain both endogenous and exogenous compounds, the exogenous ones could be a real zero if they are not in the sample (absence), while the endogenous ones are unlikely to be real zeros (since they should always be in the sample, at least in very low concentrations). In that case, users could consider zeros as real zeros and impute only the missing values in the data. However, if users do not know the exact nature of the zeros in the data, or the difference between zeros and missing values, the authors recommend considering all zeros as missing values.

Second, users are able to remove features of the data with more than a specific percentage of missing values in all the study groups (20% allowed by default). Finally, several imputation methods for dealing with the remaining missing values after the two previous steps are provided in POMAShiny. The available methods are the imputation by zero, the half minimum imputation, the imputation by median, the imputation by mean, the imputation by minimum and the k-nearest neighbours imputation (where missing values are imputed using the k-nearest neighbours algorithm [22]).

Normalization. It is generally accepted that some factors can introduce variability in metabolomics and proteomics data. Even if the data have been generated under identical experimental conditions, this introduced variability can have a critical influence on the final statistical results, making normalization a key step in the workflow [23]. These factors include: 1) differences in orders of magnitude between measured feature concentrations, 2) differences in the fold changes in feature concentration due to the induced variability, 3) large fluctuations in the concentration of some features under identical experimental conditions, 4) technical variability, and 5) heteroscedasticity [24]. POMAShiny provides six different normalization methods

widely used in this field and compared in different studies [24, 25]. Here, the normalization process comprises both the transformation and scaling of the data in a single step. The methods available for this purpose are autoscaling, level scaling, log scaling, log transformation, vast scaling, and log pareto scaling.

Outlier detection. The last step provided in this panel is outlier detection and data cleaning. Outliers are defined as observations that are not concordant with those of the vast majority of the remaining data points [20]. Outliers in metabolomics and proteomics can be separated into biological and analytical outliers [26]. On the one hand, the first group reflects random and induced biological variations that make some observations different from others. On the other hand, the second group reflects different kinds of problems during the analytical process (e.g., sampling, storage) [26]. These values can have an enormous influence on the resultant statistical analysis, making it difficult to meet all required assumptions in the most commonly applied statistical tests as well as all required assumptions in many regression techniques and predictive modelling approaches. Therefore, outlier detection procedures are a critical point on which all subsequent analysis will depend (both inference and predictive statistics).

POMAShiny allows the analysis of outliers by different plots and tables as well as the possibility of removing statistical outliers from the analysis using different customizable parameters (see Fig 2).

Fig 2. Screenshot of the "Outlier detection" panel showing the Euclidean distances in principal coordinate space between samples and their respective group centroid. ST000284 example data were used to create this plot.

Here, we propose an *ad hoc* multivariate outlier detection method based on the Euclidean distances among observations and their distances to each group centroid in a two-dimensional space (maximum, manhattan, canberra and minkowski distances are also available). Once the distances are computed, the classical univariate outlier detection formula $Q_3 + x * IQR$ is used to detect multivariate group-dependent outliers using the computed distances to each group centroid (x; the higher this value, the less sensitive the method is to outliers) (Fig 2).

Exploratory data analysis

As discussed in the preprocessing section, many uncontrolled factors can introduce bias in a systematic manner in metabolomics and proteomics experiments: different chromatographic columns, eventual repair of the LC-MS system, different laboratory conditions, etc. [27]. Exploratory data analysis (EDA) can help in evidencing some of these confounding factors and possible outliers [27]. For that reason, it is highly recommended to perform an EDA before any statistical analysis [28, 29]. Moreover, in the case of negligible confounding factors or outliers, EDA can also be useful for getting a first idea of those most interesting features in the study.

POMAShiny offers several interactive and highly customizable plots designed to facilitate this process, providing a wide range of visualization options. The specific EDA functionalities implemented in POMAShiny are the volcano plot (for two-group studies), boxplots, density plot and heatmap. However, PCA (principal component analysis) and cluster analysis should also be considered in this section.

POMAShiny interactive boxplots are designed to visualize all features at once; however, users can easily customize this plot to display only features of interest. Unlike boxplots, the interactive density plot is designed to explore the distribution of the study groups. Alternatively, users can also display the distribution of specific features.

POMAShiny also provides a clustered heatmap with a color stripe that corresponds to the group study label of each sample. Finally, for two-group studies, including a volcano plot in EDA can be very helpful for exploring those features that may be most influential in the study. POMAShiny's interactive volcano plot is based on the results of a T-test, where users can specify if the data are paired or if the variance is equal in the study groups.

PDF report. In an effort to facilitate the EDA process, POMAShiny includes a function to automatically generate a PDF report with a complete EDA, including all plots mentioned above. Users can generate the PDF report by clicking the "Exploratory report" button in the data upload panel.

The automatically generated PDF report provides information about the number of samples, features, covariates and the main study groups, as well as information on the percentage of total missing values in the data and the specific number of missing values per feature. Moreover, information on the number of zeros and features without variability is also provided. All the information provided in the PDF report is given in tables, plots and text format. A section with boxplots and density plots before and after missing value imputation and normalization (k-NN and log pareto scaling methods by default) is also included, providing users with valuable information about the preprocessing effect on the data. Furthermore, an outlier analysis, highly correlated features (r> 0.97), clustered heatmap and a PCA scores plot are also provided.

A POMAShiny PDF report helps users to have a quick and accurate description of the data. An example of this report is included as a vignette in the POMA package and it is also available at POMA's GitHub repository.

Statistical analysis

This panel encompasses several statistical methods, from the most commonly used approaches in metabolomics and proteomics data analysis to other less frequently used methodologies in these fields. All statistical methods offered in POMAShiny (Table 1) are implemented in a highly user-friendly way and generate both downloadable tables and interactive plots as outputs.

Univariate analysis. POMAShiny offers four widely used methodologies for performing classical parametric and non-parametric univariate tests. On the one hand, T-test (two-group analysis) and ANOVA (> 2 group analysis) methods are available to perform parametric tests. In the ANOVA tab, an ANCOVA (analysis of covariance) model is also computed if covariates are included in the target file. On the other hand, the Mann-Whitney U test (two-group analysis) and Kruskal-Wallis test (>2 group analysis) are available for non-parametric analysis. Each of these methodologies offers customizable parameters to adjust the analysis to users and data requirements. Due to the large number of tests performed in these types of analyses, the FDR (false discovery rate) method is used to compute adjusted p-values.

Limma. Limma (linear models for microarray data) is a univariate method created for the statistical analysis of gene expression experiments as microarrays [30]. In recent years, this approach has become the main choice for many researchers to explore and identify differential expressed genes between two conditions. Due to the many similarities between metabolomics, proteomics and microarray data (often hundreds of features and small sample sizes, quantitative data, etc.), limma can be used in metabolomics and proteomics data sets when they meet the requirements (e.g., feature normal distributions). POMAShiny allows users to perform limma models easily and include covariates in the model, if necessary. If covariates are provided in the target file

Table 1. Statistical methods provided in POMAShiny. *Methods that allow the use of covariates.

Univariate methods	Parametric	T-test (paired/unpaired)
		ANOVA
		ANCOVA*
		Limma*
	Non-parametric	Mann-Whitney U test (paired/unpaired)
		Kruskal-Wallis
Multivariate methods	Unsupervised	PCA
		k-means
		Multidimensional scaling (MDS)
	Supervised	PLS-DA
		sPLS-DA
Correlation methods	Parametric	Pearson's correlation*
	Non-parametric	Spearman's correlation*
		Kendall's correlation*
	Visualization	Gaussian graphical models (GGMs)
Statistical learning methods	Regularized regression	LASSO regression
		Ridge regression
		Elasticnet regression
	Decision trees	Random forest
Generalized linear models	Logistic regression	Odds ratio calculation*
Permutation tests	Non-parametric	Rank products

(e.g., batch effects, sex, age), two limma models – with and without covariates – are computed automatically, including the covariates in the model in the order in which they are provided in the target file (the further to the left, the more importance in the model). An interactive volcano plot based on the limma results is also generated in this tab.

Multivariate analysis. Multivariate methods focus analyses on the observation of more than one feature at a time, taking into account the different relationships between features. These methods can provide information about the structure of the data and different internal relationships that would not be observed with univariate statistics. However, the interpretation of these analyses can be more complex.

The most frequently used multivariate methodologies in metabolomics and proteomics statistical analysis are PCA, for unsupervised analysis, and PLS (partial least squares), for supervised analysis [31]. POMAShiny provides a collection of three different multivariate approaches powered by the mixOmics Bioconductor package [32]. The provided methods are PCA, PLS-DA (partial least squares discriminant analysis), and sPLS-DA (sparse partial least squares discriminant analysis).

PCA is an unsupervised method for dimension reduction that is done by calculating the data covariance matrix and performing eigenvalue decomposition on this covariance matrix without considering sample groups. In contrast, PLS-DA is a supervised method that uses the multiple linear regression method to find the direction of maximum covariance between the data and sample group [33]. sPLS-DA has been presented elsewhere [34] as an extension of sPLS (sparse partial least squares) [35] designed for classification problems. Note that while PCA is often used in exploratory data analysis, PLS-DA and sPLS-DA are used for classification and feature selection purposes, respectively. Several tuning parameters are available in all multivariate methods provided in POMAShiny. Users can define the number of components to compute, numerous graphical parameters, the number of features to select (in sPLS-DA), the VIP (variable importance in the projection) cut-off (in PLS-DA) and the cross-validation method to use, including both leave-one-out (LOO) and k-fold cross-validation.

Cluster analysis. Cluster analysis is also composed of multivariate methods, however this section is separated from multivariate methods to make POMAShiny structure clearer and more intuitive. The cluster analysis provided in POMAShiny allows users to explore different clusters in the data using the k-means algorithm [36]. k-means is an unsupervised method aimed at assigning all samples of the study to k clusters based on the sample means. By default, the optimal number of clusters (k) is determined through the popular "elbow method". Alternatively, users can define a specific number of clusters.

To provide a multivariate visualization of computed clusters, POMAShiny projects these clusters in the first two dimensions of a multidimensional scaling (MDS) plot [37]. Many user-customizable parameters are offered to define the distance used in MDS calculation. POMAShiny provides a table with the assigned cluster to each sample and an interactive MDS plot with computed clusters. This feature serves the users both in terms of cluster analysis and in calculating a classic MDS, integrating two useful functionalities within the same tab. As mentioned before, this method can also be useful in EDA.

Correlation analysis. Correlation analysis is usually one of the preferred options for evaluating the strength of relationships between different features [38].

POMAShiny provides different approaches to conducting an accurate correlation analysis. First, POMAShiny provides a highly customizable and interactive scatterplot of pairwise correlation between features (Fig 3). Here, users can select two different

features and explore them in a very comfortable way, as they are able to remove some points of the plot by clicking on them, drawing a smooth line based on a linear model, and exploring pairwise correlations within each study group and among factorial covariates (if they are provided). A downloadable table with all pairwise correlations between features is also provided. Moreover, POMAShiny provides a global correlation plot (or correlogram) and a network correlation plot. For all of the above methods and options, the three most common methods for calculating correlation coefficients – Pearson, Spearman and Kendall – are available.

Fig 3. Screenshot of the "Correlation Analysis" panel showing an interactive scatterplot of pairwise correlation between acetoacetate and epinephrine in the three different study groups "CRC", "Healthy" and "Polyp". ST000284 example data were used to create these plots.

Lastly, POMAShiny provides an alternative method for correlation network visualization in this tab. Estimation of Gaussian graphical models (GGMs) through the glasso R package [39] is also provided here. Thus, users can define the regularization parameter to estimate a sparse inverse correlation matrix using LASSO [40] and visualize the resultant GGM in a network plot.

Regularized regression. Regularized regression is a type of regression that shrinks the coefficient estimates towards zero, providing less complex and flexible models but avoiding the risk of overfitting. POMAShiny offers three different regularization strategies: LASSO, ridge regression and elasticnet.

LASSO (least absolute shrinkage and selection operator) is a regression analysis method that sets some coefficients to zero, providing more compact and interpretable models [40, 41]. Because of that, LASSO is a very good approach for the statistical analysis of metabolomics and proteomics data, both in terms of feature selection and prediction model performance.

POMAShiny provides a function based on the glmnet R package [42] that allows users to create LASSO logistic regression models (two-group analysis) both for feature selection and prediction model purposes. If the purpose is not predictive, users can set the *test* set parameter to zero and the function will return interactive plots and tables referring to the LASSO model created using all samples of the study. Otherwise, if the purpose is to build a predictive model, users can select the proportion of samples that will be used as a *test* set. In the second case, POMAShiny will fit a LASSO model without using the *test* set and using it only to perform an external validation, providing users with numerous real prediction metrics (accuracy, accuracy confidence intervals, sensitivity, specificity, etc.). Alternatively, users can also perform elasticnet (defining a penalty parameter) and ridge regression models in the same tab. For all regularized regression strategies provided in POMAShiny, the lambda parameter is chosen automatically through internal k-fold cross-validation.

Random forest. In recent years, machine learning algorithms such as random forest have become very common in the analysis of omics data. These algorithms are constantly used both to rank the importance of features and to create prediction models. POMAShiny provides a classification random forest algorithm [43] designed for the creation of prediction models to classify between two or more groups. This feature allows users to easily split data into *train* and *test* sets, where a *train* set is used to create the model and a *test* set is used only to perform an external validation. POMAShiny's random forest tab provides different tables and interactive plots with model metrics and the importance of features in the classification. In addition, the

classification algorithm provided by POMAShiny returns the model confusion matrix and errors calculated using the *test* set, providing a real measure of model accuracy.

Odds ratio. Odds ratio (OR) calculation can be very helpful in visualizing and exploring the individual feature effects on the study outcome. POMAShiny includes an option to calculate OR based on a logistic regression model (two-group analysis). By changing the function parameters, users can easily define those features that will be included in the model and the ability to include study covariates in the model.

Rank products. The rank product is a statistical non-parametric test based on ranks of fold changes. This method has been used for several years to detect differentially expressed genes in microarray experiments [44]. However, in recent years this methodology has also become popular in other omics fields such as transcriptomics, metabolomics and proteomics [45]. POMAShiny includes an option to calculate rank products both for paired and unpaired samples with a set of customizable parameters. This function provides both tables and interactive plots showing the upregulated and downregulated features, respectively.

Help and instructions

A comprehensive manual that details all POMAShiny functionalities is provided in the "Help" panel. In addition, users can also access all parameter-specific instructions by clicking on the "Help" icon available in each panel.

Example data

POMAShiny includes two example data sets that are both freely available at `https://www.metabolomicsworkbench.org`. The example data set ST000284 consists of a targeted metabolomics three-group study and the example data set ST000336 consists of a targeted metabolomics two-group study. These two data sets allow users to explore all available functionalities in POMAShiny. Both data set documentations are available at `https://github.com/nutrimetabolomics/POMA`.

Comparison with existing tools

Currently, most metabolomics and proteomics data analyses performed via web applications are conducted using the XCMS [6], MetaboAnalyst [7], Workflow4Metabolomics (W4M) [8] and Galaxy [9] tools. Among these tools, MetaboAnalyst and W4M are the most frequently used and complete in terms of statistical analysis [7].

Detailed comparisons among W4M, MetaboAnalyst and POMAShiny are exhibited in Table 2.

In terms of visualization and exploratory data analysis, only a few plots provided in MetaboAnalyst and W4M are interactive, while POMAShiny provides a whole interactive environment for almost all provided plots, offering a wide range of visualization options. As regards the importance of exploratory data analysis, POMAShiny dedicates a whole block of the workflow specifically to this issue, including an automatic PDF exploratory report. In contrast, both MetaboAnalyst and W4M provide an independent dendrogram plot, while in POMAShiny it is integrated into the heatmap.

As shown in Table 2, POMAShiny offers several methodologies for performing the three key preprocessing steps in metabolomics and proteomics: the missing value imputation, normalization and outlier detection. Being the implementation of a

methodology for outlier detection and cleaning as part of preprocessing a significant improvement in the reproducibility of the results that other tools do not provide.

Overall, the primary strength of POMAShiny is the statistical analysis. Consequently, it is shown in Table 2 that POMAShiny provides the most commonly used statistical methods for metabolomics and proteomics data analysis and other very useful methods that the MetaboAnalyst and W4M tools do not provide (e.g., regularized regression methods, rank products), as well as the opportunity to include covariates in the analysis. The increasing complexity of experimental designs has made covariates such as sex and BMI (body mass index) have a high bias in the results. Thus, the ability to use statistical methods such as ANCOVA or limma – which combine features data with other covariates – means an improvement in the accuracy and understanding of the results.

However, while POMAShiny does not offer some of the useful methods offered in MetaboAnalyst and W4M, such as orthogonal partial least squares discriminant analysis (OPLS-DA) or support vector machine (SVM), it does provide some useful methodological alternatives that these tools do not provide. These methodologies are LASSO, ridge regression, elasticnet regularization, ANCOVA, limma, rank products, odds ratio calculation and GGMs as a visualization option for correlations.

Finally, another significant advantage of POMAShiny is the predictive modelling strategy found in the regularized regression methods and in the random forest algorithm. POMAShiny allows users to easily create a random *test* set to perform an external validation of the model created with the *training* set, in contrast to MetaboAnalyst and W4M, which use the entire data set to create the models. This is a remarkable advantage as the results of POMAShiny both for regularized regression and random forest strategies provide real metrics of prediction models, allowing users to evaluate the model overfitting.

Discussion

Despite the complexity of metabolomics and proteomics data, many of the most widely used web tools for statistical analysis of these data are not very versatile in terms of the input data structure and limit the analysis to a few statistical methods. POMAShiny is a web-based tool aimed at covering some of these data analysis bottlenecks, as it is a user-friendly and intuitive complementary tool that addresses some of the issues not covered by other tools. POMAShiny offers an integrated metabolomics and proteomics data analysis workflow with a wide range of possibilities both for data preprocessing and statistical analysis, including outlier detection methods, flexible exploratory data analysis operations, downloadable reports and several statistical methods, from simpler approaches such as univariate statistics to more complex methods such as regularization and machine learning prediction algorithms. This tool requires two files as an input – the target and features file – giving users the opportunity to include important study covariates in the analysis. This intuitive and powerful web interface allows users to perform an integrated data analysis in an interactive, well-documented and extremely user-friendly web environment, making the data analysis process more accessible to a wide range of researchers not so familiar with programming and/or statistical fields.

Availability and Future Directions

The POMAShiny web application is hosted at our own server, `https://webapps.nutrimetabolomics.com/POMAShiny`, and is freely available to download at the project's GitHub repository,

Table 2. Comparison of the main features of POMAShiny with Workflow4Metabolomics (W4M) and MetaboAnalyst web-based tools. Symbols used for feature evaluations with "✓" for present and "✗" for absent.

Methods		POMAShiny	W4M	MetaboAnalyst
Visualization	Heatmap	✓	✓	✓
	Scatterplot (feature-feature)	✓	✓	✗
	Correlogram	✓	✓	✓
	Gaussian graphical models	✓	✗	✗
	Density plot (samples, features)	✓	✗	✓
	Boxplot (samples, features)	✓	✗	✓
	Volcano plot	✓	✓	✓
	Histogram	✗	✓	✗
	Dendrogram	✗	✓	✓
Preprocessing	Missing value imputation	✓	✗	✓
	Normalization	✓	✓	✓
	Outlier detection/cleaning	✓	✗	✗
Statistical analysis	T-test	✓	✓	✓
	ANOVA	✓	✓	✓
	ANCOVA	✓	✗	✗
	Limma	✓	✗	✗
	Mann-Whitney U test	✓	✓	✓
	Kruskal-Wallis	✓	✓	✓
	PCA	✓	✓	✓
	k-means	✓	✗	✓
	Multidimensional scaling	✓	✗	✗
	PLS-DA	✓	✓	✓
	OPLS(-DA)	✗	✓	✓
	sPLS-DA	✓	✗	✓
	Pearson's correlation	✓	✓	✓
	Spearman's correlation	✓	✓	✓
	Kendall's correlation	✓	✗	✓
	LASSO regression	✓	✗	✗
	Ridge regression	✓	✗	✗
	Elasticnet regression	✓	✗	✗
	Random forest	✓	✓	✓
	Support vector machine	✗	✓	✓
	Empirical bayesian analysis	✗	✗	✓
	Odds ratio calculation	✓	✗	✗
	Rank products	✓	✗	✗

`https://github.com/nutrimetabolomics/POMAShiny`, where we also use the GitHub Issues tab as a discussion channel. Additionally, users can also download the POMAShiny Docker container from DockerHub. S1 Code contains source code, documentation and the Dockerfile of POMAShiny. S2 Code contains the POMA R/Bioconductor package source code and test data sets provided in POMAShiny.

POMAShiny is an open-source project that can be readily used and enhanced by the scientific community. Further expansion of the tool to cover more preprocessing and statistical methods will certainly increase its utility. The upcoming software enhancements will be directed at implementing new statistical methods, especially focused on machine learning algorithms, to enable more diverse and robust predictive abilities in the application. In addition, new visualization methods and different statistical analysis automatic reports will also be implemented.

Supporting Information

S1 Code. Source code files and POMAShiny documentation. In addition to the source code, the archive file contains the documentation for the installation and usage of the app and the Dockerfile to create a Docker image of POMAShiny.
(ZIP)

S2 Code. Source code files and POMA Bioconductor package documentation. In addition to the source code, the archive file contains the documentation of POMA and test data sets provided in POMAShiny.
(ZIP)

Acknowledgements

We would like to acknowledge Dr Magalí Palau for her enthusiastic and valuable contribution at the start of this scientific development.

References

1. Winkler R. Processing Metabolomics and Proteomics Data with Open Software: A Practical Guide. vol. 8. Royal Society of Chemistry; 2020.
2. Wishart DS. Emerging applications of metabolomics in drug discovery and precision medicine. Nature reviews Drug discovery. 2016;15(7):473.
3. Jiang Y, Sun A, Zhao Y, Ying W, Sun H, Yang X, et al. Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. Nature. 2019;567(7747):257–261.
4. Stanstrup J, Broeckling CD, Helmus R, Hoffmann N, Mathé E, Naake T, et al. The metaRbolomics Toolbox in Bioconductor and beyond. Metabolites. 2019;9(10):200.
5. Gardinassi LG, Xia J, Safo SE, Li S. Bioinformatics tools for the interpretation of metabolomics data. Current Pharmacology Reports. 2017;3(6):374–383.
6. Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G. XCMS Online: a web-based platform to process untargeted metabolomic data. Analytical chemistry. 2012;84(11):5035–5039.

7. Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G, et al. MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. Nucleic acids research. 2018;46(W1):W486–W494.

8. Giacomoni F, Le Corguillé G, Monsoor M, Landi M, Pericard P, Pétéra M, et al. Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. Bioinformatics. 2015;31(9):1493–1495.

9. Davidson RL, Weber RJ, Liu H, Sharma-Oates A, Viant MR. Galaxy-M: A Galaxy workflow for processing and analyzing direct infusion and liquid chromatography mass spectrometry-based metabolomics data. GigaScience. 2016;5(1):s13742–016.

10. R Core Team. R: A Language and Environment for Statistical Computing; 2019. Available from: `https://www.R-project.org/`.

11. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome biology. 2004;5(10):R80.

12. Chang W, Cheng J, Allaire J, Xie Y, McPherson J. shiny: Web Application Framework for R; 2020. Available from: `https://CRAN.R-project.org/package=shiny`.

13. Castellano-Escuder P, González-Domínguez R, Andrés-Lacueva C, Sánchez-Pla A. POMA: User-friendly Workflow for Pre-processing and Statistical Analysis of Mass Spectrometry Data; 2020. Available from: `http://www.bioconductor.org/packages/release/bioc/html/POMA.html`.

14. Gatto L, Lilley K. MSnbase - an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. Bioinformatics. 2012;28:288–289.

15. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the tidyverse. Journal of Open Source Software. 2019;4(43):1686. doi:10.21105/joss.01686.

16. Granjon D. bs4Dash: A 'Bootstrap 4' Version of 'shinydashboard'; 2019. Available from: `https://CRAN.R-project.org/package=bs4Dash`.

17. Wickham H. testthat: Get Started with Testing. The R Journal. 2011;3:5–10.

18. Hester J. covr: Test Coverage for Packages; 2020. Available from: `https://CRAN.R-project.org/package=covr`.

19. Wei R, Wang J, Su M, Jia E, Chen S, Chen T, et al. Missing value imputation approach for mass spectrometry-based metabolomics data. Scientific reports. 2018;8(1):1–10.

20. Steuer R, Morgenthal K, Weckwerth W, Selbig J. A gentle guide to the analysis of metabolomic data. In: Metabolomics. Springer; 2007. p. 105–126.

21. Armitage EG, Godzien J, Alonso-Herranz V, López-Gonzálvez Á, Barbas C. Missing value imputation strategies for metabolomics data. Electrophoresis. 2015;36(24):3050–3060.

22. Hastie T, Tibshirani R, Narasimhan B, Chu G. impute: Imputation for microarray data; 2019.

23. Turck CW, Mak TD, Goudarzi M, Salek RM, Cheema AK. The ABRF Metabolomics Research Group 2016 Exploratory Study: Investigation of Data Analysis Methods for Untargeted Metabolomics. Metabolites. 2020;10(4):128.

24. van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC genomics. 2006;7(1):142.

25. Li B, Tang J, Yang Q, Li S, Cui X, Li Y, et al. NOREVA: normalization and evaluation of MS-based metabolomics data. Nucleic acids research. 2017;45(W1):W162–W170.

26. Godzien J, Ciborowski M, Angulo S, Barbas C. From numbers to a biological sense: H ow the strategy chosen for metabolomics data treatment may affect final results. A practical example based on urine fingerprints obtained by LC-MS. Electrophoresis. 2013;34(19):2812–2826.

27. Gregori J, Sanchez A, Villanueva J. msmsEDA: Exploratory Data Analysis of LC-MS/MS data by spectral counts; 2020.

28. Luo J, Schumacher M, Scherer A, Sanoudou D, Megherbi D, Davison T, et al. A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. The pharmacogenomics journal. 2010;10(4):278–291.

29. Gregori J, Villarreal L, Méndez O, Sánchez A, Baselga J, Villanueva J. Batch effects correction improves the sensitivity of significance tests in spectral counting-based comparative discovery proteomics. Journal of Proteomics. 2012;75(13):3938–3951.

30. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research. 2015;43(7):e47–e47.

31. Worley B, Powers R. Multivariate analysis in metabolomics. Current Metabolomics. 2013;1(1):92–107.

32. Rohart F, Gautier B, Singh A, Cao KAL. mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS computational biology. 2017;13(11):e1005752.

33. Xia J, Psychogios N, Young N, Wishart DS. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic acids research. 2009;37(suppl_2):W652–W660.

34. Lê Cao KA, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC bioinformatics. 2011;12(1):253.

35. Lê Cao KA, Rossouw D, Robert-Granié C, Besse P. A sparse PLS for variable selection when integrating omics data. Statistical applications in genetics and molecular biology. 2008;7(1).

36. Steinley D. K-means clustering: a half-century synthesis. British Journal of Mathematical and Statistical Psychology. 2006;59(1):1–34.

37. Hout MC, Papesh MH, Goldinger SD. Multidimensional scaling. Wiley Interdisciplinary Reviews: Cognitive Science. 2013;4(1):93–103.

38. Franzese M, Iuliano A. Correlation analysis. Encyclopedia of Bioinformatics and Computational Biology. 2019;1:706–721.

39. Friedman J, Hastie T, Tibshirani R. glasso: Graphical Lasso: Estimation of Gaussian Graphical Models; 2019. Available from: `https://CRAN.R-project.org/package=glasso`.

40. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267–288.

41. Vaarhorst AA, Verhoeven A, Weller CM, Böhringer S, Göraler S, Meissner A, et al. A metabolomic profile is associated with the risk of incident coronary heart disease. American heart journal. 2014;168(1):45–52.

42. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010;33(1):1–22.

43. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18–22.

44. Hong, F , Breitling, R , McEntee, W C , et al. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. Bioinformatics. 2006;22:2825–2827.

45. Del Carratore, F , Jankevics, A , Eisinga, R , et al. RankProd 2.0: a refactored Bioconductor package for detecting differentially expressed features in molecular profiling datasets. Bioinformatics. 2017;33:2774–2775.

A.2 Thesis software

This section includes two use cases of the software developed in the context of this thesis. The first use case corresponds to the POMA package and the second one to the fobitools package, respectively.

A.2.1 POMA use case

Use case: POMA workflow

Pol Castellano-Escuder, Cristina Andrés-Lacueva and Alex Sánchez-Pla

May, 2021

Contents

1 Installation **2**

2 Load POMA **2**

3 The POMA Workflow **2**

3.1 Data Preparation . . . 2

3.1.1 Brief Description of Example Data . . . 3

3.2 Pre Processing . . . 3

3.2.1 Missing Value Imputation . . . 3

3.2.2 Normalization . . . 4

3.2.2.1 Normalization effect . . . 4

3.2.3 Outlier Detection . . . 8

3.3 Statistical Analysis . . . 9

3.3.1 Univariate Analysis . . . 10

3.3.1.1 T-test . . . 10

3.3.1.2 Wilcoxon Test . . . 11

3.3.2 Limma . . . 11

3.3.3 Multivariate Analysis . . . 12

3.3.3.1 Principal Component Analysis . . . 12

3.3.3.2 PLS-DA . . . 13

3.3.4 Correlation Analysis . . . 15

3.3.5 Lasso, Ridge and Elasticnet . . . 16

3.3.6 Random Forest . . . 17

4 Session Information **19**

References **20**

Compiled date: 2021-05-31

Last edited: 2021-04-15

1 Installation

Run the following code to install the Bioconductor version of package.

```
# install.packages("BiocManager")
BiocManager::install("POMA")
```

2 Load POMA

```
library(POMA)
```

You can also load some additional packages that will be very useful in this vignette.

```
library(ggplot2)
library(ggraph)
library(plotly)
```

3 The POMA Workflow

`POMA` functions can be divided in three sequential well separated blocks: **Data Preparation**, **Preprocessing** and **Statistical Analysis**.

3.1 Data Preparation

The **MSnbase** Bioconductor package provides a well defined computational data structures to represent mass spectrometry (MS) experiment data types (Gatto and Lilley 2012)(Huber et al. 2015). Since data structures can mean a marked improvement in data analysis, **POMA** functions use **MSnSet** objects from **MSnbase** package, allowing the reusability of existing methods for this class and contributing to the improvement of robust and reproducible workflows.

The first step of workflow will be load or create an `MSnbase::MSnSet()` object. Often, you will have your data stored in separated matrix and/or data frames and you will have to create your own `MSnSet` object. `PomaMSnSetClass` function makes this step fast and easy building this `MSnSet` object from your independent files.

```
# create an MSnSet object from two separated data frames
target <- readr::read_csv("your_target.csv")
features <- readr::read_csv("your_features.csv")

data <- PomaMSnSetClass(target = target, features = features)
```

Alternatively, if your data is already stored in a `MSnSet` object, you can skip this step and go directly to the Pre-processing step. In this vignette we will use the example data provided in the package.

```
# load example data
data("st000336")
st000336
> MSnSet (storageMode: lockedEnvironment)
> assayData: 31 features, 57 samples
>   element names: exprs
> protocolData: none
> phenoData
>   sampleNames: DMD004.1.U02 DMD005.1.U02 ... DMD173.1.U02 (57 total)
>   varLabels: group steroids
>   varMetadata: labelDescription
> featureData: none
> experimentData: use 'experimentData(object)'
> Annotation:
> - - - Processing information - - -
>  MSnbase version: 2.12.0
```

3.1.1 Brief Description of Example Data

This example data is composed of 57 samples, 31 metabolites, 1 covariate and 2 experimental groups (Controls and DMD) from a targeted LC/MS study.

Duchenne Muscular Dystrophy (DMD) is an X-linked recessive form of muscular dystrophy that affects males via a mutation in the gene for the muscle protein, dystrophin. Progression of the disease results in severe muscle loss, ultimately leading to paralysis and death. Steroid therapy has been a commonly employed method for reducing the severity of symptoms. This study aims to quantify the urine levels of amino acids and organic acids in patients with DMD both with and without steroid treatment. Track the progression of DMD in patients who have provided multiple urine samples.

This data was collected from here.

3.2 Pre Processing

This is a critical point in the workflow because all final statistical results will depend on the decisions made here. Again, this block can be divided in 3 steps: **Missing Value Imputation**, **Normalization** and **Outlier Detection**.

3.2.1 Missing Value Imputation

Often, due to biological and technical reasons, some features can not be identified or quantified in some samples in MS (Armitage et al. 2015). **POMA** offers 7 different imputation methods to deal with this situation. Just run the following line of code to impute your missings!

```
imputed <- PomaImpute(st000336, ZerosAsNA = TRUE, RemoveNA = TRUE, cutoff = 20, method = "knn")
imputed
> MSnSet (storageMode: lockedEnvironment)
> assayData: 30 features, 57 samples
>   element names: exprs
> protocolData: none
```

```
> phenoData
>   sampleNames: DMD004.1.U02 DMD005.1.U02 ... DMD173.1.U02 (57 total)
>   varLabels: group steroids
>   varMetadata: labelDescription
> featureData: none
> experimentData: use 'experimentData(object)'
> Annotation:
> - - - Processing information - - -
> Imputed (knn): Mon May 31 14:08:14 2021
>  MSnbase version: 2.16.1
```

Note that the object has been updated with imputation information.

3.2.2 Normalization

The next step of this block is the data normalization. Often, some factors can introduce variability in some types of MS data having a critical influence on the final statistical results, making normalization a key step in the workflow (Berg et al. 2006). Again, **POMA** offers several methods to normalize the data by running just the following line of code:

```
normalized <- PomaNorm(imputed, method = "log_pareto")
normalized
> MSnSet (storageMode: lockedEnvironment)
> assayData: 30 features, 57 samples
>   element names: exprs
> protocolData: none
> phenoData
>   sampleNames: DMD004.1.U02 DMD005.1.U02 ... DMD173.1.U02 (57 total)
>   varLabels: group steroids
>   varMetadata: labelDescription
> featureData: none
> experimentData: use 'experimentData(object)'
> Annotation:
> - - - Processing information - - -
> Imputed (knn): Mon May 31 14:08:14 2021
> Normalized (log_pareto): Mon May 31 14:08:14 2021
>  MSnbase version: 2.16.1
```

Note that the object has been updated with normalization information.

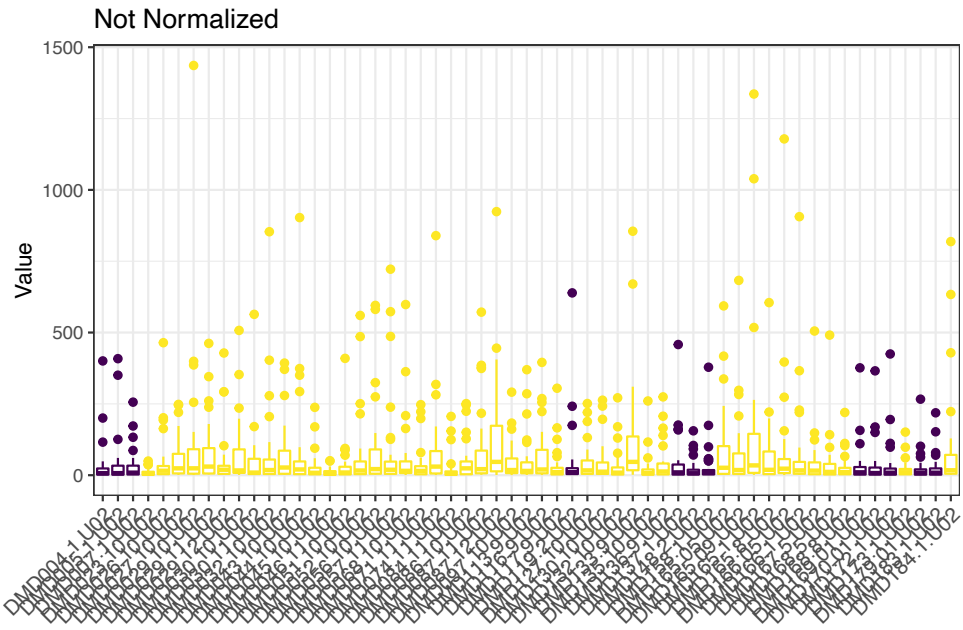
3.2.2.1 Normalization effect

Sometimes, you will be interested in *how the normalization process affect your data*?

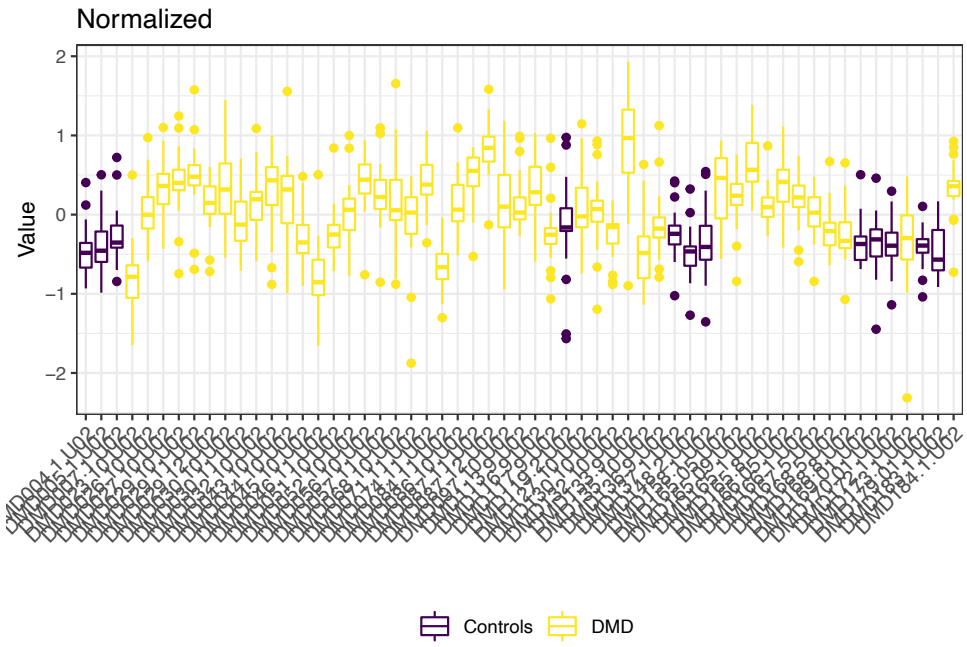
To answer this question, **POMA** offers two exploratory functions, `PomaBoxplots` and `PomaDensity`, that can help to understand the normalization process.

`PomaBoxplots` generates boxplots for all samples or features (depending on the group factor) of an `MSnSet` object. Here, we can compare objects before and after normalization step.

```
PomaBoxplots(imputed, group = "samples", jitter = FALSE) +
  ggtitle("Not Normalized") +
  theme(legend.position = "none") # data before normalization
```

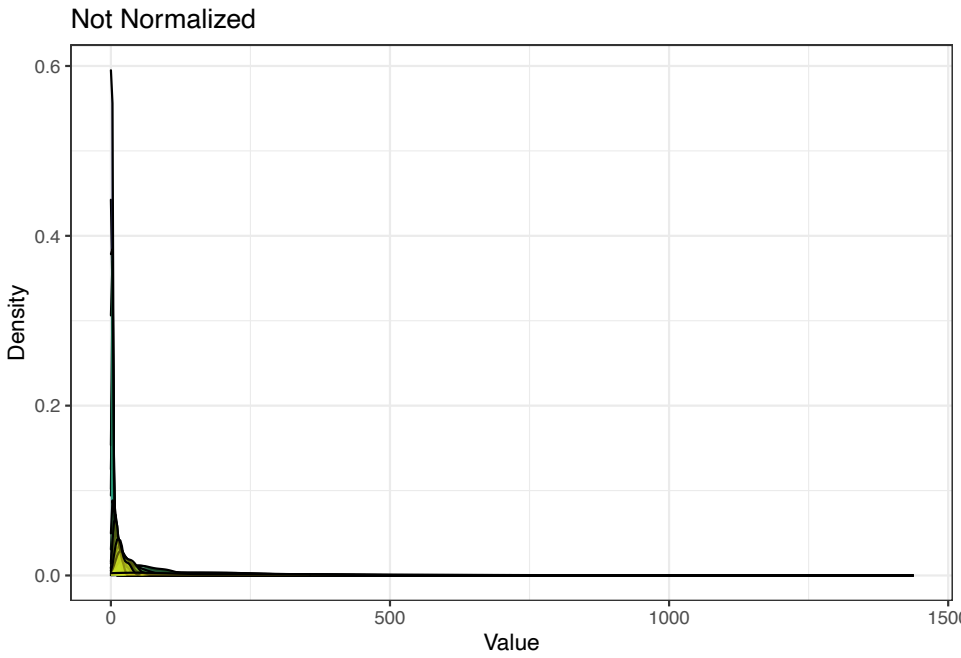



```
PomaBoxplots(normalized, group = "samples", jitter = FALSE) +
  ggtitle("Normalized") # data after normalization
```

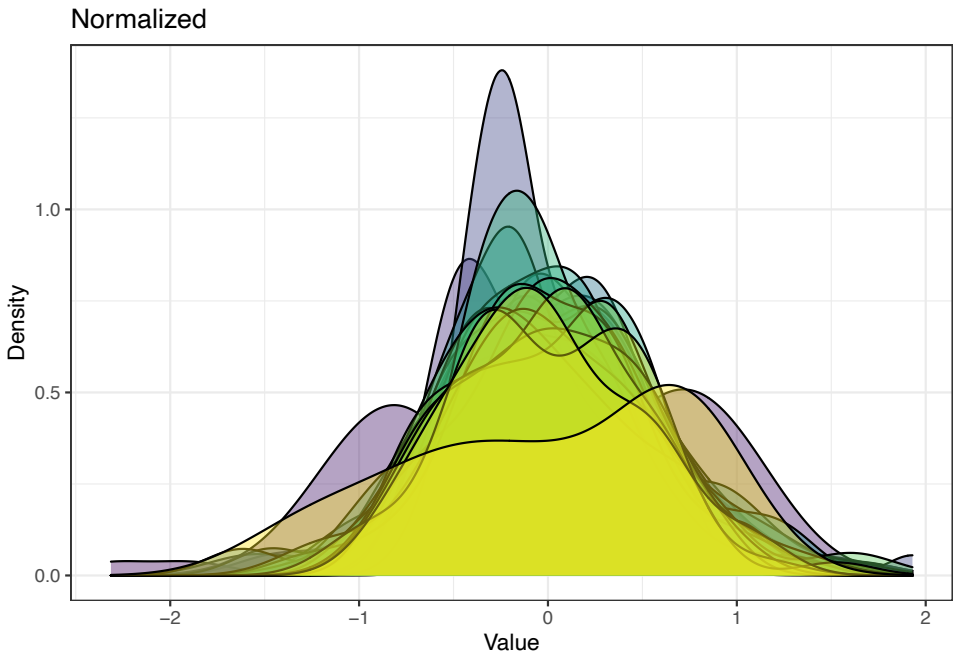


On the other hand, `PomaDensity` shows the distribution of all features before and after the normalization process.

```
PomaDensity(imputed, group = "features") +
  ggtitle("Not Normalized") +
  theme(legend.position = "none") # data before normalization
```




```
PomaDensity(normalized, group = "features") +
  ggtitle("Normalized") # data after normalization
```

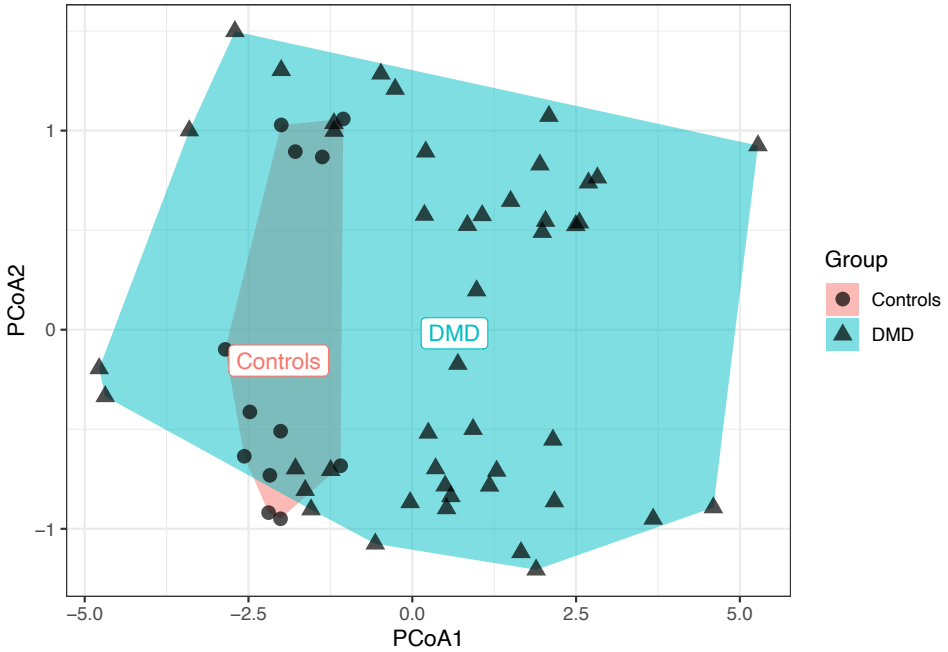


3.2.3 Outlier Detection

Finally, the last step of this block is the Outlier Detection. Outlers are defined as observations that are not concordant with those of the vast majority of the remaining data points. These values can have an enormous influence on the resultant statistical analysis, being a dangerous ground for all required assumptions in the most commonly applied parametric tests in mass spectrometry as well as for all also required assumptions in many regression techniques and predictive modeling approaches. **POMA** allows the analysis of outliers as well as the possibility to remove them from the analysis using different modulable parameters.

Analyze and remove outliers running the following two lines of code.

```
PomaOutliers(normalized, do = "analyze")$polygon_plot # to explore
```




```
pre_processed <- PomaOutliers(normalized, do = "clean") # to remove outliers
pre_processed
> MSnSet (storageMode: lockedEnvironment)
> assayData: 30 features, 50 samples
>   element names: exprs
> protocolData: none
> phenoData
>   sampleNames: DMD004.1.U02 DMD005.1.U02 ... DMD173.1.U02 (50 total)
>   varLabels: group steroids
>   varMetadata: labelDescription
> featureData: none
> experimentData: use 'experimentData(object)'
> Annotation:
> - - - Processing information - - -
> Imputed (knn): Mon May 31 14:08:14 2021
> Normalized (log_pareto): Mon May 31 14:08:14 2021
> Outliers removed (euclidean and median): Mon May 31 14:08:18 2021
>  MSnbase version: 2.16.1
```

Note that the object has been updated with outlier information.

3.3 Statistical Analysis

Once the data have been preprocessed, you can start with the statistical analysis block! **POMA** offers many different statistical methods and possible combinations to compute. However, in this vignette we will comment only some of the most used.

3.3.1 Univariate Analysis

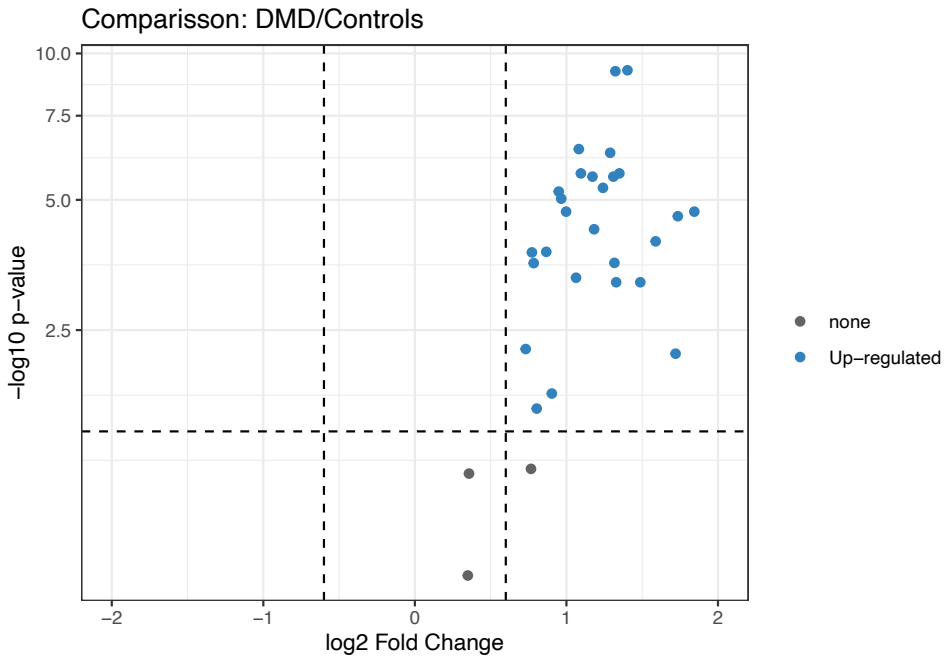
POMA allows you to perform all of the most used univariate statistical methods in MS by using only one function! `PomaUnivariate` wrap 4 different univariate methods (ttest, ANOVA and ANCOVA, Wilcoxon test and Kruskal-Wallis Rank Sum Test) that you can perform changing only the "method" argument.

```
PomaUnivariate(pre_processed, method = "ttest") %>%
  head()
>                    mean_Controls   mean_DMD Fold_Change_Ratio
> x1_methylhistidine    -0.4018182 0.16112821            -0.401
> x3_methylhistidine    -0.4195455 0.19305128            -0.460
> alanine               -0.3165455 0.10946154            -0.346
> arginine              -0.1148182 0.06405128            -0.558
> asparagine            -0.3475455 0.12764103            -0.367
> aspartic_acid         -0.2542727 0.08887179            -0.350
>                    Difference_Of_Means       pvalue    pvalueAdj
> x1_methylhistidine               0.563 9.302122e-08 3.100707e-07
> x3_methylhistidine               0.613 7.822367e-03 9.025808e-03
> alanine                          0.426 6.617797e-04 8.631909e-04
> arginine                         0.179 4.796275e-01 4.796275e-01
> asparagine                       0.475 1.279748e-05 2.399528e-05
> aspartic_acid                    0.343 3.302402e-02 3.538287e-02
```

3.3.1.1 T-test

You can also compute a volcano plot using the T-test results. *Note that we're using the non-normalized object to avoid negative values in our data.*

```
PomaVolcano(imputed, pval = "adjusted")
> Warning in PomaVolcano(imputed, pval = "adjusted"): adjust argument is empty!
> FDR will be used
```




```
PomaUnivariate(pre_processed, method = "mann") %>%
  head()
>                    mean_Controls   mean_DMD Fold_Change_Ratio
> x1_methylhistidine    -0.4018182 0.16112821            -0.401
> x3_methylhistidine    -0.4195455 0.19305128            -0.460
> alanine               -0.3165455 0.10946154            -0.346
> arginine              -0.1148182 0.06405128            -0.558
> asparagine            -0.3475455 0.12764103            -0.367
> aspartic_acid         -0.2542727 0.08887179            -0.350
>                    Difference_Of_Means       pvalue    pvalueAdj
> x1_methylhistidine               0.563 6.206369e-05 0.0001432239
> x3_methylhistidine               0.613 9.989212e-03 0.0115260136
> alanine                          0.426 2.152908e-04 0.0004036702
> arginine                         0.179 1.400885e-01 0.1449191650
> asparagine                       0.475 2.468467e-04 0.0004356118
> aspartic_acid                    0.343 5.709276e-03 0.0068511312
```

3.3.1.2 Wilcoxon Test

3.3.2 Limma

Other of the wide used statistical methods in many different omics, such as epigenomics or transcriptomics, is **limma** (Ritchie et al. 2015). **POMA** provides an easy use implementation of *limma* you only have to

specify the desired contrast to compute.

```
PomaLimma(pre_processed, contrast = "Controls-DMD", adjust = "fdr") %>%
  head()
>                logFC AveExpr         t      P.Value    adj.P.Val         B
> tryptophan -0.7749207 0.00862 -7.006568 2.691438e-09 8.074315e-08 10.982005
> valine     -0.7009883 0.01268 -6.631334 1.155051e-08 1.732576e-07  9.553920
> ornithine  -0.6327809 0.03366 -6.262549 4.794292e-08 4.794292e-07  8.160589
> isoleucine -0.6058485 0.00438 -5.948984 1.591994e-07 1.193995e-06  6.987716
> lactate    -0.7853613 0.01840 -5.687122 4.296261e-07 2.577757e-06  6.019252
> pyruvate   -0.6244615 0.01208 -5.432256 1.117445e-06 5.135059e-06  5.088552
```

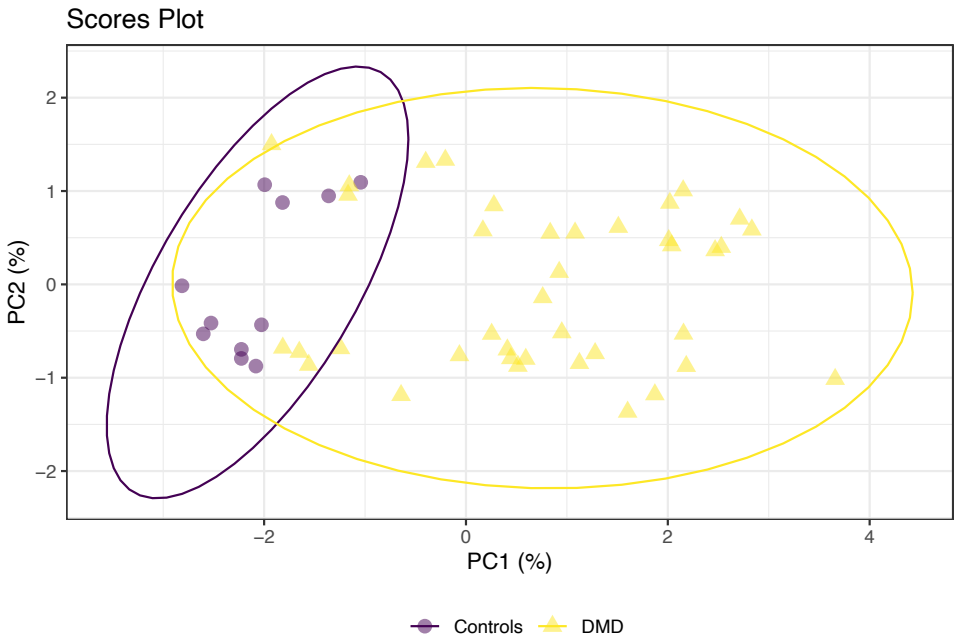
3.3.3 Multivariate Analysis

On the other hand, multivariate analysis implemented in **POMA** is quite similar to the univariate approaches. `PomaMultivariate` allows users to compute a PCA, PLS-DA or sPLS-DA by changing only the "method" parameter. This function is based on **mixOmics** package (Rohart et al. 2017).

```
poma_pca <- PomaMultivariate(pre_processed, method = "pca")
```

```
poma_pca$scoresplot +
  ggtitle("Scores Plot")
```

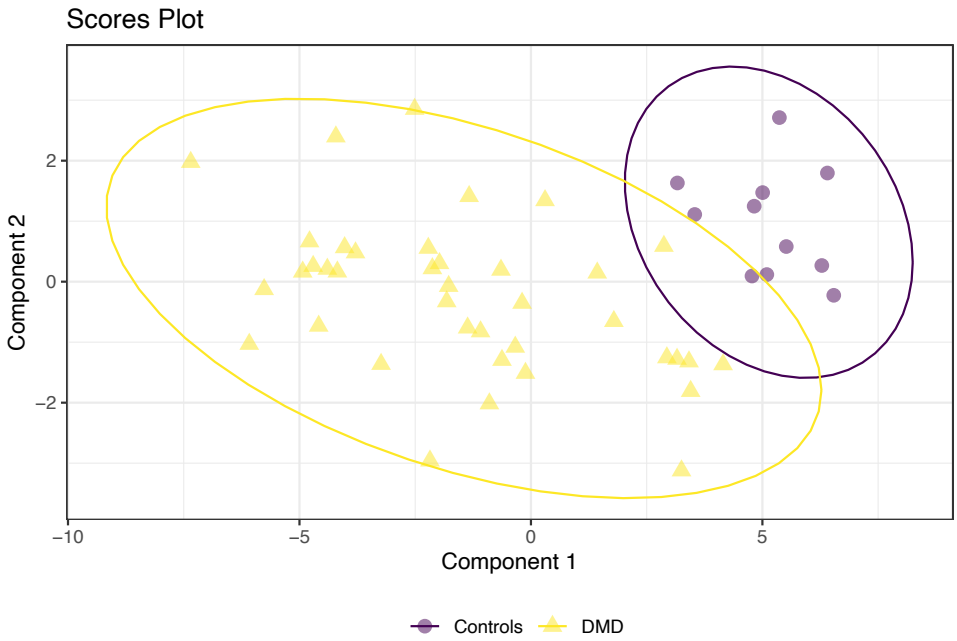
3.3.3.1 Principal Component Analysis



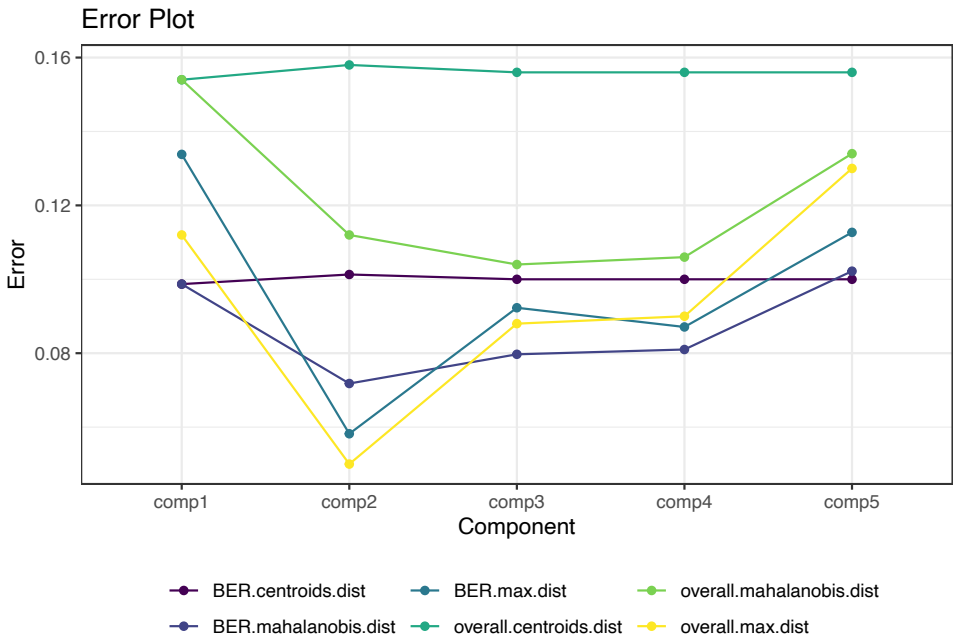

```
poma_plsda <- PomaMultivariate(pre_processed, method = "plsda")
```

```
poma_plsda$scoresplot +
  ggtitle("Scores Plot")
```

3.3.3.2 PLS-DA



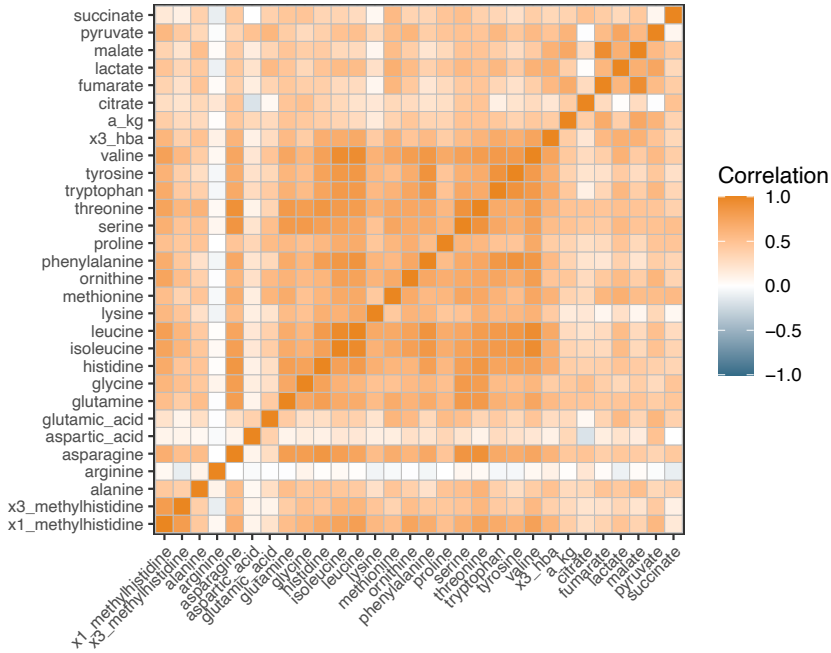

```
poma_plsda$errors_plsda_plot +
  ggtitle("Error Plot")
```



3.3.4 Correlation Analysis

Often, correlation analysis is used to explore and discover relationships and patterns within our data. `PomaCorr` provides a flexible and easy way to do that providing a table with all pairwise correlations in the data, a correlogram and a correlation graph.

```
poma_cor <- PomaCorr(pre_processed, label_size = 8, coeff = 0.6)
poma_cor$correlations %>% head()
>           Var1      Var2      corr
> 341 isoleucine   leucine 0.9631456
> 642    leucine    valine 0.9405392
> 836   fumarate    malate 0.9398663
> 641 isoleucine    valine 0.9378129
> 545 asparagine threonine 0.9067625
> 558     serine threonine 0.8931187
poma_cor$corrplot
```



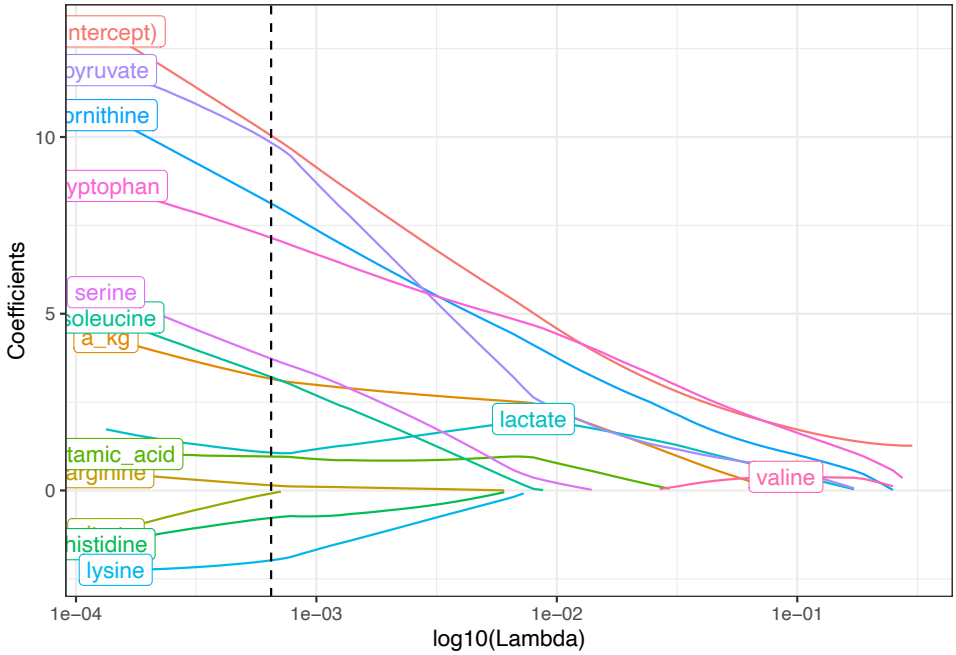
Alternatively, if you switch the "corr_type" parameter to "glasso", this function will compute a **Gaussian Graphical Model** using the **glmnet** package (Friedman, Hastie, and Tibshirani 2019).

```
PomaCorr(pre_processed, corr_type = "glasso", coeff = 0.6)$graph
```

3.3.5 Lasso, Ridge and Elasticnet

POMA also provides a function to perform a Lasso, Ridge and Elasticnet regression for binary outcomes in a very intuitive and easy way. `PomaLasso` is based on **glmnet** package (Friedman, Hastie, and Tibshirani 2010). This function allows you to create a test subset in your data, evaluate the prediction of your models and export the model computed (it could be useful to perform prediction models with MS data). If "ntest" parameter is set to NULL, `PomaLasso` will use all observations to create the model (useful for feature selection).

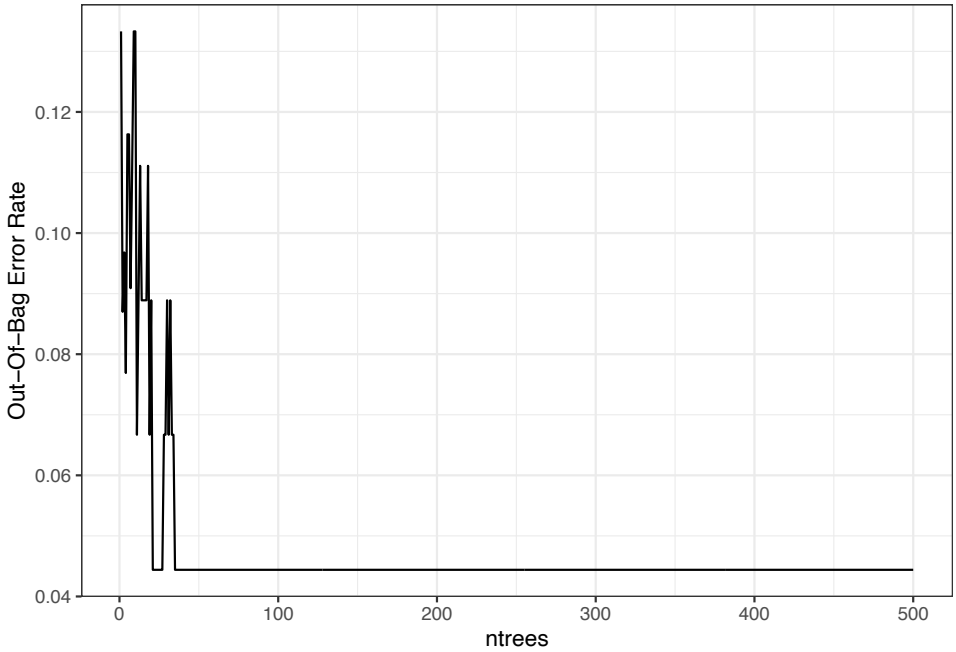
```
# alpha = 1 for Lasso
PomaLasso(pre_processed, alpha = 1, labels = TRUE)$coefficientPlot
```



3.3.6 Random Forest

Finally, the random forest algorithm is also implemented in **POMA**. `PomaRandForest` uses the **randomForest** package (Liaw and Wiener 2002) to facilitate the implementation of the algorithm and creates automatically both test and train sets to compute and evaluate the resultant models.

```
poma_rf <- PomaRandForest(pre_processed, ntest = 10, nvar = 10)
poma_rf$error_tree
```

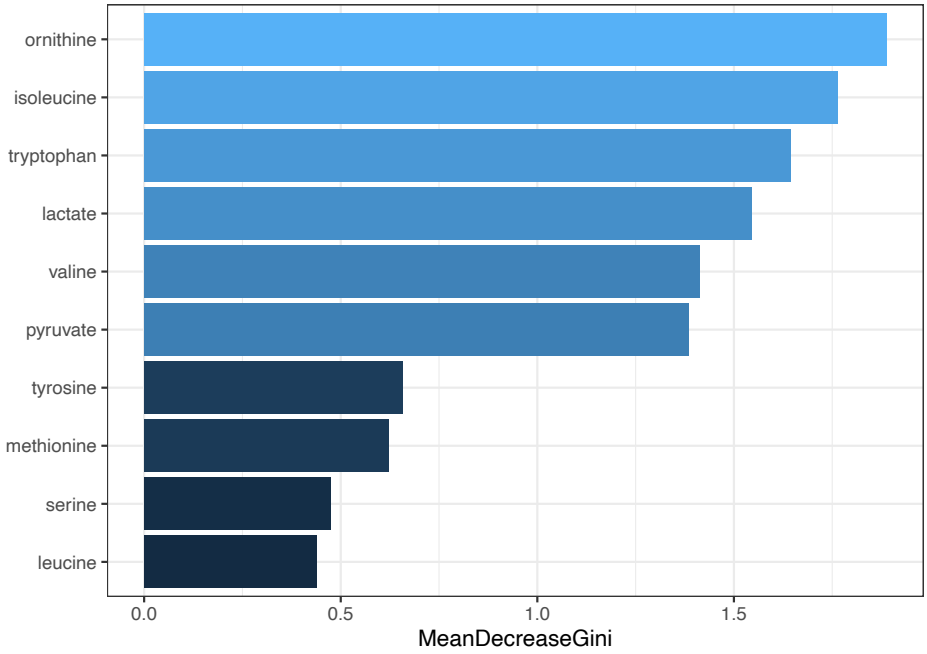


Resultant random forest model confusion matrix for **test** set:

```
poma_rf$confusion_matrix
>   1 2 class.error
> 1 1 0           0
> 2 0 4           0
```

Gini index plot for the top 10 predictors:

```
poma_rf$gini_plot
```



4 Session Information

```
sessionInfo()
> R version 4.0.2 (2020-06-22)
> Platform: x86_64-apple-darwin17.0 (64-bit)
> Running under: macOS  10.16
>
> Matrix products: default
> BLAS:   /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
> LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib
>
> locale:
> [1] es_ES.UTF-8/es_ES.UTF-8/es_ES.UTF-8/C/es_ES.UTF-8/es_ES.UTF-8
>
> attached base packages:
> [1] stats     graphics  grDevices utils     datasets  methods   base
>
> other attached packages:
> [1] plotly_4.9.3  ggraph_2.0.5  ggplot2_3.3.3 POMA_1.1.13
>
> loaded via a namespace (and not attached):
>   [1] backports_1.2.1       circlize_0.4.12       plyr_1.8.6
>   [4] igraph_1.2.6          lazyeval_0.2.2        splines_4.0.2
>   [7] gmp_0.6-2             BiocParallel_1.24.1   digest_0.6.27
```

```
>  [10] foreach_1.5.1          htmltools_0.5.1.1     viridis_0.6.1
>  [13] fansi_0.4.2            magrittr_2.0.1        cluster_2.1.2
>  [16] doParallel_1.0.16      limma_3.46.0          recipes_0.1.16
>  [19] ComplexHeatmap_2.6.2   graphlayouts_0.7.1    gower_0.2.2
>  [22] matrixStats_0.58.0     rARPACK_0.11-0        colorspace_2.0-1
>  [25] ggrepel_0.9.1          xfun_0.22             dplyr_1.0.6
>  [28] jsonlite_1.7.2         crayon_1.4.1          impute_1.64.0
>  [31] survival_3.2-11        iterators_1.0.13      glue_1.4.2
>  [34] polyclip_1.10-0        gtable_0.3.0          ipred_0.9-11
>  [37] zlibbioc_1.36.0        GetoptLong_1.0.5      RankProd_3.16.0
>  [40] shape_1.4.5            Rmpfr_0.8-4           BiocGenerics_0.36.1
>  [43] scales_1.1.1           vsn_3.58.0            DBI_1.1.1
>  [46] Rcpp_1.0.6             mzR_2.24.1            viridisLite_0.4.0
>  [49] clue_0.3-59            proxy_0.4-25          preprocessCore_1.52.1
>  [52] clisymbols_1.2.0       stats4_4.0.2          lava_1.6.9
>  [55] prodlim_2019.11.13     glmnet_4.1-1          httr_1.4.2
>  [58] htmlwidgets_1.5.3      RColorBrewer_1.1-2    ellipsis_0.3.2
>  [61] pkgconfig_2.0.3        XML_3.99-0.6          farver_2.1.0
>  [64] nnet_7.3-16            utf8_1.2.1            caret_6.0-86
>  [67] labeling_0.4.2         tidyselect_1.1.1      rlang_0.4.11
>  [70] reshape2_1.4.4         ggcorrplot_0.1.3      munsell_0.5.0
>  [73] tools_4.0.2            generics_0.1.0        broom_0.7.6
>  [76] evaluate_0.14          stringr_1.4.0         mzID_1.28.0
>  [79] yaml_2.2.1             ModelMetrics_1.2.2.2  knitr_1.33
>  [82] tidygraph_1.2.0        purrr_0.3.4           randomForest_4.6-14
>  [85] ncdf4_1.17             glasso_1.11           nlme_3.1-152
>  [88] compiler_4.0.2         png_0.1-7             e1071_1.7-6
>  [91] affyio_1.60.0          tibble_3.1.1          tweenr_1.0.2
>  [94] stringi_1.6.1          highr_0.9             RSpectra_0.16-0
>  [97] MSnbase_2.16.1         lattice_0.20-44       ProtGenerics_1.22.0
> [100] Matrix_1.3-2           vegan_2.5-7           permute_0.9-5
> [103] vctrs_0.3.8            pillar_1.6.0          lifecycle_1.0.0
> [106] BiocManager_1.30.12    MALDIquant_1.19.3     GlobalOptions_0.1.2
> [109] data.table_1.14.0      corpcor_1.6.9         patchwork_1.1.1
> [112] R6_2.5.0               pcaMethods_1.82.0     affy_1.68.0
> [115] gridExtra_2.3          IRanges_2.24.1        codetools_0.2-18
> [118] MASS_7.3-54            assertthat_0.2.1      rjson_0.2.20
> [121] withr_2.4.2            S4Vectors_0.28.1      mgcv_1.8-35
> [124] parallel_4.0.2         mixOmics_6.14.1       grid_4.0.2
> [127] rpart_4.1-15           timeDate_3043.102     tidyr_1.1.3
> [130] class_7.3-19           rmarkdown_2.7         Cairo_1.5-12.2
> [133] ggforce_0.3.3          pROC_1.17.0.1         Biobase_2.50.0
> [136] lubridate_1.7.10       ellipse_0.4.2
```

References

Armitage, Emily Grace, Joanna Godzien, Vanesa Alonso-Herranz, Ángeles López-Gonzálvez, and Coral Barbas. 2015. "Missing Value Imputation Strategies for Metabolomics Data." *Electrophoresis* 36 (24): 3050–60.

Berg, Robert A van den, Huub CJ Hoefsloot, Johan A Westerhuis, Age K Smilde, and Mariët J van der Werf. 2006. "Centering, Scaling, and Transformations: Improving the Biological Information Content of

Metabolomics Data." *BMC Genomics* 7 (1): 142.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. 2019. *Glasso: Graphical Lasso: Estimation of Gaussian Graphical Models*. https://CRAN.R-project.org/package=glasso.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22. http://www.jstatsoft.org/v33/i01/.

Gatto, Laurent, and Kathryn Lilley. 2012. "MSnbase - an R/Bioconductor Package for Isobaric Tagged Mass Spectrometry Data Visualization, Processing and Quantitation." *Bioinformatics* 28: 288–89.

Huber, W., V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, et al. 2015. "Orchestrating High-Throughput Genomic Analysis with Bioconductor." *Nature Methods* 12 (2): 115–21. http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22. https://CRAN.R-project.org/doc/Rnews/.

Ritchie, Matthew E, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. 2015. "limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." *Nucleic Acids Research* 43 (7): e47. https://doi.org/10.1093/nar/gkv007.

Rohart, Florian, Benoît Gautier, Amrit Singh, and Kim-Anh Lê Cao. 2017. "MixOmics: An R Package for 'Omics Feature Selection and Multiple Data Integration." *PLoS Computational Biology* 13 (11): e1005752. http://www.mixOmics.org.

A.2.2 fobitools use case

Use case: LC-MS Based Approaches to Investigate Metabolomic Differences in the Urine of Young Women after Drinking Cranberry Juice or Apple Juice

Pol Castellano-Escuder, Cristina Andrés-Lacueva and Alex Sánchez-Pla

May, 2021

Contents

1 Installation **2**

2 Load packages **2**

3 Download the data from Metabolomics Workbench **2**

3.1 Sumary of the study (ST000291) . . . 3

3.2 Download data . . . 3

3.3 Scraping metabolite names and identifiers with `rvest` . . . 3

4 Prepare features and metadata **5**

5 Statistical analysis with `POMA` **6**

5.1 Create a `MSnbase::MSnSet` object . . . 6

5.2 Preprocessing . . . 6

5.3 Limma model . . . 6

6 Convert PubChem IDs to FOBI IDs **6**

7 Enrichment analysis **7**

7.1 Over representation analysis (ORA) . . . 8

7.2 MSEA . . . 8

7.2.1 MSEA plot with `ggplot2` . . . 10

7.2.2 Network of metabolites found in MSEA . . . 11

8 Limitations **12**

9 Session Information **12**

References **15**

License: GPL-3

1 Installation

Run the following code to install the Bioconductor version of the package.

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("fobitools")
```

2 Load packages

```
library(fobitools)
```

We will also need some additional CRAN and Bioconductor packages for performing tasks such as statistical analysis and web scraping.

```
# CRAN
library(tidyverse)
library(rvest)
library(ggrepel)
library(kableExtra)

# Bioconductor
library(POMA)
library(metabolomicsWorkbenchR)
library(SummarizedExperiment)
```

3 Download the data from Metabolomics Workbench

The Metabolomics Workbench, available at www.metabolomicsworkbench.org, is a public repository for metabolomics metadata and experimental data spanning various species and experimental platforms, metabolite standards, metabolite structures, protocols, tutorials, and training material and other educational resources. It provides a computational platform to integrate, analyze, track, deposit and disseminate large volumes of heterogeneous data from a wide variety of metabolomics studies including mass spectrometry (MS) and nuclear magnetic resonance spectrometry (NMR) data spanning over 20 different species covering all the major taxonomic categories including humans and other mammals, plants, insects, invertebrates and microorganisms (Sud et al. 2016).

The `metabolomicsWorkbenchR` Bioconductor package allows us to obtain data from the Metabolomics Workbench repository. In this vignette we will use the sample data set ST000291.

3.1 Sumary of the study (ST000291)

Eighteen healthy female college students between 21-29 years old with a normal BMI of 18.5-25 were recruited. Each subject was provided with a list of foods that contained significant amount of procyanidins, such as **cranberries, apples, grapes, blueberries, chocolate and plums**. They were advised to avoid these foods during the 1-6th day and the rest of the study. On the morning of the 7th day, a first-morning baseline urine sample and blood sample were collected from all human subjects after overnight fasting. **Participants were then randomly allocated into two groups (n=9) to consume cranberry juice or apple juice**. Six bottles (250 ml/bottle) of juice were given to participants to drink in the morning and evening of the 7th, 8th, and 9th day. On the morning of 10th day, all subjects returned to the clinical unit to provide a first-morning urine sample after overnight fasting. The blood sample was also collected from participants 30 min later after they drank another bottle of juice in the morning. After two-weeks of wash out period, participants switched to the alternative regimen and repeated the protocol. One human subject was dropped off this study because she missed part of her appointments. Another two human subjects were removed from urine metabolomics analyses because they failed to provide required urine samples after juice drinking.The present study aimed to investigate overall metabolic changes caused by procyanidins concentrates from cranberries and apples using a global LCMS based metabolomics approach. All plasma and urine samples were stored at -80ºC until analysis.

3.2 Download data

This study is composed of two complementary MS analyses, the positive mode (AN000464) and the negative mode (AN000465). Let's download them both!

```
data_negative_mode <- do_query(
  context = "study",
  input_item = "analysis_id",
  input_value = "AN000465",
  output_item = "SummarizedExperiment")

data_positive_mode <- do_query(
  context = "study",
  input_item = "analysis_id",
  input_value = "AN000464",
  output_item = "SummarizedExperiment")
```

3.3 Scraping metabolite names and identifiers with `rvest`

In many metabolomics studies, the reproducibility of analyses is severely affected by the poor interoperability of metabolite names and their identifiers. For this reason it is important to develop tools that facilitate the process of converting one type of identifier to another. In order to use the `fobitools` package, we need some generic identifier (such as PubChem, KEGG or HMDB) that allows us to obtain the corresponding FOBI identifier for each metabolite. The Metabolomics Workbench repository provides us with this information for many of the metabolites quantified in study ST000291 (Figure @ref(fig:metabolitenames)). In order to easily obtain this information, we will perform a web scraping operation using the `rvest` package.

Below we obtain the PubChem and KEGG identifiers of the metabolites analyzed in the positive and negative mode directly from the Metabolomics Workbench website. We will then remove those duplicate identifiers.

```
metaboliteNamesURL <- "https://www.metabolomicsworkbench.org/data/show_metabolites_by_study.php?STUDY_ID
metaboliteNames <- metaboliteNamesURL %>%
  read_html() %>%
```

Return to study ST000291 main page

Analysis: ESI Positive mode

Show values for a selected metabolite or ratios for 2 selected metabolites

Select	Metabolite Name	RefMet Name/Standardized Name*	WorkBench Metabolite_ID	PubChem Compound_ID	Kegg Id	Retention time/index	m/z ratio
	10-Desacetyltaxuyunnanin C	-	ME104202	5460449	C15538	10.5	463.2694
	10-Hydroxydecanoic acid	10-Hydroxydecanoic acid	ME104203	74300	C02774	8.95	189.1488
	10-Oxodecanoate_1	10-Oxodecanoic acid	ME104204	19734156	C02217	8.7	187.133
	10-Oxodecanoate_2	10-Oxodecanoic acid	ME104205	19734156	C02217	7.62	187.1331
	11beta,21-Dihydroxy-5beta-pregnane-3,20-dione	11beta,21-Dihydroxy-5beta-pregnane-3,20-dione	ME104206	21145110	C05475	9.27	349.2372
	1,1-Dichloroethylene epoxide	-	ME104182	119521	C14857	0.11	112.9559
	11-Hydroxycanthin-6-one	-	ME104207	337601	C09212	8.07	237.0661
	11-O-Demethylpradimicinone I	-	ME104208	441170	C06776	0.97	536.119
	1-(1-Oxopropyl)-1H-imidazole	-	ME104176	77770	C03007	0.79	125.0712
	1,2,3-Trihydroxybenzene_1	1,2,3-Trihydroxybenzene	ME104183	1057	C01108	7.42	127.0393
	1,2,3-Trihydroxybenzene_2	1,2,3-Trihydroxybenzene	ME104184	1057	C01108	3.19	127.0393
	1,2,3-Trihydroxybenzene_3	1,2,3-Trihydroxybenzene	ME104185	1057	C01108	0.93	127.0392
	1,2,6-Hexanetriol	-	ME104186	7823	C20334	7.01	135.1018
	12alpha-Hydroxyamoorstatin	-	ME104209	5462416	C08767	7.12	267.1231
	1,2-Diamino-4-nitrobenzene_1	-	ME104187	5111791	C19384	6.92	154.0611
	1,2-Diamino-4-nitrobenzene_2	-	ME104188	5111791	C19384	1.71	154.0612
	1,2-Diamino-4-nitrobenzene_3	-	ME104189	5111791	C19384	8.39	154.0614
	1,2-Dihydrosantonin_1	1,2-Dihydrosantonin	ME104190	-	C02771	10.42	249.1484
	1,2-Dihydrosantonin_2	1,2-Dihydrosantonin	ME104191	-	C02771	9.32	249.1484
	1,2-Dihydroxy-5-(methylthio)pent-1-en-3-one	1,2-Dihydroxy-3-keto-5-methylthiopentene	ME104192	5462190	C15606	3.17	163.0424
	1,2-Dihydroxydibenzothiophene	1,2-Dihydroxydibenzothiophene	ME104193	441151	C06722	6.87	217.0328
	12-Hydroxydodecanoic acid_1	12-Hydroxy lauric acid	ME104210	79034	C08317	12.69	217.1799
	12-Hydroxydodecanoic acid_2	12-Hydroxy lauric acid	ME104211	79034	C08317	10.76	217.1801
	1,3,7-Trimethyluric acid	1,3,7-Trimethyluric acid	ME104194	79437	C16361	7.62	211.0826
	1,3,8-Naphthalenertriol_1	-	ME104195	439428	C01173	8.53	177.0547
	1,3,8-Naphthalenertriol_2	-	ME104196	439428	C01173	12.53	177.0548
	1,3,8-Naphthalenertriol_3	-	ME104197	439428	C01173	9.94	177.0547
	1,3-Dichloro-2-propanol	-	ME104198	7289	C14399	2.02	128.9874

Figure 1: Metabolite identifiers of the ST000291 Metabolomics Workbench study.

```
  html_nodes(".datatable")

metaboliteNames_negative <- metaboliteNames %>%
  .[[1]] %>%
  html_table() %>%
  dplyr::select(`Metabolite Name`, PubChemCompound_ID, `Kegg Id`)

metaboliteNames_positive <- metaboliteNames %>%
  .[[2]] %>%
  html_table() %>%
  dplyr::select(`Metabolite Name`, PubChemCompound_ID, `Kegg Id`)

metaboliteNames <- bind_rows(metaboliteNames_negative, metaboliteNames_positive) %>%
  dplyr::rename(names = 1, PubChem = 2, KEGG = 3) %>%
  mutate(KEGG = ifelse(KEGG == "-", "UNKNOWN", KEGG),
         PubChem = ifelse(PubChem == "-", "UNKNOWN", PubChem)) %>%
  filter(!duplicated(PubChem))
```

4 Prepare features and metadata

Now we have to prepare the metadata and features in order to proceed with the statistical analysis. In this step we assign to each metabolite its PubChem identifier obtained in the previous step.

```
## negative mode features
features_negative <- assay(data_negative_mode) %>%
  dplyr::slice(-n())
rownames(features_negative) <- rowData(data_negative_mode)$metabolite[1:(length(rowData(data_negative_m

## positive mode features
features_positive <- assay(data_positive_mode) %>%
  dplyr::slice(-n())
rownames(features_positive) <- rowData(data_positive_mode)$metabolite[1:(length(rowData(data_positive_m

## combine positive and negative mode and set PubChem IDs as feature names
features <- bind_rows(features_negative, features_positive) %>%
  tibble::rownames_to_column("names") %>%
  right_join(metaboliteNames, by = "names") %>%
  select(-names, -KEGG) %>%
  tibble::column_to_rownames("PubChem")

## metadata
pdata <- colData(data_negative_mode) %>% # or "data_positive_mode". They are equal
  as.data.frame() %>%
  tibble::rownames_to_column("ID") %>%
  mutate(Treatment = case_when(Treatment == "Baseline urine" ~ "Baseline",
                               Treatment == "Urine after drinking apple juice" ~ "Apple",
                               Treatment == "Urine after drinking cranberry juice" ~ "Cranberry"))
```

5 Statistical analysis with POMA

POMA provides a structured, reproducible and easy-to-use workflow for the visualization, preprocessing, exploration, and statistical analysis of metabolomics and proteomics data. The main aim of this package is to enable a flexible data cleaning and statistical analysis processes in one comprehensible and user-friendly R package. POMA uses the standardized MSnbase data structures, to achieve the maximum flexibility and reproducibility and makes POMA compatible with other Bioconductor packages (Castellano-Escuder, Andrés-Lacueva, and Sánchez-Pla 2021).

5.1 Create a MSnbase::MSnSet object

First, we create a MSnSet object that integrates both metadata and features in the same data structure.

```
data_msnset <- PomaMSnSetClass(target = pdata, features = t(features))
```

5.2 Preprocessing

Second, we perform the preprocessing step. This step includes the missing value imputation unsing the k-NN algorithm, log Pareto normalization (transformation and scaling) and outlier detection and cleaning. Once these steps are completed, we can proceed to the statistical analysis of these data.

```
data_preprocessed <- data_msnset %>%
  PomaImpute(ZerosAsNA = TRUE, cutoff = 20, method = "knn") %>%
  PomaNorm(method = "log_pareto") %>%
  PomaOutliers(coef = 3)
```

5.3 Limma model

We use a limma model (Ritchie et al. 2015) to identify those most significant metabolites between the "**Baseline urine**" and "**Urine after drinking cranberry juice**" groups. With this analysis we expect to find metabolites related to cranberry intake.

```
limma_res <- data_preprocessed %>%
  PomaLimma(contrast = "Baseline-Cranberry", adjust = "fdr") %>%
  tibble::rownames_to_column("PubChemCID")

# show the first 10 features
limma_res %>%
  dplyr::slice(1L:10L) %>%
  kbl(row.names = FALSE, booktabs = TRUE) %>%
  kable_styling(latex_options = c("striped", "hold_position"))
```

6 Convert PubChem IDs to FOBI IDs

Once we have the results of the statistical analysis and generic identifiers recognized in the FOBI ontology (Castellano-Escuder et al. 2020), we can proceed to perform one of the main functions provided by the fobitools package, the ID conversion. With the fobitools::id_convert() command, users can convert different IDs between FOBI, HMDB, KEGG, PubChem, InChIKey, InChICode, ChemSpider, and chemical

PubChemCID	logFC	AveExpr	t	P.Value	adj.P.Val	B
54678503	-2.015867	-2.22e-05	-7.319346	0.0e+00	0.0000035	11.031428
94214	-1.436933	-4.44e-05	-5.720525	7.0e-07	0.0004658	5.827603
5378303	-1.591333	0.00e+00	-5.578285	1.2e-06	0.0005079	5.367712
71485	-1.520600	4.44e-05	-5.417065	2.0e-06	0.0005552	4.848693
5486800	1.175733	0.00e+00	5.403865	2.1e-06	0.0005552	4.806317
17531	-1.490867	2.22e-05	-5.350936	2.6e-06	0.0005552	4.636603
3035199	-1.298667	-2.22e-05	-5.074320	6.6e-06	0.0011642	3.755554
439361	-1.061267	-2.22e-05	-5.030209	7.6e-06	0.0011642	3.616094
5353	-1.349133	4.44e-05	-5.010060	8.2e-06	0.0011642	3.552497
1132	-1.579533	2.22e-05	-4.983350	8.9e-06	0.0011642	3.468300

names. We will then obtain the FOBI IDs from the PubChem IDs (obtained in the previous sections) and add them as a new column to the results of the limma model.

```
limma_FOBI_names <- limma_res %>%
  dplyr::pull("PubChemCID") %>%
  fobitools::id_convert()

# show the ID conversion results
limma_FOBI_names %>%
  head() %>%
  kbl(row.names = FALSE, booktabs = TRUE) %>%
  kable_styling(latex_options = c("striped", "hold_position"))
```

FOBI	PubChemCID	ChemSpider
FOBI:030415	91	89
FOBI:030711	1145	1113
FOBI:030555	5280445	4444102
FOBI:030709	1123	10675782
FOBI:030625	7533	15484224
FOBI:030397	1794427	1405788

```
limma_FOBI_names <- limma_FOBI_names %>%
  right_join(limma_res, by = "PubChemCID") %>%
  dplyr::arrange(-dplyr::desc(P.Value))
```

7 Enrichment analysis

Enrichment analysis denotes any method that benefits from biological pathway or network information to gain insight into a biological system (Creixell et al. 2015). In other words, these type of analyses integrate the existing biological knowledge (from different biological sources such as databases and ontologies) and the statistical results of *omics* studies, obtaining a deeper understanding of biological systems.

In most metabolomics studies, the output of statistical analysis is usually a list of features selected as statistically significant or statistically relevant according to a pre-defined statistical criteria. Enrichment analysis methods use these selected features to explore associated biologically relevant pathways, diseases,

etc., depending on the nature of the input feature list (genes, metabolites, etc.) and the source used to extract the biological knowledge (GO, KEGG, **FOBI**, etc.).

Here, we present a tool that uses the FOBI information to perform different types of enrichment analyses. Therefore, the presented methods allow researchers to move from lists of metabolites to chemical classes and food groups associated with those lists, and consequently, to the study design.

Currently, the most popular used approaches for enrichment analysis are the over representation analysis (ORA) and the gene set enrichment analysis (GSEA), with its variants for other fields such as the metabolite set enrichment analysis (MSEA) (Xia and Wishart 2010).

7.1 Over representation analysis (ORA)

ORA is one of the most used methods to perform enrichment analysis in metabolomics studies due to its simplicity and easy understanding. This method statistically evaluates the fraction of metabolites in a particular pathway found among the set of metabolites statistically selected. Thus, ORA is used to test if certain groups of metabolites are represented more than expected by chance given a feature list.

However, ORA has a number of limitations. The most important one is the need of using a certain threshold or criteria to select the feature list. This means that metabolites do not meet the selection criteria must be discarded. The second big limitation of ORA is that this method assumes independence of sets and features. In ORA, is assumed that each feature is independent of the other features and each set is independent of the other sets.

Here, we perform an ORA with the `fobitools` package, where we will use as a universe all the metabolites of the study present in FOBI and as a list those metabolites with a raw p-value < 0.01 in the limma results table.

```
metaboliteList <- limma_FOBI_names$FOBI[limma_FOBI_names$P.Value < 0.01]
metaboliteUniverse <- limma_FOBI_names$FOBI

fobitools::ora(metaboliteList = metaboliteList,
               metaboliteUniverse = metaboliteUniverse,
               pvalCutoff = 0.5) %>%
  kbl(row.names = FALSE, booktabs = TRUE) %>%
  kable_styling(latex_options = c("striped", "hold_position"))
```

className	classSize	overlap	pval	padj	overlapMetabolites
soft drink (dietetic)	2	1	0.1038407	1	FOBI:030627
olive oil	10	1	0.4623044	1	FOBI:030340

As we can see, due to the limitations of this methodology and the small number of metabolites that meet the set statistical criterion, the results do not show a clear and obvious relationship with the design of the study, as the food groups that appear in the ORA results do not correspond to those foods administered in the intervention.

7.2 MSEA

Gene Set Enrichment Analysis (GSEA) methodology was proposed for the first time in 2005, with the aim of improving the interpretation of gene expression data. The main purpose of GSEA is to determine whether members of a gene set S tend to occur toward the top (or bottom) of the gene list L, in which case the gene set is correlated with the phenotypic class distinction (Subramanian et al. 2005).

This type of analysis basically consists of three key steps (Subramanian et al. 2005):

The first step consists on the calculation of an enrichment score (ES). This value indicates the degree to which a set S is overrepresented at the extremes (top or bottom) of the entire ranked gene list L. The ES is calculated by walking down the list L, increasing a running-sum statistic when a gene is found in S and decreasing it when a gene is not found in S. The magnitude of the increment depends on the correlation of the gene with the phenotype. The ES is the maximum deviation from zero encountered in the random walk.

The second step is the estimation of significance level of ES. The statistical significance (nominal p-value) of the ES is estimated by using an empirical phenotype-based permutation test that preserves the complex correlation structure of the gene expression data. The phenotype labels (L) are permuted and the ES of the S is recomputed for the permuted data, which generates a null distribution for the ES. The empirical, nominal p-value of the observed ES is then calculated relative to this null distribution. The permutation of class labels (groups) preserves gene-gene correlations and, thus, provides a more biologically reasonable assessment of significance than would be obtained by permuting genes.

Finally, the third step consist on the adjustment for multiple hypothesis testing. When an entire database of gene sets is evaluated, the estimated significance level is adjusted for multiple hypothesis testing. First, the ES is normalized for each gene set to account for the size of the set, yielding a normalized enrichment score (NES). Then, the proportion of false positives is controlled by calculating the FDR corresponding to each NES.

In 2010, a modification of the GSEA methodology was presented for metabolomics studies. This method was called Metabolite Set Enrichment Analysis (MSEA) and its main aim was to help researchers identify and interpret patterns of human and mammalian metabolite concentration changes in a biologically meaningful context (Xia and Wishart 2010). MSEA is currently widely used in the metabolomics community and it is implemented and freely available at the known MetaboAnalyst web-based tool (Xia and Wishart 2010).

As can be seen, GSEA approach is more complex than the ORA methodology, both in terms of methodological aspects and understanding of the method.

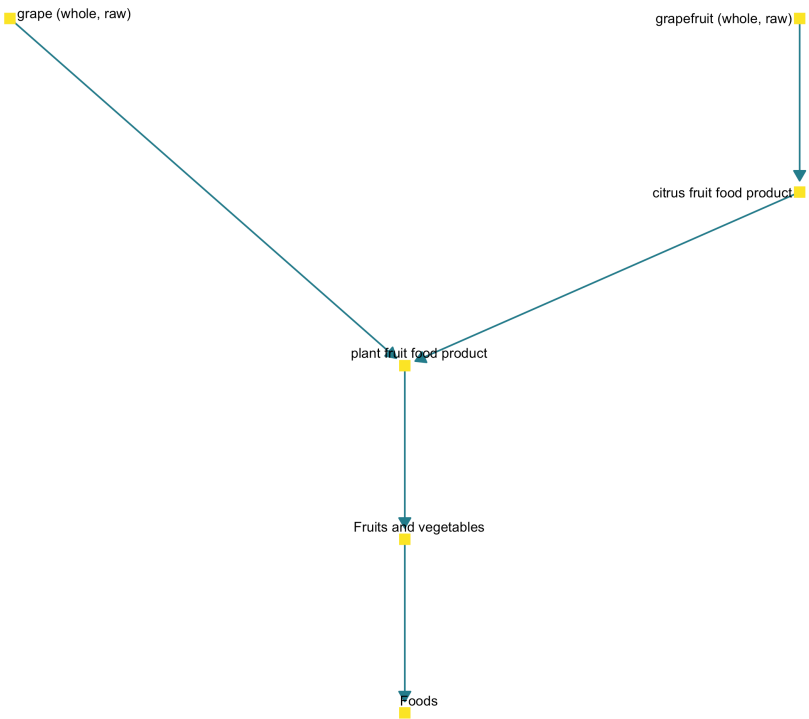
The `fobitools` package provides a function to perform MSEA using the FOBI information. This function requires a ranked list. Here, we will use the metabolites obtained in the limma model ranked by raw p-values.

```
limma_FOBI_msea <- limma_FOBI_names %>%
  select(FOBI, P.Value) %>%
  filter(!is.na(FOBI)) %>%
  dplyr::arrange(-dplyr::desc(abs(P.Value)))

FOBI_msea <- as.vector(limma_FOBI_msea$P.Value)
names(FOBI_msea) <- limma_FOBI_msea$FOBI

msea_res <- fobitools::msea(FOBI_msea, pvalCutoff = 0.06)

msea_res %>%
  kbl(row.names = FALSE, booktabs = TRUE) %>%
  kable_styling(latex_options = c("striped", "hold_position"))
```

className	classSize	log2err	ES	NES	pval	padj	leadingEdge
grape (whole, raw)	1	0.3807304	1.0000000	2.037782	0.0154703	0.2797203	FOBI:030590
grapefruit (whole, raw)	1	0.1978220	0.9729730	1.982706	0.0509491	0.2797203	FOBI:030523
dairy food product	5	0.1882041	0.6383555	1.663786	0.0559441	0.2797203	FOBI:030701, FOBI:030
egg food product	5	0.1882041	0.6383555	1.663786	0.0559441	0.2797203	FOBI:030701, FOBI:030
meat food product	5	0.1882041	0.6383555	1.663786	0.0559441	0.2797203	FOBI:030701, FOBI:030

As we can see, the enrichment analysis with the MSEA method seems to be much more accurate than the ORA method, since the two classes that head the results table ("*grape (whole, raw)*" and "*grapefruit (whole, raw)*") are clearly within the FOBI food group (set) "*plant fruit food product*", which is aligned with the study intervention, cranberry juice intake.

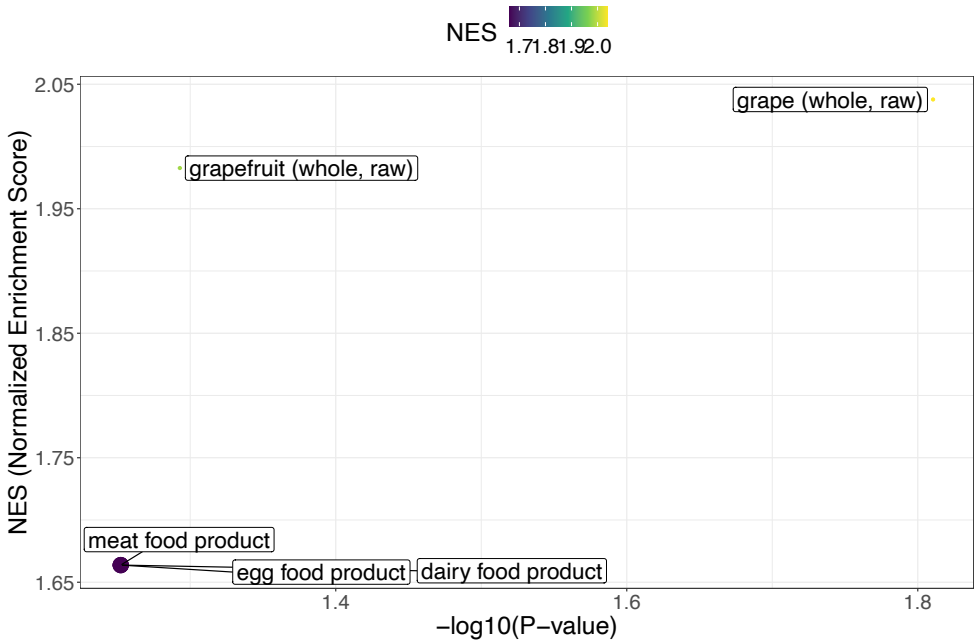
```
fobi_graph(terms = c("FOODON:03301123", "FOODON:03301702"),
           get = "anc",
           labels = TRUE,
           labelsize = 6)
```



7.2.1 MSEA plot with ggplot2

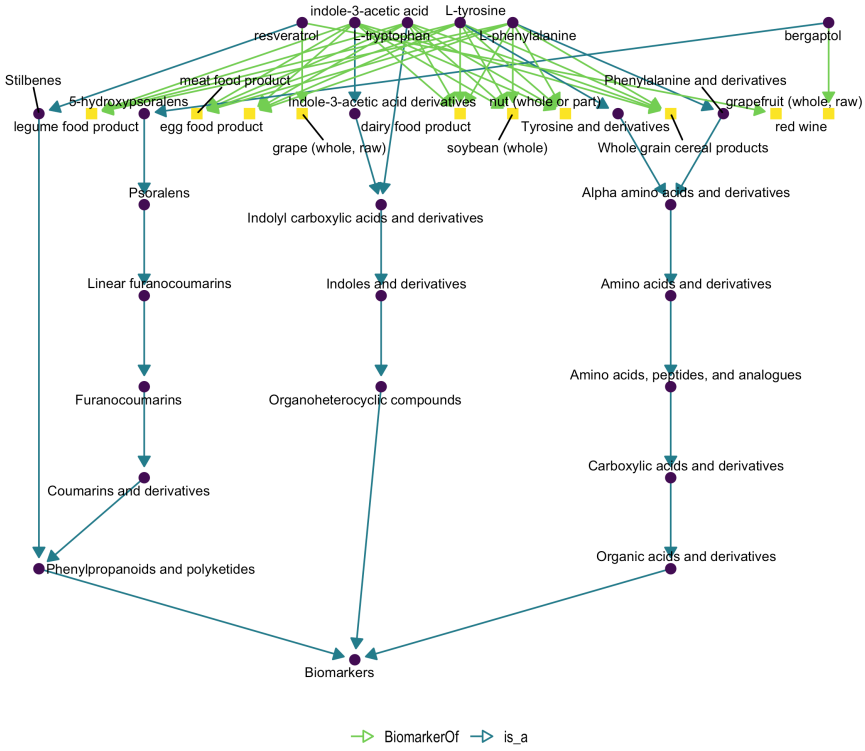
```
ggplot(msea_res, aes(x = -log10(pval), y = NES, color = NES, size = classSize, label = className)) +
  xlab("-log10(P-value)") +
  ylab("NES (Normalized Enrichment Score)") +
  geom_point() +
```

```
  ggrepel::geom_label_repel(color = "black", size = 7) +
  theme_bw() +
  theme(legend.position = "top",
        text = element_text(size = 22)) +
  scale_color_viridis_c() +
  scale_size(guide = "none")
```



7.2.2 Network of metabolites found in MSEA

```
FOBI_terms <- msea_res %>%
  unnest(cols = leadingEdge)

fobitools::fobi %>%
  filter(FOBI %in% FOBI_terms$leadingEdge) %>%
  pull(id_code) %>%
  fobi_graph(get = "anc",
             labels = TRUE,
             legend = TRUE,
             labelsize = 6,
             legendSize = 20)
```



8 Limitations

The FOBI ontology is currently in its first release version, so it does not yet include information on many metabolites and food relationships. All future efforts will be directed at expanding this ontology, leading to a significant increase in the number of metabolites and metabolite-food relationships. The `fobitools` package provides the methodology for easy use of the FOBI ontology regardless of the amount of information it contains. Therefore, future FOBI improvements will also have a direct impact on the `fobitools` package, increasing its utility and allowing to perform, among others, more accurate, complete and robust enrichment analyses.

9 Session Information

```
sessionInfo()
#> R version 4.0.2 (2020-06-22)
```

```
#> Platform: x86_64-apple-darwin17.0 (64-bit)
#> Running under: macOS  10.16
#>
#> Matrix products: default
#> BLAS:   /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
#> LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib
#>
#> locale:
#> [1] es_ES.UTF-8/es_ES.UTF-8/es_ES.UTF-8/C/es_ES.UTF-8/es_ES.UTF-8
#>
#> attached base packages:
#> [1] parallel  stats4    stats     graphics  grDevices utils     datasets
#> [8] methods   base
#>
#> other attached packages:
#>  [1] SummarizedExperiment_1.20.0  Biobase_2.50.0
#>  [3] GenomicRanges_1.42.0         GenomeInfoDb_1.26.7
#>  [5] IRanges_2.24.1               S4Vectors_0.28.1
#>  [7] BiocGenerics_0.36.1          MatrixGenerics_1.2.1
#>  [9] matrixStats_0.58.0           metabolomicsWorkbenchR_1.0.0
#> [11] POMA_1.1.13                  kableExtra_1.3.4
#> [13] ggrepel_0.9.1                rvest_1.0.0
#> [15] forcats_0.5.1                stringr_1.4.0
#> [17] dplyr_1.0.6                  purrr_0.3.4
#> [19] readr_1.4.0                  tidyr_1.1.3
#> [21] tibble_3.1.1                 ggplot2_3.3.3
#> [23] tidyverse_1.3.1              fobitools_0.99.56
#>
#> loaded via a namespace (and not attached):
#>   [1] utf8_1.2.1                  tidyselect_1.1.1
#>   [3] RSQLite_2.2.7               grid_4.0.2
#>   [5] BiocParallel_1.24.1         gmp_0.6-2
#>   [7] pROC_1.17.0.1               munsell_0.5.0
#>   [9] codetools_0.2-18            preprocessCore_1.52.1
#>  [11] withr_2.4.2                 colorspace_2.0-1
#>  [13] highr_0.9                   knitr_1.33
#>  [15] rstudioapi_0.13             mzID_1.28.0
#>  [17] labeling_0.4.2              GenomeInfoDbData_1.2.4
#>  [19] polyclip_1.10-0             bit64_4.0.5
#>  [21] farver_2.1.0                vctrs_0.3.8
#>  [23] generics_0.1.0              ipred_0.9-11
#>  [25] xfun_0.22                   randomForest_4.6-14
#>  [27] R6_2.5.0                    doParallel_1.0.16
#>  [29] clue_0.3-59                 graphlayouts_0.7.1
#>  [31] DelayedArray_0.16.3         bitops_1.0-7
#>  [33] cachem_1.0.4                fgsea_1.16.0
#>  [35] assertthat_0.2.1            scales_1.1.1
#>  [37] vroom_1.4.0                 ggraph_2.0.5
#>  [39] nnet_7.3-16                 gtable_0.3.0
#>  [41] Cairo_1.5-12.2              affy_1.68.0
#>  [43] tidygraph_1.2.0             timeDate_3043.102
#>  [45] tictoc_1.0.1                rlang_0.4.11
#>  [47] clisymbols_1.2.0            systemfonts_1.0.1
```

```
#>  [49] mzR_2.24.1                  GlobalOptions_0.1.2
#>  [51] splines_4.0.2               ModelMetrics_1.2.2.2
#>  [53] impute_1.64.0               selectr_0.4-2
#>  [55] broom_0.7.6                 RecordLinkage_0.4-12.1
#>  [57] BiocManager_1.30.12         yaml_2.2.1
#>  [59] reshape2_1.4.4              modelr_0.1.8
#>  [61] backports_1.2.1             caret_6.0-86
#>  [63] tools_4.0.2                 lava_1.6.9
#>  [65] affyio_1.60.0               ellipsis_0.3.2
#>  [67] ff_4.0.4                    RColorBrewer_1.1-2
#>  [69] proxy_0.4-25                MSnbase_2.16.1
#>  [71] MultiAssayExperiment_1.16.0 Rcpp_1.0.6
#>  [73] plyr_1.8.6                  zlibbioc_1.36.0
#>  [75] RCurl_1.98-1.3              rpart_4.1-15
#>  [77] GetoptLong_1.0.5            viridis_0.6.1
#>  [79] haven_2.4.1                 cluster_2.1.2
#>  [81] fs_1.5.0                    magrittr_2.0.1
#>  [83] RSpectra_0.16-0             data.table_1.14.0
#>  [85] circlize_0.4.12             reprex_2.0.0
#>  [87] pcaMethods_1.82.0           ProtGenerics_1.22.0
#>  [89] hms_1.0.0                   patchwork_1.1.1
#>  [91] evaluate_0.14               xtable_1.8-4
#>  [93] XML_3.99-0.6                readxl_1.3.1
#>  [95] gridExtra_2.3               shape_1.4.5
#>  [97] compiler_4.0.2              ellipse_0.4.2
#>  [99] ncdf4_1.17                  crayon_1.4.1
#> [101] htmltools_0.5.1.1           mgcv_1.8-35
#> [103] corpcor_1.6.9               qdapRegex_0.7.2
#> [105] lubridate_1.7.10            DBI_1.1.1
#> [107] tweenr_1.0.2                dbplyr_2.1.1
#> [109] ComplexHeatmap_2.6.2        MASS_7.3-54
#> [111] Matrix_1.3-2                permute_0.9-5
#> [113] cli_2.5.0                   vsn_3.58.0
#> [115] textclean_0.9.3             evd_2.3-3
#> [117] RankProd_3.16.0             gower_0.2.2
#> [119] igraph_1.2.6                pkgconfig_2.0.3
#> [121] recipes_0.1.16              MALDIquant_1.19.3
#> [123] xml2_1.3.2                  foreach_1.5.1
#> [125] rARPACK_0.11-0              svglite_2.0.0
#> [127] ggcorrplot_0.1.3            XVector_0.30.0
#> [129] webshot_0.5.2               prodlim_2019.11.13
#> [131] ada_2.0-5                   digest_0.6.27
#> [133] vegan_2.5-7                 rmarkdown_2.7
#> [135] cellranger_1.1.0            fastmatch_1.1-0
#> [137] curl_4.3.1                  rjson_0.2.20
#> [139] glasso_1.11                 lifecycle_1.0.0
#> [141] nlme_3.1-152                jsonlite_1.7.2
#> [143] mixOmics_6.14.1             viridisLite_0.4.0
#> [145] limma_3.46.0                fansi_0.4.2
#> [147] pillar_1.6.0                ontologyIndex_2.7
#> [149] lattice_0.20-44             fastmap_1.1.0
#> [151] httr_1.4.2                  survival_3.2-11
#> [153] glue_1.4.2                  png_0.1-7
```

```
#> [155] iterators_1.0.13          glmnet_4.1-1
#> [157] bit_4.0.4                 ggforce_0.3.3
#> [159] class_7.3-19              stringi_1.6.1
#> [161] struct_1.2.0              blob_1.2.1
#> [163] memoise_2.0.0             Rmpfr_0.8-4
#> [165] e1071_1.7-6
```

References

Castellano-Escuder, Pol, Cristina Andrés-Lacueva, and Alex Sánchez-Pla. 2021. *POMA: User-Friendly Workflow for Pre-Processing and Statistical Analysis of Mass Spectrometry Data*. https://github.com/pcastellanoescuder/POMA.

Castellano-Escuder, Pol, Raúl González-Domínguez, David S Wishart, Cristina Andrés-Lacueva, and Alex Sánchez-Pla. 2020. "FOBI: An Ontology to Represent Food Intake Data and Associate It with Metabolomic Data." *Database* 2020.

Creixell, Pau, Jüri Reimand, Syed Haider, Guanming Wu, Tatsuhiro Shibata, Miguel Vazquez, Ville Mustonen, et al. 2015. "Pathway and Network Analysis of Cancer Genomes." *Nature Methods* 12 (7): 615.

Ritchie, Matthew E, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. 2015. "limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." *Nucleic Acids Research* 43 (7): e47. https://doi.org/10.1093/nar/gkv007.

Subramanian, Aravind, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, et al. 2005. "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *Proceedings of the National Academy of Sciences* 102 (43): 15545–50.

Sud, Manish, Eoin Fahy, Dawn Cotter, Kenan Azam, Ilango Vadivelu, Charles Burant, Arthur Edison, et al. 2016. "Metabolomics Workbench: An International Repository for Metabolomics Data and Metadata, Metabolite Standards, Protocols, Tutorials and Training, and Analysis Tools." *Nucleic Acids Research* 44 (D1): D463–D470.

Xia, Jianguo, and David S Wishart. 2010. "MSEA: A Web-Based Tool to Identify Biologically Meaningful Patterns in Quantitative Metabolomic Data." *Nucleic Acids Research* 38 (suppl_2): W71–W77.

Appendix B

B.1 Other publications

This section includes scientific publications directly or indirectly related to the content of this thesis and carried out during the period of its realization. All publications mentioned in this section have already been sent to different international scientific journals but have not yet been accepted for publication. Each publication includes the author list and its corresponding abstract.

B.1.1 Paper 6: A polyphenol-rich diet causes increase in the gut microbiota metabolite indole 3-propionic acid in older adults with preserved kidney function

Gregorio Peron*, Tomás Meroño*, Giorgio Gargari, Nicole Hidalgo-Liberona, Antonio Miñarro, Esteban Vegas Lozano, **Pol Castellano-Escuder**, Cristian Del Bo', Stefano Bernardi, Paul A. Kroon, Antonio Cherubini, Patrizia Riso, Simone Guglielmetti, Cristina Andrés-Lacueva. *A polyphenol-rich diet causes increase in the gut microbiota metabolite indole 3-propionic acid in older adults with preserved kidney function: a randomized, controlled, crossover trial.*

*Authors equally contributed to this work.

- **ABSTRACT**

Dietary polyphenols can trigger the production of microbial bioactive metabolites by altering the gut microbiota (GM). Here, our aim was to determine if a polyphenol-rich (PR) diet could affect the production of specific bioactive GM-tryptophan metabolites in older adults involved in an 8-week randomized, controlled, crossover trial. Subgroup analyses based on kidney function (normal glomerular filtration rate: 90-120 ml/min/1.73m^2) were performed. The PR-diet significantly increased serum indole 3-propionic acid (IPA) in subjects with normal renal function (NRF), although the same effect was not observed in subjects with impaired renal function. Other GM-tryptophan metabolites were not affected. Comparison of baseline GM composition showed shifts in Bacteroidales order members as well as higher abundance of Clostridiales in participants with NRF. During the trial, variations of IPA were associated with changes in C-reactive protein (β = 0.32, p = 0.010) and GM, particularly with the Clostridiales (r = 0.35, p $<$ 0.001) and Enterobacteriales (r = -0.15, p $<$ 0.05) orders.

- **MaPLE STUDY DESIGN**

The "MaPLE" project (Guglielmetti et al., 2020) is a crossover interventional study of older people (>65 years) where participants toke both polyphenol-rich and control diets during two different periods of time. Halfway through the study, participants were subjected to a wash-out period in order to invert the control and intervention groups (Figure B.1). Urine metabolites were measured and compared at four different times of the study.

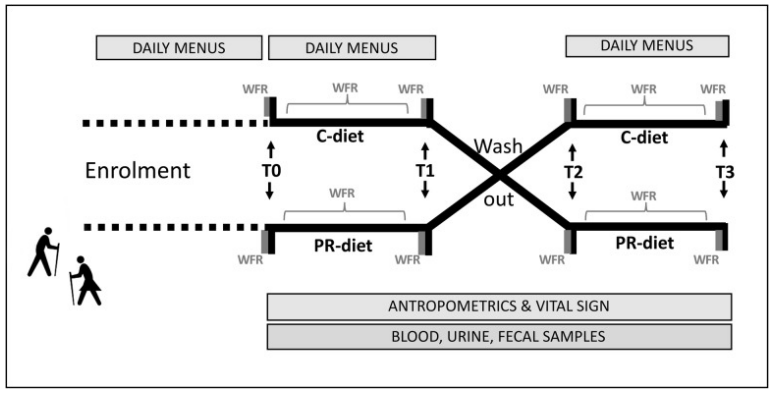


Figure B.1: MaPLE project study design (Guglielmetti et al., 2020).

B.1.2 Paper 7: Apolipoprotein E and sex modulate fatty acid metabolism in early cognitive decline

Raúl González-Domínguez, **Pol Castellano-Escuder**, Sophie Lefèvre-Arbogast, Dorrain Y. Low, Andrea Du Preez, Silvie R. Ruigrok, Hyunah Lee, Catherine Helmer, Mercè Pallàs, Mireia Urpi-Sarda, Alex Sánchez-Pla, Aniko Korosi, Paul J. Lucassen, Ludwig Aigner, Claudine Manach, Sandrine Thuret, Cécilia Samieri, Cristina Andres-Lacueva. *Apolipoprotein E and sex modulate fatty acid metabolism in early cognitive decline.*

- **ABSTRACT**

Fatty acids and related pathways are known to be disturbed in cognitive decline, but the involvement of common risk factors, namely the $\epsilon 4$ allele of the apolipoprotein E (ApoE-$\epsilon 4$) gene and sex, remains elusive. Targeted metabolomics analysis was performed on serum samples from a nested case-control study (N=368), part of a prospective population cohort on dementia. Circulating levels of free fatty acids, acyl-carnitines and pantothenic acid were increased among participants who had greater odds of cognitive decline over a 12-year follow-up. Stratified analyses indicated that these alterations were

specific for ApoE-ϵ4 non-carriers and women. Our results highlight that the regulation of fatty acids and related metabolic pathways during cognitive decline depends on the ApoE-ϵ4 genotype and sex. A better understanding of this intertwined modulation would help to elucidate the impact of individual variability in the onset of cognitive decline and to develop personalized therapeutic approaches.

B.1.3 Paper 8: A mixture of four dietary fibres ameliorates adiposity, and improves metabolic profile and intestinal health in cafeteria-fed obese rats: an integrative multi-omics approach

Núria Estanyol-Torres, Cristina Domenech-Coca, Raúl González-Domínguez, Antonio Miñarro, Ferran Reverter, Jose Antonio Moreno-Muñoz, Jesús Jiménez, Manel Martín-Palomas, **Pol Castellano-Escuder**, Hamza Mostafa, Santi García-Vallvé, Nerea Abasolo, Miguel A. Rodríguez, Helena Torrell, Josep M del Bas, Alex Sanchez-Pla, Antoni Caimari, Anna Mas-Capdevila, Cristina Andres-Lacueva, Anna Crescenti. *A mixture of four dietary fibres ameliorates adiposity, and improves metabolic profile and intestinal health in cafeteria-fed obese rats: an integrative multi-omics approach.*

- **ABSTRACT**

Dietary fibre is a health-promoting nutrient well-known to lower risk for obesity and to improve intestinal and metabolic health, although its effects depend on the properties of each fibre. The aim of this study was to assess the effects of a mixture of the fibres inulin, hydrolysed guar gum, resistant maltodextrin and dehydrated plum, using a daily dose extrapolated to human consumption, against cafeteria diet-induced obesity in rats. We studied a wide number of biometric and biochemical parameters, conducted a multi-omics approach based on transcriptomics, metagenomics and metabolomics analysis and applied an integrative multivariate

analysis. The intervention reduced the body weight and adiposity of animals, which was probably mediated by an increase of energy expenditure and lipid oxidation. Fibre supplementation reduced HbA1c and adiponectin blood levels and liver cholesterol levels. Fibre intake improved the intestinal health and endotoxemia (i.e., increased caecal weight and small intestine length/weight ratio, reduced LPS serum levels and MPO activity in the colon), which was in turn reflected at the metabolomics (i.e., production of short chain fatty acids and phenolic acids), metagenomics (i.e., modulation of *Ruminococcus* species) and transcriptomics levels (i.e., expression of tight junctions). Transcriptomics analysis showed downregulated proteolysis in response to fibre, in line with the decrease of amino acid levels observed in serum and urine and with the increase of *Lactobacillus* counts. Altogether, our integrative multi-omics approach highlights the great potential of the supplementation with the mixture of fibres to ameliorate the impairments in adiposity, metabolic and intestinal health triggered by obesity.

B.2 Other software

This section includes software applications not directly related to the content of this thesis but generated during its development.

B.2.1 Lheuristic

Lheuristic is a web-based tool for exploring correlations between gene methylation and expression in *omics* studies. Lheuristic tool provides a heuristic algorithm for selecting those genes with a L pattern in the expression-methylation scatterplot. Thus, the main aim of this tool is to detect genes potentially regulated by methylation.

Available documents:

- Lheuristic R package GitHub repository: `https://github.com/ASPresearch/Selecting_GRM`
- Lheuristic Shiny app GitHub repository: `https://github.com/pcastellanoescuder/Lheuristic`

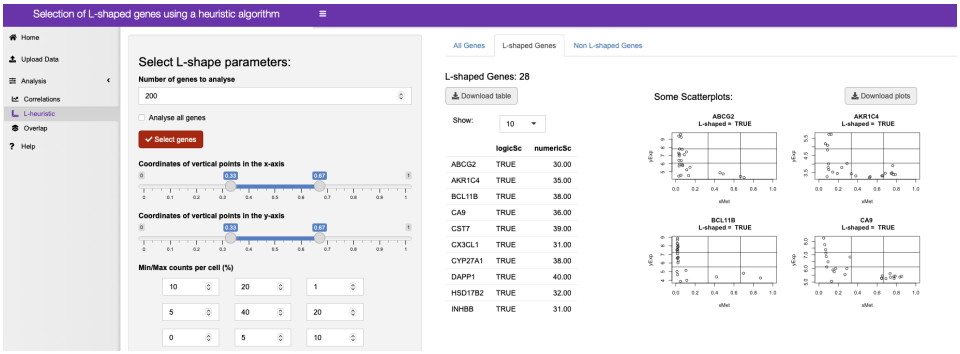


Figure B.2: Screenshot of the Lheuristic *Home* page.

B.2.2 Covid19Explorer

Covid19Explorer is a web-based application that provides a set of tools for visualization, exploration, and statistical analysis of complex multivariate COVID-19 data. This tool was developed in the context of the COVID-19 pandemic to facilitate the analysis of these data. Covid19Explorer was made in collaboration with different researchers from the Vall d'Hebron hospital in Barcelona.

Available documents:

- Covid19Explorer URL: `http://uebshiny.vhir.org:3838/Covid19Explorer/`
- Covid19Explorer GitHub repository: `https://github.com/pcastellanoescuder/Covid19Explorer`

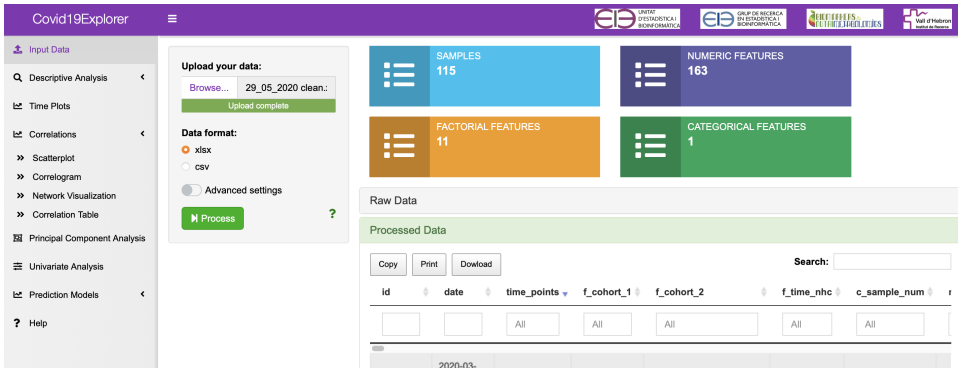


Figure B.3: Screenshot of the Covid19Explorer *Home* page.