Miriam Navarro Sanz

Proteomic and metabolomic approaches to study diabetic retinopahty

Ph Thesis Dissertation

Supervised by

Dr. Oscar Yanes Torrado

Dr. Maria Vinaixa

Department of Medicine and Surgery

UNIVERSITAT ROVIRA I VIRGILI

Reus

2018

UNIVERSITAT
ROVIRA I VIRGILI

I STATE that the present study, entitled “Proteomic and metabolomic approaches to study diabetic retinopathy”, presented by Míriam Navarro Sanz for the award of the degree of Doctor, has been carried out under my supervision at the Department of Medicina i Cirurgia of this university.

Reus, 4th September 2018

Doctoral Thesis Supervisor/s

OSCAR YANES TORRADO
Digitally signed by OSCAR YANES TORRADO
Date: 2018.09.04 16:03:49 +02'00'

Óscar Yanes Torrado

MARIA VINAIXA CREVILLENT
Digitally signed by MARIA VINAIXA CREVILLENT
Date: 2018.09.04 02:01:47 +02'00'

Maria Vinaixa Crevillent

“Research is to see what everybody else has seen,

and to think what nobody else has thought”

Albert Szent-Gyorgyi

Aknowledgments

Después de estos 5 años me gustaría agradecer a todos los que habéis formado parte de esta aventura que sin duda me ha marcado tanto profesional como personalmente.

En primer lugar, me gustaría dar las gracias a mis dos directores de tesi, a Óscar por haberme dado la oportunidad de tomar parte de este proyecto. Gracias por compartir tu gran pasión por la ciencia y prepararme para abordar todos los retos que conlleva la investigación. A Mariona, gracias por compartir conmigo todo el conocimiento que has ido adquiridiendo a base de una constante lucha por superarte cada día un poco más.

También quería agradecer a Xavier Correig, director de la plataforma de metabolómica, por haber creado un grupo tan multidisciplinario y rico en conocimiento y aceptarme en él.

Y no me puedo olvidar de mis compañeros y amigos de tesis. Gracias Sara por tu tiempo dedicado a que aprendiera el workflow de la metabolómica y sobretodo por tu buena actitud ante los retos. A Miguel por enseñarme tu gran maestria con el RMN y saber tanto de tantas cosas. A Sandra, que aunque llegaste cuando yo estaba a punto de irme te dio tiempo a demostrame lo buena compañera y profesional que eres. A Jordi, mi compañero de batallas y sobretodo amigo. Gracias por tu generosidad, tu tiempo y compartir conmigo todos tus conocimientos.

Agradecer al resto de estudiantes y personal del grupo, a Ruben, Roger, Josep, Xavi, Dídac, Judith, Sonia, y Noelia por haberme acompañado en esta etapa, sin vosotros el camino no hubiera sido tan emocionante.

A Rosa y Silvia por su impecable profesionalidad.

A SEES lab, Roger, Marta, Toni y Oriol, por darme la oportunidad de estar en vuestro lab y enseñarme vuestro asombroso mundo de las networks.

A Shabaz por haberme concedido el privilegio de realizar mi estancia en la Universidad de Oxford y haberme iniciado en el campo de la proteómica.

A Lore, gracias por lo extraordiaria persona que eres, por compartir tantos momentos juntas, Reus sin ti no hubiera sido ni la mitad de interesante. Gracias a esas amigas que han hecho que mi vida en Reus fuera de todo menos aburrida. Rocío y María, las mejores compis de piso pero sobretodo, buenas amigas. A Núria por enseñarme tu asombroso mundo y compartir tantos buenos momentos juntas. A Miriam, por llevar siempre contigo esa buena onda que tanto te caracteriza.

Y aunque la tesis la haya hecho en Reus no me quiero olvidar de mi familia y amigos de Barcelona. A Gisela mi compañera de aventuras en la vida, gracias por escucharme y darme tantos consejos. A

Miriam, Marta y Mireia, por la buena química que hay entre nosotras.

Y a mi familia, gracias por vuestro cariño, paciencia y fuerza. Sin vosotros todo esto no hubiera sido posible.

Finalmente, gracias Miguel por estar siempre ahí, por tus sabios consejos, por tu tiempo escuchándome y animándome a conseguir todos los retos que me proponga.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS **17**

ABSTRACT **25**

CHAPTER 1. Introduction **31**

1.1 Systems biology in life sciences 33

1.1.1 Components data. 35

1.1.2 Interactions data 38

1.1.3 Functional-states data 38

1.2 Metabolomics 39

1.2.1 Metabolomics history 41

1.2.2 Metabolomics strategies 48

1.2.3 Untargeted metabolomics 51

1.2.3.1 Untargeted metabolomics workflow 51

1.2.3.1.1 Experimental design 55

1.2.3.1.2 Sample preparation 55

1.2.3.1.3 Data acquisition 65

1.2.3.1.4 Data processing 84

1.2.3.1.5 Data analysis 105

1.2.3.1.6 Biological interpretation 111

1.3 Proteomics 116

1.3.1 Shotgun quantitative proteomics workflow 117

1.3.1.1 Proteins extraction 118

1.3.1.2 Protein depletion and denaturalization 119

1.3.1.3 Proteolytic Digestion 120

1.3.1.4 Data acquisition by LC-MS 122

1.3.1.5 Data processing: protein identification and quantification 125

1.3.1.6 Data analysis and biological interpretation 132

1.4 Diabetic retinopathy 135

1.4.1 DR stages (NPDR, PDR) 137

1.4.2 Metabolomics and diabetes 141

1.4.3 Metabolomics and Diabetic Retinopathy 145

1.4.4 Proteomics and Diabetic Retinopathy 148

CHAPTER 2. Objectives 151

CHAPTER 3. Results 155

3.1 LC-MS processing: from MS1 to MS2 to metabolite ID .. 157

3.1.1 NIST database for LC/MS-based metabolomics 158

3.1.2 CliqueMS: A computational tool for annotating in-source MS1 untargeted metabolomics data based on a coelution similarity network........162

3.1.3 iMet: A Network-Based Computational Tool To Assist in the Annotation of Metabolites from Tandem Mass Spectra. 184

3.2 Alterations in the proteome and metabolome of human retinal pigment epithelium cells exposed to hyperglycemic and/or hypoxic conditions........188

3.2.1 Analysis of a protein-protein interaction network in retinal pigment epithelial cells exposed to hyperglycemic and hypoxic conditions........188

3.2.2 The connectivity of the human metabolic network reveals altered metabolites in ARPE-19 cells and vitreous humor of diabetic retinopathy patients........218

CHAPTER 4. Discussion........251

CHAPTER 5. Conclusions........261

CHAPTER 6. References........267

LIST OF ABBREVIATIONS

List of abbreviations

1D-NMR	One dimensional NMR
2D-NMR	Two dimensional NMR
AAA	Aromatic amino acids
ArMet	Architecture for Metabolomics consortium
ARPE-19	Human retinal pigment epithelial cell line
BCAA	Branched chain amino acid
BD2K	Big Data to Knowledge
BM	Basement membrane
BRB	Blood retinal barrier
BSS	Blind source separation
CAD	Coronary artery disease
CAS	Chemical Abstracts Service
CASMI	Critical Assessment of Small Molecule Identification
CE	Capillary electrophoresis
CI	Chemical ionization
COSY	Correlation spectrometry

CSF	Cerebrospinal fluid
CV	Coefficient of variation
DDA	Data-dependent acquisition
DLLs	Dynamic-Link Libraries
DM	Diabetes mellitus
DMO	Diabetic macular oedema
DN	Diabetic nephropathy
DPN	Diabetic peripheral neuropathy
DR	Diabetic retinopathy
EBI	European Bioinformatics Institute
EI	Electron impact
EIC	Extracted ion chromatograms
ESI	Electrospray ionization
EtOH	Ethanol
FDR	False discovery rate
FFA	Free fatty acids

List of abbreviations

FTICR	Fourier transform ion cyclotron resonance
GC	Gas chromatography
HGP	Human Genome Project
HILIC	Hydrophilic interaction chromatography
HMDB	Human Metabolome Database
IR	Insulin resistance
IRMA	Intraretinal microvascular abnormalities
Itraq	Isobaric tags
LC	Liquid chromatography
LIT	Linear ion trap
LTQ	Linear quadrupole ion trap
m/z	mass-to-charge ratio
MeOH	Methanol
MI	Minimal information
MPA	Meta-phosphoric acid
MS	Mass Spectrometry

List of abbreviations

MS/MS	Tandem mass spectrometry
MSI	Metabolomics Standards Initiative
MSSC	Metabolite Standards Synthesis Core
MW	Molecular weight
mzRT	Peak with unique m/z and a specific retention time
NIH	National Institutes of Health
NIST	National Institute of Standards and Technology
NMR	Nuclear magnetic resonance
NOESY	Nuclear Overhauser effect spectroscopy
NOS	Nitric oxide synthase
NPDR	Non proliferative diabetic retinopathy
ORA	Overrepresentation analysis
PCA	Principal component analysis
PDR	Proliferative diabetic retinopathy
PPI	Protein-protein interaction
Q	Quadrupole

QC	Quality control
QIT	Quadrupole ion trap
Q-TOF	Quadrupole-TOF
QTrap	Triple-quadrupole ion trap
RF	Radio frequency
RP	Reversed-phase
RT	Retention time
SAGE	Serial analysis of gene expression
SDS	Anionic sodium dodecyl sulfate
SEA	Set enrichment analysis
SILAC	Stable isotope labeling by amino acids in cell culture
T2DM	Type 2 diabetes mellitus
TAG	Triacylglycerol
TCA	Tricarboxylic acid
TFA	Trifluoroacetic acid
TMT	Tandem mass tags

TOCSY	Total correlation spectroscopy
TOF	Time of flight
TQ	Triple quadrupole
XIC	Extracted ion chromatogram
YM3	Metabolite extraction method from Yanes et al.[1]

ABSTRACT

Abstract

The general objective of this doctoral thesis was to develop, analyse and validate new bioinformatic tools for converting raw MS-based metabolomics data into biological knowledge, in order to study alterations in the proteome and metabolome of human retinal pigment epithelium cells exposed to hyperglycemic and/or hypoxic conditions.

To reach this general objective, I have structured my thesis in two blocks:

- Methodological aims: (i) analyse mass spectral databases for LC/MS-based untargeted metabolomics, and (ii) generate and improve the characterization of LC/MS metabolomics data focusing on MS1 and MS2 annotation.
- Biological aims: (iii) detect and analyse changes in protein-protein interaction (PPI) networks by hyperglycemic and/or hypoxic conditions, and (iv) predict and validate metabolite alterations due to hyperglycemic and/or hypoxic conditions integrating protein expression data in metabolic networks.

My contribution to aim (i) was a thorough study of the NIST14 mass spectral database, proving it superior to other available databases due to its larger number of metabolites with spectral data acquired from different adducts and using a wide range of mass spectrometers. I showed the importance of adduct formation for metabolite identification, analysing MS/MS data for different adducts in my lab.

This is particularly important because predominant adducts in an ESI spectrum vary from one metabolite to another as well as on the mobile phase used (*Vinaixa M et al. TrAC Trends in Analytical Chemistry, 2016*).

My contribution to aim (ii) was to evaluate the performance of two new computational tools: CliqueMS and iMet, which were developed in collaboration with SEES lab led by Dr. Roger Guimerà and Marta Sales-Pardo (URV). CliqueMS annotates in-source MS1 data based on a coelution similarity network. I have experimentally proven that CliqueMS correctly identifies and annotates a large number of adducts from pure standards and complex biological samples, leading to more correctly parental ion neutral masses than CAMERA, the most widely-used approach (*Senan et al. Bioinformatics, under revision*).

iMet facilitates structural annotation of metabolites based on MS2 data not described in mass spectral databases. I simulated a real scenario of metabolites not present in a database by testing iMet using metabolites proposed in the Critical Assessment of Small Molecule Identification (CASMI) challenges from years 2012-2016. In addition, I compared iMet's performance to other tools such as CFM-ID, MetFrag and MS-Finder. I could demonstrate the potential of iMet for annotating metabolites that are not present in databases, as well as to compare the performance of other methods to assist the structural annotation of known metabolites lacking MS/MS spectra in databases (*Aguilar-Mogas et al. Analytical Chemistry, 2017*).

On the other hand, the biological applications of my thesis aimed at studying diabetic retinopathy (DR) using metabolomics and proteomics approaches. It is important to highlight that the training and results obtained above was key to improve the metabolome coverage in ARPE-19 cells and vitreous humour samples.

My contribution to aim (iii) was to generate proteomics data and develop a novel approach that integrates PPI, module analysis and protein expression for detecting dysregulated groups of interacting proteins involved in similar biological processes. This work was performed in collaborations with the Structural Bioinformatics and Network Biology group led by Dr. Patrick Aloy at the IRB Barcelona, and with the SEES lab at the URV. Using this new approach, it has been possible to capture slight but consistent protein changes occurring in a protein module which are impossible to detect considering only individual proteins. (*Navarro M et al. In preparation, 2018*).

My contribution to aim (iv) was to validate a novel proteomics data analysis workflow based on a human genome-scale metabolic network that predicted metabolic alterations in an *in vitro* model of DR. I have generated and analysed metabolomics data on ARPE-19 cells cultured at low and high glucose concentrations, and normoxic or hypoxic conditions, also fed with 13C-glucose for isotopic label tracking (flux analysis), to validate the predictions made by our novel data analysis workflow. In addition, I also analysed human vitreous humor from DR patients and controls for clinical validation.

CHAPTER 1. Introduction

In this thesis we present the structure of systems biology and the role that it plays in the life sciences study. We have focused on two important blocks that build the systems biology scaffold: metabolomics and proteomics, explaining the state of the art of their respective methodological workflows. Finally, we show the application and integration of these two omics levels to understand the diabetic retinopathy highlighting the biological mechanism leading to neurodegenerative state.

1.1 Systems biology in life sciences

Systems biology goal consist in describing the structure of the biological system and its response to genetic, biological or chemical disturbance through mathematical models implementation. This goal can only be achieved by monitoring and integrating the genome, transcriptome, proteome and metabolome response in a network of regulatory interactions[2]. Nowadays, due to the incredible advances in new technologies such as DNA sequencers, microarrays, and high-throughput proteomics and metabolomics have enabled us to gain information of the underlying molecules and collect comprehensive data set on system performance. All these different system levels forms the omics cascade (Figure 1) and can be hierarchically described in nature: DNA → mRNA → protein → protein interactions → informational pathways → informational networks → cells → tissues → an organism → populations → ecologies.

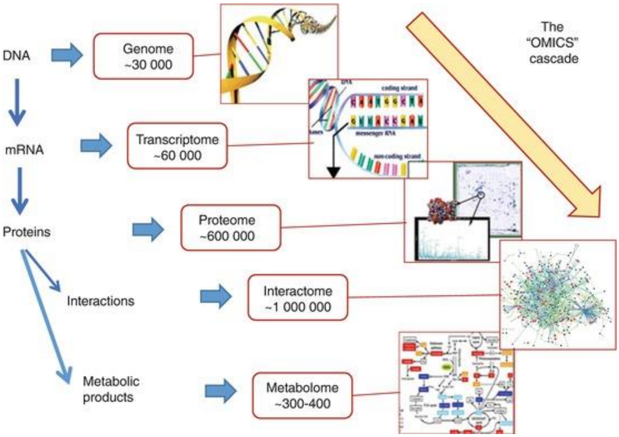


Figure 1. Omics cascade (adopted from [3])

Systems biology was first introduced more than one decade ago after the scientific community realized[2,4] the need of understanding biology at the system levels not examining the characteristics of isolated parts but studying the structure and dynamics of cellular and organismal functions of a cell. This transition in biology from the molecular levels to the system levels to understand complex biological regulatory systems was possible thanks to the Human Genome Project (HGP), one of the first modern biological endeavors to practice discovery science. The objective of HGP was the complete mapping and understanding of all the genes of human beings that are known as our "genome."

It is analogous to a static roadmap, whereas what we really seek to know are the traffic patterns, why such traffic patterns emerge, and how we can control them. Thus, system-level understanding requires a shift in our notion of “what to look for” in biology focusing on understanding a system’s structure and dynamics[4, 5, 6, 7]. This goal can be achieved through Big Data to Knowledge (BD2K), whose aim is to connect multiple disparate data types to obtain a meaningful understanding of the biological functions of an organism. The unprecedented growth in the type, size and complexity of biological data sets over the past couple of decades has led to a pressing grand challenge in biology referred to as BD2K.

As it has been reviewed by Joyce and Palsson[8], the model organism can be studied as a system by integrating ‘omics’ data sets. The description of the cellular network that these omics data provide for a given time and/or condition can be classified into three broad categories: components, interactions and functional states.

1.1.1 Components data.

Components refer to the specific molecular content of the cell or system: genomics, transcriptomics, proteomics and metabolomics.

Briefly, genomics is defined as the study of the whole genome sequence and the information contained therein, is clearly the most mature of the different omics fields. Since 1995, nearly 300 genome-sequencing projects, with representative species from each of the three kingdoms of life, have been completed and hundreds more are underway. The raw sequence data themselves are facilitating many fascinating comparative genomics studies that are designed to identify gene-regulatory elements, to understand speciation, and to refine our idea of the evolutionary tree of life. The field of transcriptomics provides crucial information regarding the expression state, or primary genomics readout of the cell through the measurement of the abundance of RNA transcripts, thereby indicating the active components within the cell. Microarrays and serial analysis of gene expression (SAGE) represent the well-used approaches and have been applied to many model systems, as well as to the study of genes that are predominantly expressed in specific cells. Proteomics is defined as the large-scale characterization of the entire protein complement of a cell line, tissue, or organism and embraces different areas of study such as protein-protein interaction studies, protein modifications, protein function, and protein localization studies. The proteome of a cell is dynamic and will reflect the immediate environment in which it is studied in response to internal or external stimuli. A given genome can potentially give rise to an infinite number of proteomes since proteins can be modified by posttranslational modifications, undergo translocations within the cell, or be synthesized or degraded[9]. A typical proteomics experiment can be broken down into the following categories: (i) the separation and isolation of proteins from a cell line, tissue, or organism; (ii) the acquisition of protein structural information for the purposes of protein identification and characterization; and (iii) database utilization. Finally, metabolomics, considered as the newest 'omics', is focused on high-throughput characterization of small molecule metabolites (metabolome) in biological matrices. The metabolome represents the output that results from the cellular integration of the transcriptome, proteome and interactome, and therefore provides not only a list of metabolite components but also a functional readout of the cellular state that better characterize the organismal phenotype[10]. Most reactions involve more than one product and substrate, resulting in very tightly connected metabolic networks which can be sensitive to even small perturbations in the proteome. Thus, global or untargeted metabolomics is a promising means of conducting hypothesis-generating discovery studies to advance our understanding of an organism's response to normal and abnormal biological processes and to external stimuli[11]. Metabolic profiling is able to measure biological changes that are no possible to detect in genome sequencing, but could point out important pathways. The most commonly used analytical platforms are mass spectrometry (MS) coupled with separation techniques, such as gas chromatography (GC) and liquid chromatography (LC), and nuclear magnetic resonance (NMR).

1.1.2 Interactions data.

Interactions are based on the connectivity that exists among the molecular species that define the network ‘scaffold’ within the cell or system. For instance, the protein–DNA interactome covers data concerning the interactions between proteins and DNA, particularly between transcription factors and their target promoters, fundamentally define the genetic regulatory network of the cell. Determining the structure of this network is important to understand how cells modify their transcriptional state during developmental processes and in response to environmental, extracellular, intracellular and intercellular signals. Another type of interactions is the protein–protein interactions — in signalling cascades and enzyme-complex formation, for example — dictate many cellular processes. Identifying all functional protein–protein interactions will be important for understanding the structure and function of the integrated cellular network[12,13,14].

1.1.3 Functional-states data

Functional-states data reveal the overall phenotype of the cell or system and may be viewed as the outcome of the execution of the genetic program written in the DNA[8, 15]. Different functional states can be associated to only one network (formed by multiple links between components). One interesting feature of biochemical networks as they grow in size is that due to combinatorics, the number of possible functional states that they can take can grow faster than the number of components in a network. Therefore, the number of phenotypic functions derivable from a genome does not linearly scale with the number of genes. For instance, the human genome may only have 50% more genes than the genome of Caenorhabditis elegans, a small worm, but nevertheless, human beings display much more complicated phenotypes and in greater variety[15]. Thus, in general, it is hard to correlate organism complexity and functions to the number of genes its genome contains. The fundamental property of biochemical networks having many possible functional states leads to the possibility of having the same network displays many different phenotypic behaviours. To achieve the functional-state of the cell or system is essential to combine and integrate multiple *omic* platforms.

1.2 Metabolomics

The focus of metabolomic studies is shifting from cataloguing chemical structures to finding biological stories through the comprehensive insight into the metabolic status of a cellular or biological system by detecting the metabolites (metabolome) which are typically recognized as small molecules that are involved in cellular reactions. Unlike other ‘omics’ measures, metabolomics best represents the molecular phenotype (the phenotype – is produced by the genotype in juxtaposition with the environment) since metabolites and their concentrations directly reflect the underlying biochemical activity and state of cells and tissues. The metabolite profiling is complementary to the upstream biochemical information obtained from genes, transcripts and proteins and can be linked to the genotype through knowledge about biochemical pathways and gene regulatory networks[16, 17, 18].

A metabolite (or small molecule) is a low molecular weight (50 – 1500 Da) organic compound, typically involved in a biological process as a substrate or product. Some examples of metabolites include: sugars, lipids, amino acids, fatty acids, phenolic compounds, alkaloids and many more. The metabolome is inherently very dynamic: small molecules are continuously absorbed, synthesized, degraded and interact with other molecules, both within and between biological systems, and with the environment.

Metabolomics has an important impact in scientific research, the annual numbers of documents (articles, reviews, book chapter, …) with the term metabolom* in their title or abstract continue to rise, and in 2017 amounted, to 4555 (in a total exceeding 20,000) as it can be seen in Figure 2.

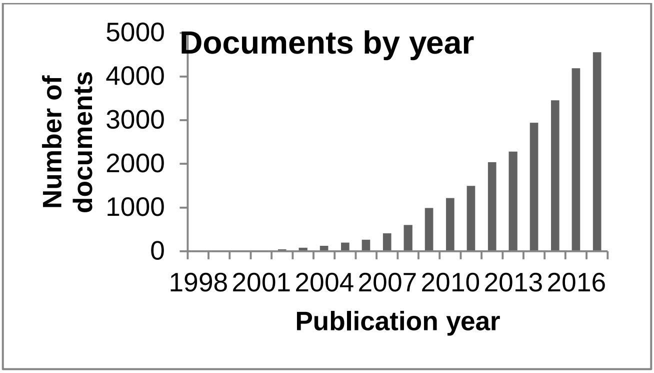


Figure 2. PubMed “Metabolom*” search results.

1.2.1 Metabolomics history

Metabolites in Ancient History

The word metabolism originates from the Greek ‘‘metabolismo′’’, which means ‘‘change.’’ The concept of metabolism was mentioned by Ibn al-Nafis (1213–1288), who stated that ‘‘the body and its parts are in a continuous state of dissolution and nourishment, so they are inevitably undergoing permanent change.’’ Studies on individual changes during different daily activities were performed by Santorio in 1614 and were mentioned in his book Ars de Statica Medecina[19].

From another perspective, metabolites have a long association with sweet flavours that extends to ancient times and predate the development of metabolic nomenclature. Ancient Chinese cultures (1500–2000 BC) recognized urine as an important source of health-related information and sweet-tasting urine as indicative of a disease (now known as diabetes). At that time, ‘‘clinical testing’’ involved actual urine tasting. Advanced ‘‘biosensor’’ ants could also be used to test differences between sample and reference urines. The association between sweet urine and disease was contemporaneously made in India by Ayurveda Hindus[19].

Metabolites in Modern History (1770–Present)

The analysis of metabolites (and in general, small organic molecules) in biological systems has been possible due to advances in organic chemistry and analytical instrumentation. The best way to understand which are the limitations and the potential of metabolomics is knowing the historical context where this new field in omic sciences has emerged (Figure 3)[20].

The first chemical composition determinations of small organic molecules, such as, lactic acid, citric acid and oxalic acid, resulted from applying analytical techniques developed by Lavoisier between 1777 and 1790, and then they were improved by Gay-Lussac and Thenard between 1810 and 1812. The compounds were separated and purified by distillation and crystallization from animal and vegetable tissues which had high content of these molecules.

During the 19th century the publication of the book written by Justus von Liebig *Animal Chemistry* (1842) represented a big advance for stablishing the basis of biochemistry reactions. Liebig inferred by first time metabolic equations based on his knowledges of organic chemistry, without having evidence of any *in vivo* reactions.

The discovery of enzymes by German chemist Eduard Buchner (1860–1917) at the beginning of the 20th century led to a focus on intracellular chemical reactions and inspired development of the field of biochemistry. The biochemical understanding of metabolism developed rapidly due to new insights into enzymatic reactions and intracellular biochemical pathways.

In 1945, practically all the analytical techniques necessary for the biochemical investigation were available for the next generation of researches. In fact, before 1957, metabolic pathways had been elucidated for almost all types of biological molecules, including lipids, carbohydrates, nucleic acid bases, amino acids and vitamins. Donald Nicholson compiled in 1955 all the metabolic reactions which were known at that time in only one map made up of about 20 metabolic pathways[21].

New technological developments in the early 20th century in the domain of mass spectrometry (MS) enabled researchers to measure the molecules involved in biochemical pathways, and to investigate their roles in disease states. As early as 1948, Williams and his associates introduced the concept that individuals might have a "metabolic pattern" that would be reflected in the constituents of their biological fluids. Utilizing data from over 200,000 paper chromatograms, many run with techniques developed in his own laboratory for this purpose, Williams was able to show convincingly that the taste thresholds and the excretion patterns for a variety of substances varied greatly from individual to individual, but that these patterns were relatively constant for a given individual[22, 19].

While Williams, an early advocate of what we would now call 'precision medicine' (Williams 1956), recognized the potential utility of such methods, the Hornings and their colleagues were at the forefront of instrumental implementations[22]. First mass spectrometer was constructed in 1905 by J.J. Thomson at the University of Cambridge (then called a parabola spectrograph)[23]. To enable the routine measurement of molecules in biological matrices, researchers in the 1950s coupled gas chromatography (GC) with MS. Early commercial GC-MS instruments enabled practical clinical investigations and the measurement of profiles of urine and blood samples. Such profiles were typically limited to a specific class of compounds, such as organic acids, because the available technology required volatile components or compounds made volatile by chemical derivatization[19]. In 1989 the introduction of electrospray introduction (ESI) technique by John B. Fenn was decisive in improving the coverage of metabolomics studies and still it continues being one of the most widely used techniques. Soon, in 1994, Richard Lerner from The Scripps Research Institute performed probably the first study of untargeted metabolomics based on LC-MS. In that study they compared the cerebrospinal fluid (CSF) from sleep-deprived felines to normal cats and they found a lipid that was unknown at the time which was accumulated in the CSF from sleep-deprived cats[24]. Finally, at the end of the 1990s the term metabolomics was coined to describe the development of approaches which aim was to measure all the metabolites that are present within a cell, tissue or organism during a genetic modification or physiological stimulus[25, 26].

To date, metabolomics has been evolving continuously and it has emerged the need for stablishing guidelines to ensure reliable metabolomics results, The Metabolomics Standards Initiative (MSI) was conceived in 2005, as an initiative of Metabolomics Society activities, now coordinated by the Data Standards Task Group of the Society. The MSI is an academic policy provider, to support the development of open data and metadata formats for metabolomics. MSI followed on earlier work by the Standard Metabolic Reporting Structure initiative and the Architecture for Metabolomics consortium (ArMet). The early efforts of MSI were focused on community-agreed reporting standards, the so called minimal information (MI) checklists and data exchange formats to support the MIs reporting standards. MSI aims were to provide a clear description of the biological system studied and all components of a metabolomics study as well as to allow data to be efficiently applied, shared and reused. Out of 2004-2007 society meetings and discussions five working groups were established. For example, the chemical analysis group proposed minimum information for reporting chemical analysis, including minimum metadata reporting related to metabolite identification. Therefore, several cross-communicating working groups were built to cover the development of different aspects of metabolomics experiment data standardisation[27].

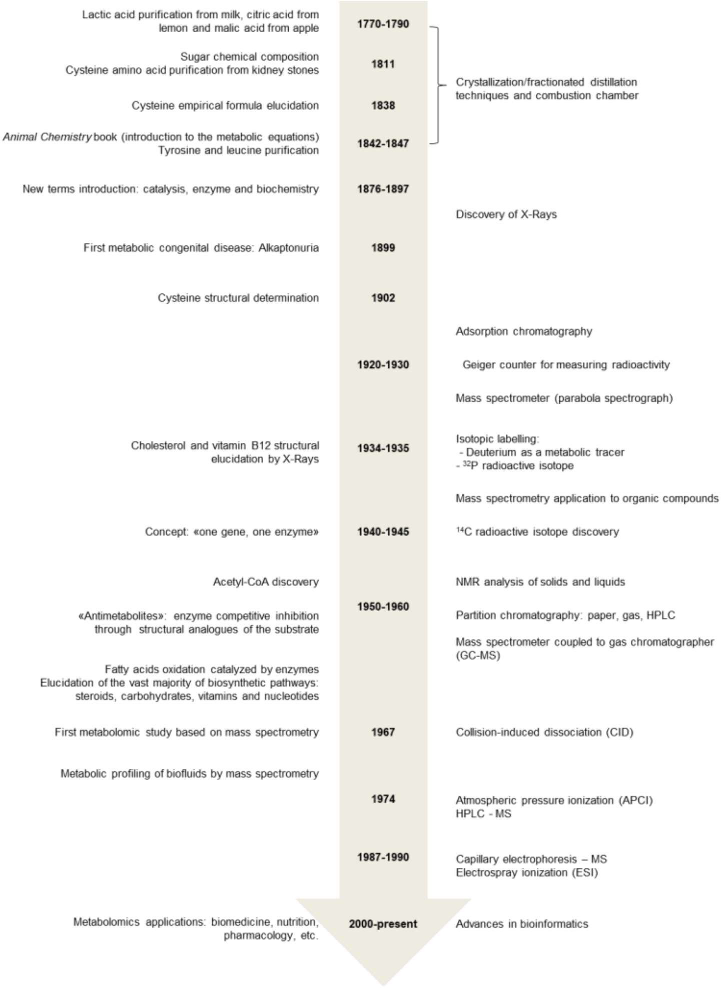

Lactic acid purification from milk, citric acid from lemon and malic acid from apple
1770-1790
Sugar chemical composition
Cysteine amino acid purification from kidney stones
1811
Crystallization/fractionated distillation techniques and combustion chamber
Cysteine empirical formula elucidation
1838
Animal Chemistry book (introduction to the metabolic equations)
Tyrosine and leucine purification
1842-1847
New terms introduction: catalysis, enzyme and biochemistry
1876-1897
Discovery of X-Rays
First metabolic congenital disease: Alkaptonuria
1899
Cysteine structural determination
1902
Adsorption chromatography
1920-1930
Geiger counter for measuring radioactivity
Mass spectrometer (parabola spectrograph)
Cholesterol and vitamin B12 structural elucidation by X-Rays
1934-1935
Isotopic labelling:
- Deuterium as a metabolic tracer
- ^{32}P radioactive isotope
Mass spectrometry application to organic compounds
Concept: «one gene, one enzyme»
1940-1945
^{14}C radioactive isotope discovery
Acetyl-CoA discovery
NMR analysis of solids and liquids
1950-1960
«Antimetabolites»: enzyme competitive inhibition through structural analogues of the substrate
Partition chromatography: paper, gas, HPLC
Mass spectrometer coupled to gas chromatographer (GC-MS)
Fatty acids oxidation catalyzed by enzymes
Elucidation of the vast majority of biosynthetic pathways: steroids, carbohydrates, vitamins and nucleotides
First metabolomic study based on mass spectrometry
1967
Collision-induced dissociation (CID)
Metabolic profiling of biofluids by mass spectrometry
1974
Atmospheric pressure ionization (APCI)
HPLC - MS
1987-1990
Capillary electrophoresis – MS
Electrospray ionization (ESI)
Metabolomics applications: biomedicine, nutrition, pharmacology, etc.
2000-present
Advances in bioinformatics

Figure 3. Main analytical advances and milestones related to metabolism study.

1.2.2 Metabolomics strategies

Metabolites constitute a diverse set of atomic arrangements when compared to the proteome (arrangement of 20 amino acids) and transcriptome (arrangement of four nucleotide bases bonded with sugar and phosphate backbone) and this provides wide variations in chemical (molecular weight, polarity, solubility) and physical (volatility) properties. The degree of diversity is indicated by the analysis of low molecular weight (MW), polar, volatile organic metabolites, such as ethanol or isoprene to the higher MW, polar (carbohydrates) and non-polar (terpenoids and lipids) metabolites. The metabolome also extends over an estimated 7–9 magnitudes of concentration (pmol–mmol). For this reason, is not technologically possible to analyse all metabolites in a single analysis. Along these lines, metabolomics analysis encompasses different strategies depending on the situation under study, as it can be seen in Table 1[28]:

Table 1. Strategies for metabolomic analysis

Metabolomic strategy	Analysis method
Metabolomics	Non-biased identification and quantification of all metabolites in a biological system. Sample preparation must not exclude metabolites, and selectivity and sensitivity of the analytical technique must be high.
Metabolite profiling	Identification and quantification of a selected number of pre-defined metabolites, generally related to a specific metabolic pathway(s). Sample preparation and instrumentation are employed so to isolate those compounds of interest from possible matrix effects prior to detection, normally with chromatographic separation prior to detection with MS. In the pharmaceutical industry, this is widely used to study drug candidates, drug metabolic products and the effects of therapeutic treatments[29, 30].
Metabolic fingerprinting	High-throughput, rapid, global analysis of samples to provide sample classification. Quantification and metabolic identification are generally not employed. A screening tool to

	discriminate between samples of different biological status or origin.
Metabolite target analysis	Qualitative and quantitative analysis of one or a few metabolites related to a specific metabolic reaction or pathway. Extensive sample preparation and separation from other metabolites is required and this approach is especially employed when low limits of detection are required. Generally, chromatographic separation is used followed by sensitive MS or UV detection.

Definitely metabolomics coverage is always a compromise between quality of data and throughput of analysis. Thus, while global metabolic profiling and metabolic fingerprinting are tools for large-metabolite coverage, the quality of the data is lower than in targeted profiling, where the method developed is exclusively optimized for one metabolite or a few metabolites [31, 32].

Although the above table shows different strategies classified according to specific and complementary aims, the term “untargeted metabolomics” has been used along this thesis encompassing more than one strategy and it has been applied to study the biological mechanism leading to diabetic retinopathy.

1.2.3 Untargeted metabolomics

Untargeted metabolomic methods are global in scope and have the aim to simultaneously measure as many metabolites as possible from biological samples without bias

Global metabolomics profiling can cover around several hundred metabolites, with concentrations encompassing several orders of magnitude. The purpose of metabolomics profiling is usually to obtain qualitative information, although quantitative or semi-quantitative analyses are also possible when different subjects or biological states are compared[33].

Although untargeted metabolomics is still an emerging science, a general workflow can be applied to carry out exploratory studies. One advantage of untargeted studies is the ability to observe changes in unknown metabolites or in metabolites not commonly reported or detected. In untargeted studies, data provide relative comparisons between samples (metabolite concentrations are not reported) compared with targeted studies that usually provide quantitative data related to metabolite concentrations.

1.2.3.1 Untargeted metabolomics workflow

Metabolomics approach has evolved the last decade with the aim of profiling the complete metabolome being able to measure more metabolites more reliably. Having a good experimental design is crucial to ensure robust scientific conclusions are reached[34]. A correct metabolomic study must guarantee the biological observations obtained from the study are significantly greater than the variation introduced by performing the study, thus, the data must be robust to avoid false biological conclusions.

The most feasible approach for comprehensive analysis of an organism metabolome is to integrate the information obtained from different analytical methodologies, aiming to increase metabolite coverage. The wide range of metabolites – which includes ionic (organic acids), polar, neutral (sugars) and non-polar (lipid) compounds – usually demands application of complementary sample-preparation protocols[31].

Figure 4 shows the typical methodological pipeline of an untargeted metabolomic study. Briefly, this methodological pipeline starts stablishing the appropriate design of scientific studies which is critical to ensure robust scientific conclusions are reached. Then, sample preparation whose aim is to extract from the biological material the widest range of metabolites which will be analysed using the most suitable analytical technique (e.g., NMR, LC-MS, or GC-MS). After acquiring the spectral data, the pre-processing will be performed in order to generate a complete set of metabolomic features. These features will be treated statistically applying univariant and/or multivariant data analysis methods to investigate: (a) the general structure of the metabolomics data in the dataset and (b) how the different metabolic features are related with the phenotypic data associated with the samples. Finally, to reach biological interpretation from the obtained results, bioinformatic resources such as pathway enrichment analysis and mapping will be very useful and the conclusions obtained from them will help to generate hypothesis[35].

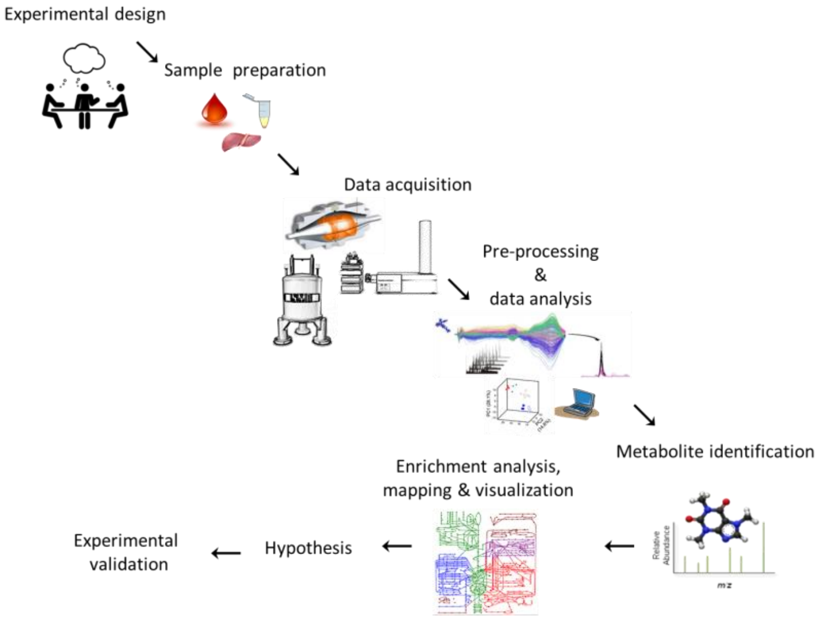


Figure 4. Metabolomics analysis workflow. Metabolomics experiment involves sample preparation, data acquisition using MS and/or NMR, pre-processing and data analysis, metabolite identification of dysregulated peaks between conditions and once a set of metabolites of interest have been identified, enrichment analysis, mapping and visualization tools can be used to gain biological insight into experimental results. The final goal after this procedure is generating hypothesis which will be further validated experimentally.

Following, more detailed explanation of each step that builds the well-known metabolic workflow is described below:

1.2.3.1.1 Experimental design

By definition, experimental design is the plan constructed to perform data-gathering studies; in other words, providing the appropriate foresight to plan a study to ensure that the variation related to the biological observations are significantly greater than process variation – the variation introduced by performing the study. Without foresight and design, large experimental datasets could provide no relevance to the biological objectives or data what are not robust and can lead to false observations and biological conclusions. As a summary, the following issues must be taken into account in order to obtain reliable biological observations[34]:

- Reproducibility of sample collection and preparation across single or multiple sites and over long periods of time.
- The requirement for multiple analytical experiments.
- Randomization of sample preparation and analysis.
- QC samples to control the process variation.

1.2.3.1.2 Sample preparation

The choice of sample-preparation method is extremely important in metabolomic studies because it affects both the observed metabolite content and biological interpretation of the data. An ideal sample-preparation method for global metabolomics should (i) be as non-selective as possible to ensure adequate depth of metabolite coverage; (ii) be simple and fast to prevent metabolite loss and/or degradation during the preparation procedure and enable high-throughput; (iii) be reproducible; and (iv) incorporate a metabolism-quenching step to represent true metabolome composition at the time of sampling[11]. In order to achieve all these requirements the next steps must be studied:

- Sample selection

One of the critical steps in metabolomics is sample selection, as the results generated will depend on its suitability. Thus, metabolomics analysis requires previous knowledge of the biological system. This information aids in the design of a suitable analytical protocol based on the nature and the particular characteristics of the sample[31].

The existence of more than one biological material available for sampling, such as, biological fluids, tissue or cells, enables the selection of the sample most suited for the analytical problem under study. Most clinical analyses are performed on biological fluids. Specifically, plasma, serum and urine have traditionally been used for diagnosis of many diseases, as they are easily collected, reflect directly the global state of an individual and allow the biological response to drug therapy to be monitored. In addition, plasma and urine provide complementary information about the state of an organism36. Thus, plasma gives an ‘‘instantaneous’’ readout of the metabolic state at the time of collection, and its composition directly reflects catabolic and anabolic processes occurring in the whole organism. Urine provides an ‘‘averaged’’ pattern of easily-excreted polar metabolites discarded from the body as a result of catabolic processes37. There are also other biofluids (e.g., saliva, amniotic fluid, cerebral spinal fluid, breast milk, synovial fluid, seminal plasma, bile, digestive fluids, or breathing air) that can provide valuable information, especially in the discovery of biomarkers for certain diseases31.

Within the field of tissue metabolomics, the analysis of these materials offers particular benefits over biofluids as the spatial description of metabolite distribution can be accomplished. For example, direct study of tumour tissues results in the profile of existing metabolites distributed in the affected tissue. Monitoring some drugs in certain tissues (e.g., brain) gives information about their mechanisms of action and effects38. In the field of biomarkers, the greatest chance of discovering a novel biomarker resides in its screening within the target tissue38. Some remarkable drawbacks of tissue metabolomics are sample heterogeneity, the small availability of tissues and the invasive character of sampling techniques31.

Finally, the study of cell cultures is one of the most extended approaches in metabolomics. Metabolome analysis of cells usually distinguishes between extracellular and intracellular metabolites, which constitute the endometabolome and the exometabolome, respectively. Thus, the samples are separated into two fractions, usually by a filtration step. Analysis of the liquid fraction (i.e., media) obtains the profile of extracellular metabolites; meanwhile, the separated cells, containing the intracellular metabolites, are subjected to extraction steps[39].

- Sample variability

Another critical factor is the variability between subjects which is more noticeable in some types of biological samples. Some matrices (e.g., plasma, serum, tissues and cerebral spinal fluid) are physiologically regulated by homeostatic control, the metabolite profile being relatively constant within healthy specimens. However, composition of urine samples can widely vary inter-individually and even intra-individually, depending on urine volume, water and food intake and other physiological conditions (e.g., age, sex, weight and environment)[40].

- Quenching and sample storage

The control of the potential sources of variability exposed above is crucial to avoid errors in data interpretation. After inter-individual and intra-individual variations, the main source of error in metabolomics is associated with sampling and post-collection procedures (e.g., freeze-thaw cycles or inadequate storage conditions). It is known that unsuitable sampling and sample pretreatment protocols can lead to biased results due to conversion or degradation of metabolites. There is increased interest in rapid collection and handling of samples for metabolomics purposes, while turnover kinetics of some metabolites is known to be extremely fast. For example, for intermediates in energy metabolism (e.g., ATP, ADP and glucose-6-phosphate), turnover rates are 1.5–2.0 s for Saccharomyces cerevisiae41. Thus, sampling techniques, particularly in the case of cells and tissues, need to be fast enough to ensure that the metabolic profile reflects in vivo conditions. Accordingly, the time window between sampling and analysis has to be as short as possible, with this aim, a quenching step (rapid change in temperature or pH) is performed.

. Nevertheless, immediate analysis of the samples is impossible in some occasions so storage is required (e.g., in banks of biological samples for research purposes)[31]. Sample storage is another critical reason for errors in metabolomics analysis. Most metabolites are preserved if samples are immediately frozen at below -80C (e.g., by using liquid nitrogen). However, it is worth noting that differences in storage time or frequent thaw/freeze cycles may have a strong influence on the development of metabolomic models.

The influence of sample storage has been widely studied in biofluids (e.g., serum and plasma). The strategies for sampling and storage of biofluids for metabolomics studies are especially important, compared to proteomics and transcriptomics. This is justified by the metabolic activity time-scale (metabolic reaction half-lives are often <1 s). Metabolic activity during sampling and storage requires stopping or minimizing changes in the metabolic profile either in concentration or structure. For this purpose, reduced temperatures during sample preparation (4^{o}C) and storage (-80^{o}C) are common[36].

- Metabolite extraction

Due to the large number of molecules with chemical and structural diversity constituting the metabolome, there is not a unique metabolite extraction protocol but a combination of them. Moreover, efforts to survey the complete metabolome rely on the implementation of multiplatform approaches based on nuclear magnetic resonance and mass spectrometry.. However, these two analytical platforms are characterized by distinctive physicochemical principles making even more difficult to use a unique metabolite extraction protocol.

For metabolomics applications, a fast, reproducible, unselective extraction method is preferred for detecting the widest range of metabolites, avoiding unforeseen chemical modifications. In general, metabolites of interest are extracted by liquid extraction with one solvent, aqueous or organic, or with a combination of solvents (liquid-liquid extraction), implying that the type of metabolites extracted depends on the chemical properties of the solvent used. For a certain class of metabolites, a particular solvent can be more adequate, yet not unique for its extraction. The modification of some parameters, such as changing the temperature or composition of the extraction solvent will have effects on the metabolites that can be detected and studied. Hence, the protocol for metabolite extraction plays an extremely important role in determining the scope of metabolites that can be studied. Better understanding of extraction parameters will allow optimization of metabolite extraction protocols, leading to more robust metabolomics studies.

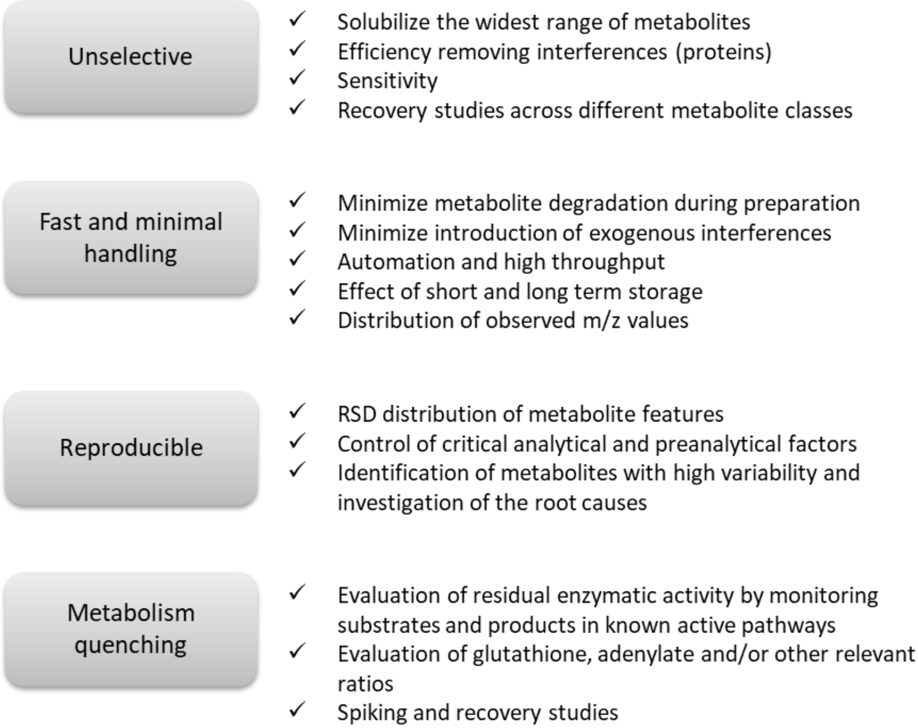
Generally polar organic solvents such as methanol, ethanol, acetonitrile and acetone, are typically mixed with water to extract hydrophilic metabolites while chloroform can be used to extract hydrophobic metabolites. However a major advantage of the methanol/chloroform/water method is the simultaneous extraction of both hydrophilic and hydrophobic metabolites into two different compartments, which is especially important when addressing the lipid metabolome.

A large number of studies among others Yanes et al. [1], Ser et al. [42], Dietmair et al. [43], El Rammouz [44] and Masson et al. [45], have focused on examining the effects of critical extraction parameters on the quantitation of a wide variety of metabolites. Yanes et al. [1] carried out a study in which they compared different metabolite extraction methods for LC-MS analysis in which different fundamental conditions for metabolite solubility and stability were varied such as solvent polarity, temperature, pH, and molecular weight cutoff filtering. Their results could be interpreted in decreasing order of efficiency in extracting both polar and nonpolar metabolites simultaneously as follows: Hot EtOH/Ammonium Acetate < BoilingWater < Cold EtOH/AmmoniumAcetate < MPA < Hot MeOH < Acetone/MeOH = YM3 – cold aqueous solvent buffered at pH 7.2 and centrifugal filter unit with a cutoff of 3 kDa.

While LC-MS can cover a wider range of metabolite classes, GC-MS analysis is more focused on detecting organic acids, sugars, amino acids, and steroids which are nonvolatile and must be derivatized prior to analysis by GC to reduce their polarity and facilitate chromatographic separation on a column of low polarity. Organic acids were commonly esterified by silylation, predominantly trimethylsilylation or tert-butyldimethylsilylation. In addition, keto-(oxo-) groups are usually oximated in order to improve their GC properties and prevent enolization reactions which can introduce multiple products, thereby complicating the chromatograms[46].

Samples for NMR-based metabolomics can be analyzed without extraction of metabolites provided that the active volume of the NMR probe is filled, typically with 500 µL of biofluid for 5 mm NMR tubes. However, the intrinsic low resolution of NMR greatly hinders accurate assignments and quantification of a large number of metabolites from biofluids[47, 48, 49] or intact tissues[50, 51]. In this context, extraction of metabolites may reduce broad signals in the spectra arising from the high overlap of chemical shifts for metabolites and macromolecules (e.g., proteins) and generate narrower and better-resolved NMR resonances that allow reliable quantification of metabolites[52]. In general, most of the same extractions protocols for LC-MS could be also applied in NMR.

To conclude sample preparation section, Figure 5 summarizes the characteristics of an ideal method for untargeted metabolomics and overview of aspects to consider during method development and evaluation[11].



Unselective

- ✓ Solubilize the widest range of metabolites
- ✓ Efficiency removing interferences (proteins)
- ✓ Sensitivity
- ✓ Recovery studies across different metabolite classes

Fast and minimal handling

- ✓ Minimize metabolite degradation during preparation
- ✓ Minimize introduction of exogenous interferences
- ✓ Automation and high throughput
- ✓ Effect of short and long term storage
- ✓ Distribution of observed m/z values

Reproducible

- ✓ RSD distribution of metabolite features
- ✓ Control of critical analytical and preanalytical factors
- ✓ Identification of metabolites with high variability and investigation of the root causes

Metabolism quenching

- ✓ Evaluation of residual enzymatic activity by monitoring substrates and products in known active pathways
- ✓ Evaluation of glutathione, adenylate and/or other relevant ratios
- ✓ Spiking and recovery studies

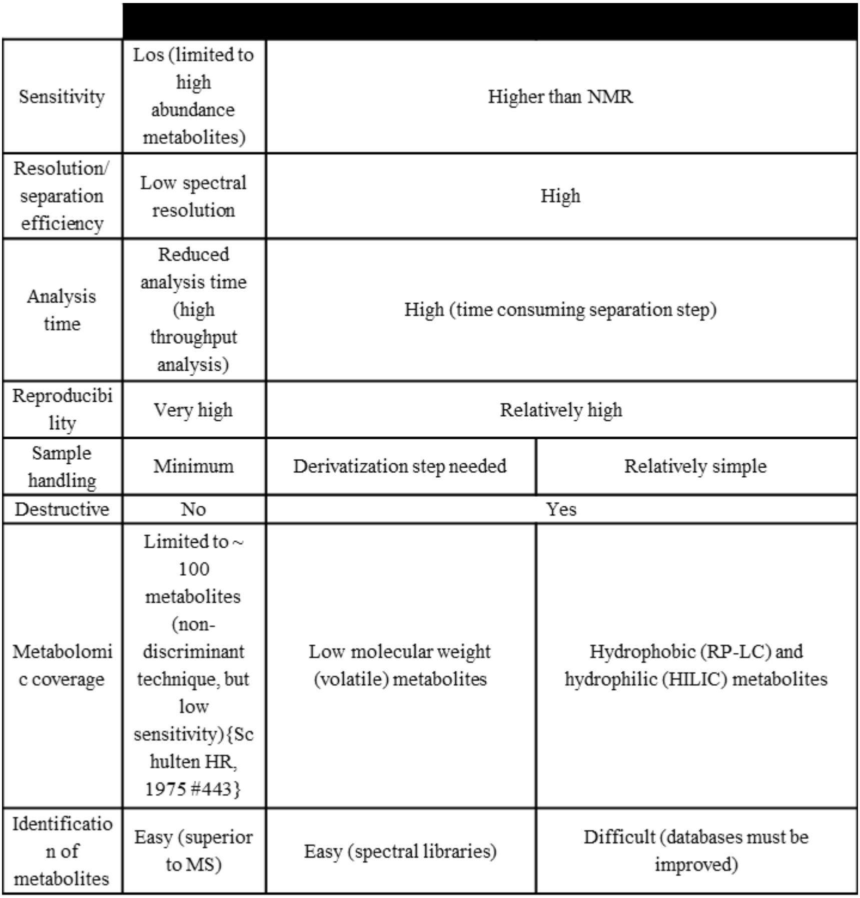
Figure 5. Summary of the characteristics of an ideal sample-preparation method for untargeted metabolomics and overview of aspects to consider during method development and evaluation.

1.2.3.1.3 Data acquisition

The human metabolome represents the complete collection of small molecules found in the human body including peptides, lipids, amino acids, nucleic acids, carbohydrates, organic acids, biogenic amines, vitamins, minerals, food additives, drugs, cosmetics, contaminants, pollutants, and just about any other chemical that humans ingest, metabolize, catabolize or come into contact with. All these small molecules result in more than 100,000metabolites[53]. Considering the large number of possible metabolites, their chemical diversity and the large range of possible concentrations (11 orders of magnitude observed for human serum[54]), there is no single analytical technique that can achieve full coverage of the entire metabolome since extraction, separation and analytic techniques that work for one class of metabolites are often useless for others. With nucleic acids and proteins one detection technique will usually suffice, but the metabolomic community must rely on a suite of detection techniques. A range of analytical techniques, including GC-MS, LC-MS and NMR are required in order to maximize the number of metabolites that can be identified in a matrix[31].

Table 2 shows the main characteristics that define the three major technologies (NMR, GC-MS, and LC-MS) used in modern metabolomic studies[55].

Table 2. A comparison of different metabolomics technologies



	[redacted]	[redacted]	[redacted]
Sensitivity	Los (limited to high abundance metabolites)	Higher than NMR	
Resolution/ separation efficiency	Low spectral resolution	High	
Analysis time	Reduced analysis time (high throughput analysis)	High (time consuming separation step)	
Reproducibility	Very high	Relatively high	
Sample handling	Minimum	Derivatization step needed	Relatively simple
Destructive	No	Yes	
Metabolomic coverage	Limited to ~ 100 metabolites (non-discriminant technique, but low sensitivity){Schulten HR, 1975 #443}	Low molecular weight (volatile) metabolites	Hydrophobic (RP-LC) and hydrophilic (HILIC) metabolites
Identification of metabolites	Easy (superior to MS)	Easy (spectral libraries)	Difficult (databases must be improved)

- LC/GC-MS data acquisition

Chromatography

Most MS applications in metabolomics use a separation method before mass detection, typically LC, GC or capillary electrophoresis (CE). GC-MS and LC-MS are widely used techniques and can detect a wide variety of compounds. However, the configuration of MS instruments for these two methods is distinct due to the ionization procedures used; GC-MS instruments make use of the hard-ionization method, electron-impact (EI) ionization, while LC-MS mostly uses soft-ionization sources (e.g., atmospheric pressure ionization (API) (e.g., electrospray ionization (ESI)) and atmospheric pressure chemical ionization (APCI))[56]Liquid chromatography

LC is probably the most versatile separation method, as it allows separation of compounds of a wide range of polarity with little effort in sample preparation (compared to GC-MS). Given the nature of the analytical problem in untargeted metabolite profiling (analysis of samples of unknown composition comprising a complex mixture of unknown analytes covering a wide range of structures and divergent physicochemical properties) the successful application of LC–MS in metabolomics requires the employment of highest possible resolution power available to the investigator. As a result utilisation of the smallest particle size (e.g. sub-2 m particles), as used for U(H)PLC, has become the gold standard for LC as it increases peak capacity and sensitivity compared to HPLC while decreasing the risk of matrix effects[57, 58]. In metabolite profiling these parameters play a significant role because superior resolution and/or sensitivity mean better analyte (metabolome) coverage[59,]. Typically 2.1 mm i.d. columns are used as these represent the mostly used U(H)PLC column format[60, 61].

Using reverse-phase silica-based particles columns, semi-polar compounds (phenolic acids, flavonoids, glycosylated steroids, alkaloids and other glycosylated species) can be separated, and, using hydrophilic columns, polar compounds can be measured (sugars, amino sugars, amino acids, vitamins, carboxylic acids and nucleotides)[56].

Reversed-phase (RP) liquid chromatography using gradient elution is currently by far the mostly used separation mode for metabolome characterisation and analysis as it is directly compatible with the analysis of aqueous samples[62, 1]. RP separation media cover a large part of the metabolome, and at the same time provide the most reliable, robust and sophisticated LC stationary phases, while the number of available chemistries, geometrical characteristics of particles and columns surpasses the corresponding numbers for all other modes. Typical published protocols on the analysis of urine, plasma and tissue extracts are based on RPLC[59-60, 61].

However RPLC faces limitations in the analysis of polar and/or ionic species due to poor retention of such molecules. This means that many polar primary metabolites (including numerous aminoacids, amines, organic acids, sugars and carbohydrates) cannot be effectively analysed in RPLC. These metabolites are of the highest importance as they are involved in a multitude of biochemical pathways and fluxes. Alternative separation mechanisms are needed to provide sufficient retention and thus resolution of polar analytes. As such hydrophilic interaction chromatography (HILIC), ion-exchange LC or aqueous normal phase chromatography can be applied [63, 64, 65].

HILIC-MS provides separations complementary to those obtained by RPLC–MS in that early eluting analytes in the RP mode are often well retained by HILIC. This enables analysis of many of the polar primary metabolites mentioned above. In addition, as HILIC employs mobile phases rich in organic solvents, high MS sensitivity can be achieved as a result of increased ionisation efficiency. On the other hand when using HILIC there are constrains such as allowing sufficient re-equilbration time to reach acceptable analytical repeatability[66, 61].Gas chromatography

Comparing to LC-MS and NMR, gas chromatography–mass spectrometry (GC-MS) is the most standardized method in metabolomics, with almost 50 years of established protocols for analysis of metabolites such as sugars[67], amino acids[68], sterols[69], hormones[70], fatty acids[71], aromatics[72], and many other intermediates of primary metabolism[73].

GC-MS is a combined system where volatile and thermally stable compounds are first separated by GC and then eluting compounds are detected traditionally by electron-impact mass spectrometers. In metabolomics, GC-MS has been described as the gold standard[74], although it is biased against non-volatile, high-MW metabolites. Volatile, low-MW metabolites can be sampled and subsequently analysed directly. However, the majority of metabolites analysed require chemical derivatisation at room or elevated temperatures to provide volatility and thermal stability prior to analysis[28].

In most of the metabolomic applications, the derivatized metabolites which are predominantly analyzed as TMS derivatives, are introduced into a heated injector (200–250 °C), where rapid vaporization and mixing with the carrier gas occurs (usually helium), followed by chromatographic separation of metabolites on the GC column and subsequent MS detection[75, 76]. Sample can be injected in either split or splitless mode. In a splitless system, the advantage is that a larger amount of sample can be introduced into the column. However, a split system is preferred when the detector is sensitive to trace amounts of analyte and there is concern about sample overloading of the column. Therefore, in metabolomic studies, split mode is generally preferred because metabolites are present in wide range of concentrations. Chromatograms in metabolomic studies are complex due to large number of metabolite peaks as wells as multiple derivatization products. Therefore, long analysis time (up to 60 min) may be needed for satisfactory chromatographic separation. The most important factors which influence chromatographic separation include column properties (length of the column, stationary phase, internal diameter (i.d.)), carrier gas type, carrier gas velocity and oven temperature program[77].

Capillary GC columns made of fused silica are commonly employed in GC/MS based metabolomic studies. These columns have a thin film of liquid phase bonded to the walls of a narrow i.d. (0.25mm or 1.8mm) column. Capillary GC columns can operate at very higher temperatures and provide significantly higher chromatographic resolution. Because of the small i.d. of these columns, the sample capacity of the 0.25mm columns is limited to about 50–100 ng per component of a mixture. Columns with varying polarity (DB-1 to DB-50), varying chemical composition of stationary phases, and varying lengths (10 to 60 meters) have been utilized in metabolomic analysis, in particular DB-5MS[78] columns or columns with equivalent stationary phase (HP-5MS[79] and RTX-5MS[80]). A DB-5MS column is a fused silica capillary column, chemically bonded with a 5% diphenyl cross-linked 95% dimethylpolysiloxane stationary phase (0.25m film thickness). The capillary column is held in an oven that can be ramped continuously or in steps to achieve desired separation. As the temperature increases, those compounds that have low boiling points elute from the column sooner than those that have higher boiling points. Therefore, there are actually two distinct separation forces, temperature and stationary phase interactions. Typical column oven temperatures range from 40 to 325 ^{o}C[81]. Maximum temperature that can be used on a particular column should always be verified with the manufacturer's instructions. The rate at which a sample passes through the column is directly proportional to the temperature of the column. The higher the column temperature, the faster the sample moves through the column. However, the faster a sample moves through the column, the less it interacts with the stationary phase and the chromatographic separation is poor. Similarly, the carrier gas flow rate also affects the analysis. The higher the flow rates the faster the analysis, but the lower the separation between analytes. Selecting the flow rate is therefore the same compromise between the level of separation and length of analysis as selecting the column temperature. Column flow rate between 0.8 and 2 mL/min is commonly used for metabolic profiling[82]. By optimizing various GC parameters, it is possible to separate most if not all of the endogenous metabolites before they enter the mass spectrometer for detection[77].

Mass spectrometry

Mass spectrometry (MS) is intrinsically a highly sensitive method for detection, quantitation, and structure elucidation of up to several hundred metabolites in a single measurement. The sensitivity and accuracy of detection by MS are dependent on the nature of the experimental conditions and the instrumental settings; major factors that contribute include the nature of metabolite extraction, separation, ionization (and possibly ion suppression), and detection approaches[83]. Ion source, mass analyzer, and detector are the three basic components necessary to perform metabolites detection (see Figure 6). The molecular or fragment ions are first produced in the ionization chamber, and then they are transferred in the mass analyzer region via several ion optics (electromagnetic elements like skimmer, focusing lens, multipole, etc.), which basically focus the ion stream to maintain a stable trajectory of the ions. The mass analyzer sorts and separates the ions according to their mass to charge ratio (m/z value). The separated ions are then passed to the detector systems to measure their concentration, and the results are displayed on a chart called a mass spectrum (see Figure 7). Since the ions in the gas phase are very reactive and often short lived, their formation and manipulation should be conducted in high vacuum. For this reason, the ion optics, analyzer, and also the detectors are kept at very high vacuum (typically from 10−3 torr to 10−6 torr pressure). Mass spectrometers typically use either oil diffusion pumps or turbomolecular pumps to achieve the high vacuum required to operate the instrument.

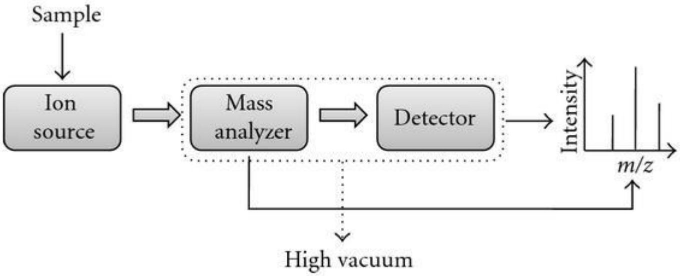


Figure 6. The basic components of the mass spectrometer

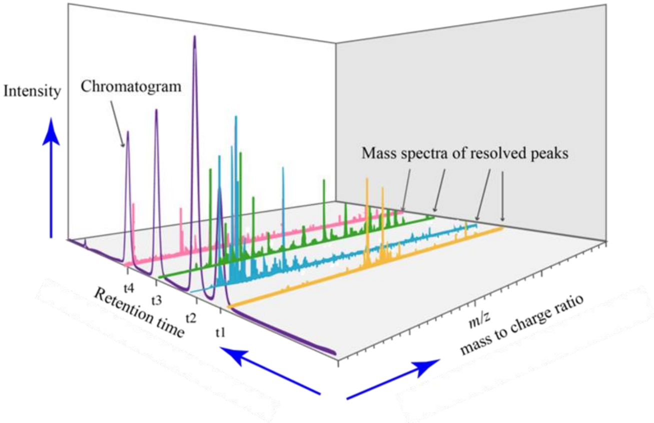


Figure 7. LC/GC - Mass Spectrum.

Ionization.

Ionization is one of the most critical steps in MS-based metabolite measurements. The degree of its ionization determines the ability to detect and quantify a metabolite. The most often used ionization methods in the field of metabolomics are electrospray ionization (ESI) and electron impact (EI) ionization.

ESI is the favourite ionization technique for LC-MS for multiple reasons. It adequately ionizes molecules in the liquid phase and can universally be used for small molecules (<1,000 amu) as well as for large molecules such as peptides and proteins. Moreover, ESI is a soft ionization technique, so it does not induce a significant fragmentation of the molecular ions.

Generally, a dilute analyte solution is injected by a mechanical syringe pump through a hypodermic needle or stainless steel capillary at low flow rate (typically 1–20 μL/min). A very high voltage (2–6 kV) is applied to the tip of the metal capillary relative to the surrounding source-sampling cone or heated capillary (typically located at 1–3 cm from the spray needle tip). This strong electric field causes the dispersion of the sample solution into an aerosol of highly charged electrospray (ES) droplets (see Figure 8). A coaxial sheath gas (dry N_2) flow around the capillary results in better nebulization. This gas flow also helps to direct the spray emerging from the capillary tip towards the mass spectrometer. The charged droplets diminish in size by solvent evaporation, assisted by the flow of nitrogen (drying gas). Finally the charged analytes are released from the droplets, some of which pass through a sampling cone or the orifice of a heated capillary (kept in the interface of atmospheric pressure and the high vacuum) into the analyser of the mass spectrometer, which is held under high vacuum[84].

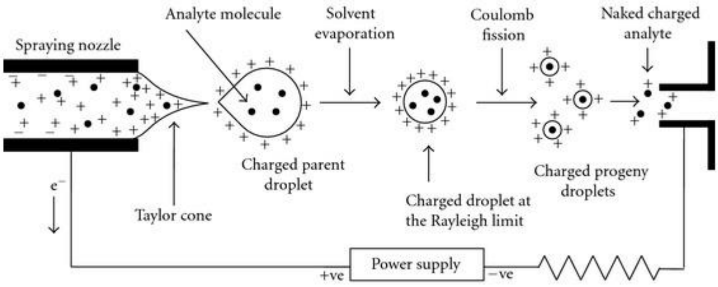


Figure 8. Schematic representation of the electrospray ionization process (adopted from [84]).

A drawback in using ESI is that its ionization efficiency is deleteriously affected by the presence of salts, so the chromatography methods are limited to the use of only volatile buffers such as ammonium acetate or ammonium formate. In addition, ion suppression can occur when co-eluting metabolites compete for a limited number of molecular ions with low electron or proton affinity metabolites are obscured, or not detected at all[83].

On the other hand, EI is the widely used ionization method for GC. It is performed in a high-vacuum ion source (10−7 to 10−5 mbar, 200–250 ◦C), where analytes in vapor state are bombarded with electrons at 70 eV[82]. This gives the sample molecules a great deal of excess energy and many fragment ions are formed. Fragmentation pattern is characteristic to a particular molecule and therefore can be useful in determining the structure of the analyte. Unfortunately, some compounds fragment completely and do not give molecular ions. Therefore, chemical ionization (CI) has also been utilized in some of the metabolomic studies[85]. CI is a relatively softer ionization technique. Thus, CI produces less fragmentation compared to EI. CI can produce molecular ions for some volatile compounds that do not give molecular ions in EI. For metabolomic applications, MS is typically utilized in full scan mode. Broad range mass fragments of m/z from 50 to 700 are generally monitored.

Mass analyser and detection

Metabolite detection with high resolution and sensitivity is generally desired. However, achieving both goals in a single MS detection mode is challenging because as a general rule higher sensitivity leads to lower resolution and vice versa. There are various options including single (MS) or tandem (MS/MS) mass analyzers to choose from, each of which has different sensitivity and resolution performance. The single-configuration mass analyzers include the quadrupole (Q), linear ion trap (LIT), quadrupole ion trap (QIT), time of flight (TOF), Fourier transform ion cyclotron resonance (FTICR), and Orbitrap. Quadrupole and ion trap analyzers offer high sensitivity, but limited resolution whereas TOF, FTICR, and Orbitrap offer high mass resolution. Mass analyzers arranged in a tandem configuration include triple-quadrupole ion trap (QTrap), triple quadrupole (TQ), quadrupole-TOF (Q-TOF), and linear quadrupole ion trap-Orbitrap (LTQ-Orbitrap). Because of their high sensitivity and selectivity, TQ and Qtrap analyzers are the most common MS spectrometers hyphenated to LC and employed in targeted metabolic studies, while Q-TOF and Q- Orbitrap (Q-Exactive)are more suitable for global profiling and metabolite identification (including isotopomer analysis) due to their higher mass-resolving power. Mass analyzers used with GC are usually single quadrupoles or TOFs, but some recent GC-MS instruments are now equipped with QTOF or TQ mass spectrometers[83, 86, 87]. Figure 9 shows a schematic representation of three analyzers: QIT, TQ and Q-TOF.

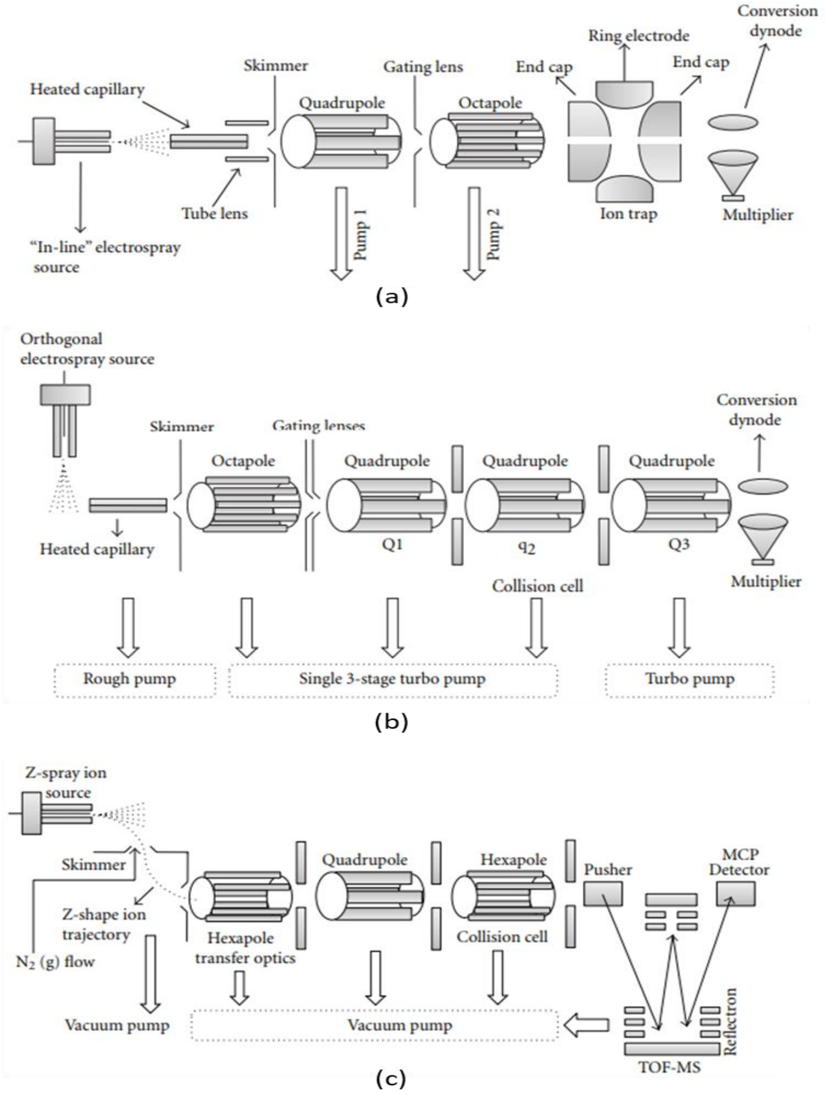


Figure 9. (a) quadrupole ion trap (QIT), triple quadrupole (TQ) and quadrupole-TOF (Q-TOF) (adopted from [84]).

- NMR data acquisition

In metabolomics NMR spectroscopy provides a rapid, non-destructive, high-throughput method that requires minimal sample preparation[88]. NMR spectroscopy functions by the application of strong magnetic fields and radio frequency (RF) pulses to the nuclei of atoms. For atoms with either an odd atomic number (e.g., 1H) or odd mass number (e.g., 13C), the presence of a magnetic field will cause the nucleus to possess spin, termed nuclear spin. Absorption of RF energy will then allow the nuclei to be promoted from low-energy to high-energy spin states, and the subsequent emission of radiation during the relaxation process is detected[28].

The majority of applications employ 1H (proton) NMR for clinical studies and, as the majority of known metabolites contain hydrogen atoms, the system is nonbiased to particular metabolites, unlike other techniques (i.e. all metabolites at concentrations greater than instrument limit of detection are detected). Although less frequent, other atoms like carbon (13C-NMR) and phosphorus (31P NMR) are also targeted by NMR, providing additional information on specific metabolite types (Reo, 2002)[35]. Initially, sensitivity depends on the natural abundance of the atom studied (1H, 31P, 19F 100%; 13C 1.10%; 15N 0.37%) though improvements in sensitivity can be obtained by longer analysis times, application of higher magnetic fields and the use of cryogenic probes [89, 28].

The NMR spectrum (specifically the chemical shift) depends on the effect of shielding by electrons orbiting the nucleus and it is calculated as the difference between the resonance frequency and that of a reference substance, subsequently divided by the operating frequency of the spectrometer (Blümich and Callaghan, 1995)[35]. The chemical shift for 1H NMR is determined as the difference (in ppm) between the resonance frequency of the observed proton and that of a reference proton present in a reference compound (for 1H NMR experiments, tetramethylsilane in solution, set at 0 ppm). The measured chemical shifts vary: 0–10 ppm for 1H (Figure 10); and, 0–250 ppm for 13C. The signal intensity depends on the number of identical nuclei, and the presence of complex samples does not interfere with measured intensity as ionisation suppression does with electrospray ionisation. This allows quantification to be performed[28].

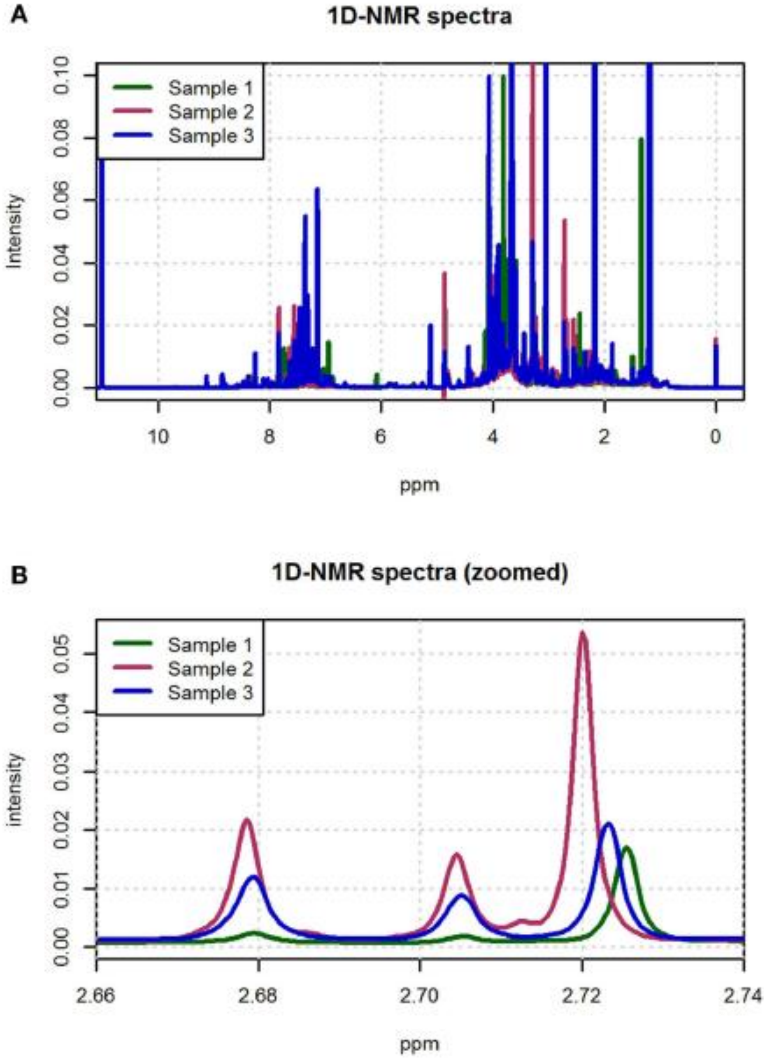

A
1D-NMR spectra
Sample 1
Sample 2
Sample 3
Intensity
0.00
0.02
0.04
0.06
0.08
0.10
10
8
6
4
2
0
ppm
B
1D-NMR spectra (zoomed)
Sample 1
Sample 2
Sample 3
intensity
0.00
0.01
0.02
0.03
0.04
0.05
2.66
2.68
2.70
2.72
2.74
ppm

Figure 10. Examples of spectra obtained with 1H-NMR. (A) An example of three spectra obtained with 1D 1H-NMR. (B) A zoomed view of the spectra in (A) in the 2.66–2.74 ppm range (adopted from 35).

Although 1D-NMR is the most commonly used method in high-throughput metabolomics studies, conversely, two dimensional NMR (2D-NMR), which is based on two frequency axes, is often restricted to the characterization of those compounds that cannot be identified with 1D-NMR spectra. The second dimension in 2D-NMR allows to separate otherwise overlapping spectral peaks and, therefore, gives additional and important information on the chemical properties of the metabolite[90]. Although 2D-NMR generates a large number of different spectra, these can be globally classified into homonuclear (i.e., 1H–1H-NMR) and heteronuclear (i.e., 1H–13C or 1H–15N) spectra[91]. There are also different pulse sequences used to generate the 2D-NMR spectra such as correlation spectrometry (COSY), total correlation spectroscopy (TOCSY), and nuclear Overhauser effect spectroscopy (NOESY). NMR spectroscopy is a high-throughput fingerprinting technique. Crude samples are mixed with a reference compound solution (e.g., tetramethylsilane dissolved in D2O for 1H NMR), added to an NMR probe (generally less than 2 mL), inserted into the instrument and analysed. Normally, the application of 96-well plates and flow injection probes allows the analysis of hundreds of samples per day. NMR probes are generally high µl volume based and that adds constraints on sample volume required. However, the introduction of 1 mm µl probes has enabled 2 µl volumes to be analysed, hence allowing smaller volume invasive sampling of study subjects, which is important for small-animal-based studies[89, 28].

1.2.3.1.4 Data processing

Metabolomics is basically focused on the study of biological systems responses resulting in metabolite level regulation related to genetic variation or multitude of environmental changes, therefore it is important to separate interesting biological variation from obscuring sources of variability introduced in studies of metabolites, including at various stages of data processing. The quality of data processing is an essential step for our ability to properly analyze and interpret metabolomic data[92].

Once the samples have been analyzed by some of the analytical platforms, the acquired data must be processed using bioinformatic software that performs quantitative analyses to find compounds that are significantly different between sample groups. Recent technological advances in nuclear magnetic resonance and mass spectrometry are significantly improving our capacity to obtain more data from each biological sample. Consequently, there is a need for fast and accurate statistical and bioinformatic tools that can deal with the complexity and volume of the data generated in metabolomic studies[35], [93].

- LC/GC – MS data processing

Mass spectrometry is an analytical technique that acquires spectral data in the form of a mass-to-charge ratio (m/z), retention time and a relative intensity of the measured compounds which are defined as features[35].

In mass spectrometry-based metabolomics, the starting point for data processing is a set of rawdata files, each file corresponding to a single biological sample. A single LC/GC–MS data file is a collection of successively recorded histograms, each representing hits of ionized molecules on the detector during a small time frame which are characterized by a number of m/z and intensity data points (see Figure 11)[92].

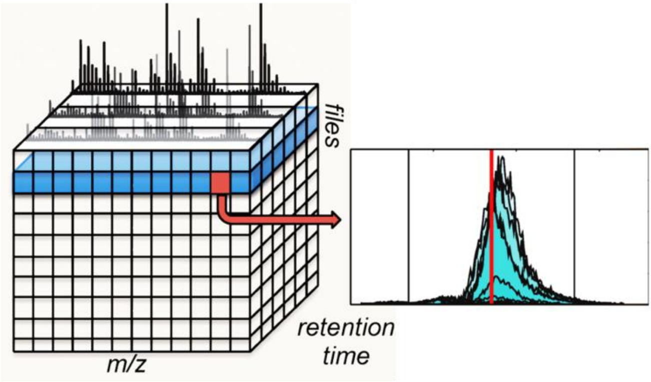


Figure 11. Collection of LC/GC–MS data files (adopted from [94]).

The basic aim of data processing is to transform raw data files into representation that facilitates easy access to characteristics of each observed ion. These characteristics include m/z and retention time of the ion and an ion intensity measurement from each raw data file. In addition to these basic features, data processing can extract additional information like isotope distribution of the ion[92].

As it can be seen in Figure 12, a typical data processing pipeline usually proceeds through multiple stages, which are explained below.

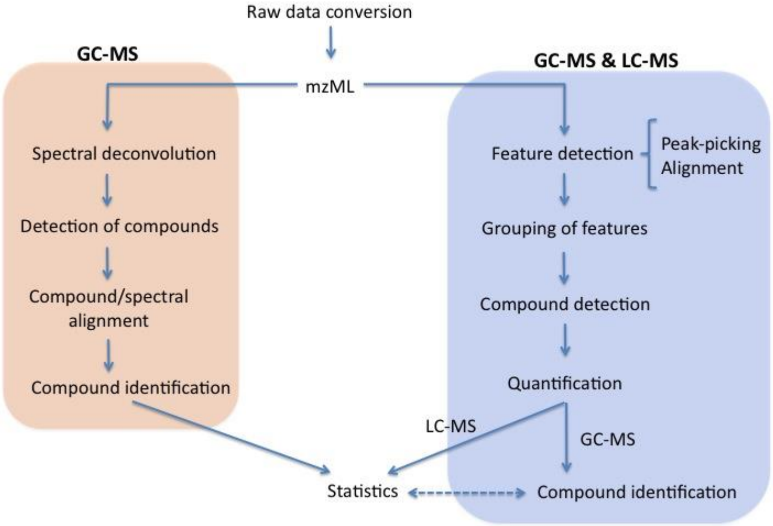


Figure 12. Flowchart for data processing in LC-MS and GC-MS mass spectrometry

The metabolomic data processing steps are explained below:

i. 1. Raw data conversion

Since different instrument vendors utilize different proprietary data formats it is necessary to convert such raw proprietary data into common raw data format such as mzXML[95] as a preliminary step for data processing.

Usually, mzML files can be produced by the following two routes. The first approach is to use vendor provided converters and exporters. Sometimes they are easy to find as part of the main data analysis software (e.g., File→Export as…→mzML), sometimes they consist of dedicated command line tools, which usually have the advantage of being capable of doing batch conversion for many files at once. The second approach to perform file conversion is to rely on community projects, which have developed independent converters. The Proteowizard (http://proteowizard.sourceforge.net/) team have obtained licenses to redistribute most of the required vendor Dynamic-Link Libraries (DLLs). The Proteowizard msconvert tool can convert from most of the LC-MS data, as well as some of the GC-MS formats, to mzML. The msconvert tool also allows a set of operations (like simple processing or filtering) on the input data.

i.2. Feature detection

It is known as “feature” all raw data points that originate from one particular ion species. A *feature* is characterised by a retention time, an m/z value and an intensity (RT, mz, intensity).

The purpose of the feature detection stage is to identify all signals caused by true ions and avoid detection of spurious signal. This step also aims to provide as accurate quantitative information about ion abundance as possible. Feature detection is an essential step in the metabolomic data processing pipeline, yet in practice rarely performed perfectly. This is therefore an important area for further method development[92, 35].

Feature detection is usually performed on individual input files (samples), and can easily be parallelized. The general question, to detect a signal among potentially noisy data, can be tackled with concepts commonly used in signal processing applications. One of the most intuitive feature detection algorithms is the filter approach, where the raw data are processed through filter functions that resemble the shape of the feature to be detected: where the signal will show a good similarity with the shape of the filter, a feature will be detected. Examples of typical filters are the mexican hat filter function, or more general wavelet transformations. Due to the large number of raw data points, the algorithms often select regions of interest from the raw data with fast algorithms first, before the (computationally expensive) filter functions are applied. The last step in the feature detection is the calculation of the m/z, RT and intensity characterizing each feature. The actual m/z can be derived, e.g., as an intensity-weighted average, the retention time is the position of the maximum intensity (the so-called peak apex), and the intensity could be calculated as maximum count (peak height) or the area under the curve of the feature (peak area).

One of the most important characteristics of a feature detection algorithm is the ability to distinguish features from the noise. Typical noise in mass spectrometry consists of electronic and chemical components. Electronic noise is an inherent characteristic of the mass spectrometer and is — depending on the mass trace — relatively constant across the chromatography. Noise reduction methods aim at removing electronic noise from the measured signals. These methods are typically implemented using traditional signal processing techniques such as filtering with moving average window[96], median filter[97] in chromatographic direction and wavelet transformation[97] in m/z direction[92].

Chemical noise originates mostly from the chromatographic system, and is generated by phenomena such as column bleeding or by the presence of contaminants within the mobile phase. Chemical noise is typically dependent on the mass trace and contributes to a noisy baseline of a chromatogram. Depending on the software package, several quite different approaches can be used for feature detection.

Baseline removal is typically a two-step process: (1) finding the baseline shape and (2) subtracting the shape from the raw signal. For example, Haimi et al. [98] estimate the baseline by first segmenting a spectrum and then performing a linear regression through the lowest points of smoothed spectrum segments[98, 92].

i.3. Feature alignment

Due to the nature of LC/GC - MS data, it is normal to find slight analytical differences between the instrumental measurements resulting in small deviations of the m/z values or slight shifts in the retention times. As a consequence, the "same" feature will be characterized by slightly different m/z and RT values in different samples. For this reason, in order to be sure of comparing the same feature across the different samples, it is necessary to perform feature grouping (including optionally a retention time correction step) to go from individual feature lists to a rectangular data matrix (**¡Error! No e encuentra el origen de la referencia.**) which can be required for subsequent statistical analysis.

To determine which features in a sample should be linked to corresponding features in other samples, several approaches have been proposed, implemented and compared[99]. Most of the approaches are variations of some sort of clustering based on m/z and retention time. This is implemented in e.g., OpenMS[100] as FeatureLinker using the "unlabeled_qt" algorithm[101]. A different approach is implemented in XCMS[102, 103], which uses extracted ion chromatograms (EIC): all features from all files are mapped to m/z slices. Then it is possible to calculate the density of features along the retention time, and group all features which have “the same” retention time. Here “same” retention time depends on the stability of the chromatography. The important parameters are the width of the EIC-like slices and the amount of smoothing in the density estimation. Once the groups of features are collected, filtering is performed. The most important parameter is the absolute or relative minimum occurrence of features across the samples, to avoid groups with just one or very few features. If the experiment contains different sample classes, it might be that for instance a metabolite is present in the wild-type, while it is absent in the mutant. So the minimum number of occurrences should be calculated on a per-class level.

As mentioned above, the retention time for the same metabolite can vary across different runs. To correct for this, one can try to align or warp all measurements, so that the retention deviations are minimized. Several approaches are possible. The first approach operates directly on the raw spectra: given two sequences of spectra, the pairs of retention times that maximise the similarity of the corresponding spectra are found. Then, a global alignment is created from the individual pairwise alignments. The second class of algorithms requires the extracted feature lists and uses so-called landmark peaks, i.e. features that can be reliably found across (almost) all samples. Then a smoothed curve can be fitted to adjust the retention time in each sample to minimize this retention deviation. A third approach uses optimization to find a deformation (called a warping) of the time axis that leads to maximal overlap of EICs, without first defining spectra or landmark peaks. An example of the latter is Parametric Time Warping[104], implemented in the R package ptw[105], [106]. This can be used to align both EICs and peak lists.

i.4. Parameter optimisation and performance metrics

Several metrics can be used to calculate and visualise the performance of the LC-MS data. Run-order effects can be detected in PCA plots if the samples are coloured in e.g. a rainbow color gradient (following the injection order).

In order to remove the unwanted systematic bias in ion intensities between measurements, while retaining the interesting biological variation, normalization can be applied. Chemical diversity of metabolites, leading, for example, to different recoveries during extraction or responses during ionization in mass spectrometer, makes separation of interesting biological variation and unwanted systematic bias a difficult task. Strategies for normalization of metabolic profile data can be divided into two major categories[92]:

- Statistical models used to derive optimal scaling factors for each sample based on complete dataset[107], such as normalization by unit norm[108] or median[109] of intensities, or the maximum likelihood method[110].
- Normalization by a single or multiple internal (i.e., added to sample prior to extraction) or external (i.e., added to sample after extraction) standard compounds based on empirical rules, such as specific regions of retention time[111], [112].

i.5. Feature annotation

a) LC-MS deconvolution:

Only a small fraction of the hundreds of metabolites that can be present in a sample can be annotated with an acceptable level of confidence[113]. Library searching of all statistically significant features without prior knowledge of monoisotopic accurate masses of the underlying metabolites might lead to missanotations if adducts or in-source fragments are present[114]. In addition, accurate mass library searches – considering expected adducts – can lead to a large number of potential molecular formulas and thus, molecular entities. Computational feature grouping and annotation is therefore a necessary step to reduce the list of putative identities. In this context, annotation is defined as the process of “noting” each observed feature with a putative identity. Annotation generally refers to assigning each feature with a putative metabolite name or molecular formula, but it also includes assigning each observed feature with the identity of formed adducts, neutral losses, isotopes and in-source fragments. This, ultimately facilitates the accurate characterization and identification of annotated adduct peaks via tandem mass spectrometry (MS/MS)[115].Existing computational annotation tools usually start from the input provided by the above peak picking tools, the output of which is a list of features: a peak, or a set of aligned peaks across samples with a unique m/z and a specific retention time (mzRT).

The two main grouping principles for detecting and annotating features related to a metabolite are chromatographic peak-shape similarity (i.e., co-eluting features) and peak-abundance correlation, or a combination thereof. Pairwise intensity correlation analysis across multiple samples is the basis of computational tools such as AStream[116], MSClust[117], RAMClust[118], MS-FLO[119], compMS2Miner[120] or findMAIN[121] among other similar approaches. On the other hand, peak shape similarity is used by CAMERA[122], MET-COFEA[123], ALLocator[123] or MZmine2[124]. MetAssign[125] or xMSannotator[126] have also included a probabilistic score to measure the confidence in particular assignments based on statistical clustering.

b) GC-MS deconvolution:

GC–EI MS metabolomics experiments produce large and complex datasets characterized by co-eluting compounds and extensive fragmentation of molecular ions caused by the hard electron ionization. GC-MS deconvolution aims to identify and extract quantitative information of metabolites across multiple biological samples. Computational approaches performing deconvolution are classified into two categories: tools based on peak-picking, and tools for compound extraction through the so-called curve resolution or spectral deconvolution.

The first category involves detecting all relevant fragment ion peaks in the spectra, and align them across multiple samples[127] to subsequently discover statistical peak variations between experimental groups. Representative tools from this category include MZmine[128], MetAlign[129] and XCMS[102-103]. Although these tools were initially intended for liquid chromatography mass spectrometry (LC–MS) data processing, they can also be used for GC–MS data analysis. The quantitative variables provided by these methods are not based on the compound spectra, but the m/z value, retention time window and area of fragment ion peaks. Thus, compound identification is the main bottleneck of peak-picking approaches. In this regard, tools such as metaMS[130], TagFinder[131], MetaboliteDetector[132] and PyMS[133] attempt to overcome this limitation by grouping the different peaks (based on their shape similarity or peak correlations) into partial compound spectra, allowing the putative identification of compounds by comparing their mass spectra with a reference MS library.

The second category focuses on the compound as the analysis entity, as opposed to the use of individual fragment ion peaks. Compounds are quantified and identified on the basis of a multivariate deconvolution process that extracts and constructs pure compound spectra from raw data. Representative tools falling into this category include TNO-DECO[134] or ADAP-GC[135]. TNO-DECO uses multivariate curve resolution to extract the compound spectra, whereas the deconvolution algorithm of ADAP-GC is based on an hierarchical clustering of fragment ions. Other free software, such as AMDIS[136] or BinBase[137] perform parts of the GC–MS metabolomics workflow. AMDIS is used to identify compounds by using the NIST library, but it processes samples independently and it does not include spectral alignment. BinBase uses the spectral deconvolution provided by a proprietary algorithm in the commercial software ChromaTOF (LECO Corporation) in order to align compounds across samples, and it provides compound quantification and identification based on selfconstructed libraries[138]. Finally, another free computational tool such as eRah[139] with an integrated design that not only incorporates a spectral deconvolution method using multivariate techniques based on blind source separation (BSS) but also alignment of spectra across samples, quantification, and automated identification of metabolites by spectral library matching.

i.6. Metabolite identification

Small molecules are less tractable to cataloging than are the objects of other ‘omics’. Unlike genes, transcripts and proteins, metabolites are not encoded in the genome. They are also chemically diverse, consisting of carbohydrates, amino acids, lipids, nucleotides and more. Researchers can make informed guesses and usually deduce compounds’ molecular formulas, but typically only about 25% of observed compounds can be tentatively identified [18].

Identification of metabolites is still evolving within the community, with active discussion on defining what constitutes a valid metabolite identification[140]. The Metabolomics Society, for instance, is currently assessing and developing an improved set of reporting standards[140], [141], [142]. The Chemical Analysis Working Group of the Metabolomics Standards Initiative (MSI; http:// www.metabolomics-msi.org/) established four levels of identification so far[143]. Level 1 identification requires at least two orthogonal molecular properties of the putative metabolite to be confirmed with an authentic pure compound analyzed under identical analytical conditions. By contrast, for levels 2 and 3, the comparison against literature and database data is sufficient, and therefore rather than identifications, only annotations are achieved. Level 4 identification refers to unknown compounds[144].

Proper usage and development of MS-based spectral databases, therefore, is essential for metabolomics to reach the status of other omic sciences[145]. Unfortunately, current databases are still far from containing experimental data of all known metabolites, despite attempts to increase and improve their content; one such example is the Metabolite Standards Synthesis Core (MSSC) initiative by the National Institutes of Health (NIH), aiming to generate new compound standards (http://www.metabolomicsworkbench.org/standards/nominatecompounds.php). The major limitation is the relatively small number of metabolites commercially available as pure standards, not to mention the large number of metabolites with unknown chemical structures that remain to be identified and characterized[146]. In addition, the transferability of mass spectral databases, particularly MS/MS, between MS instruments can impose some limitations, restricting the structural assignments of metabolites by empirically matching spectral values from pure standard compounds. Despite these limitations, the use of reference spectral databases is still one of the best approaches to annotate the structure of known metabolites when full isolation and structure determination by NMR or X-ray crystallography is not possible. Alternatively, novel computational tools that heuristically predict MS fragmentation patterns in silico have been developed to assist with identification of metabolites for which tandem MS data are not available yet in databases[147, 148, 148, [149, 150]. For electron ionization–mass spectrometry (EI–MS), it has been shown that fragmentation spectra can be simulated with quantum chemical and molecular dynamics methods[151], although the runtime is still too large (several thousand CPU hours per molecule) to simulate spectra for many compounds[144].

Freely accessible and/or commercially available compound databases currently used in the field of metabolomics provide information on chemical structures, physicochemical properties, spectral profiles, biological functions, and pathway mapping of metabolites. On the basis of these annotations, Fiehn et al.[152] classified these databases into two categories: (i) pathway-centric databases such as KEGG[153], Reactome[154], WikiPathways[155], or BioCyc[156] and (ii) compound-centric databases such as PubChem[157], ChemSpider[158], METLIN[159], MassBank[160], GMD[160], or Human Metabolome Database (HMDB)[161]. While PubChem, ChemSpider, and Chemical Abstracts Service (CAS) provide >60 million, > 35 million, and >100 million chemical compounds, respectively, they are not typically used in metabolomics because of the limited biological relevance of the vast majority of chemicals and the lack of mass spectral information. By contrast, some other compound-centric databases are also enriched with mass spectral information, which enables annotation of metabolites by matching mass spectral features of the unknown compounds to curated spectra of reference standards. Although these are much smaller repositories than

PubChem or ChemSpider, mass spectral databases represent the first step in converting raw spectral data into metabolite annotations and thus biological knowledge[144].

- NMR data processing

The resulting spectral data in NMR not only allows the quantification of the concentration of metabolites but also provides information about its chemical structure. The spectral peak areas generated by each molecule are used as an indirect measure of the quantity of the metabolite in the sample, while the pattern of spectral peaks informing on the physical properties of the molecule is used to identify the type of metabolite.

Types of processing:

- Spectral binning: also called the chemometric approach, the compounds are not initially identified – only their spectral patterns and intensities are recorded, statistically compared and used to identify the relevant spectral features that distinguish sample classes. Once these features have been identified, a variety of approaches may then be used to identify the corresponding metabolites[162], [162]. Specifically, the NMR spectrum of the biosample of interest is divided into smaller regions or bins so that specific features, peaks or peak clusters in a multi-peak spectrum can be mapped and compared across many different spectra (Figure 13). Once binned, the peak intensities (or total area under the curve) in each bin are tabulated and analyzed using multivariate statistical analysis[163].

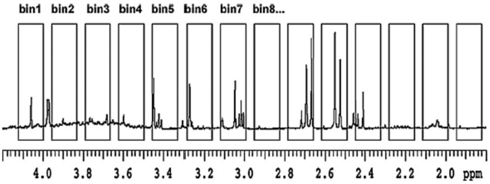


Figure 13. An example of how an NMR spectrum would be binned or partitioned prior to analysis by principal component analysis (PCA) or other statistical methods (adopted from [163]).

- Targeted profiling: the compounds are identified and quantified by comparing the NMR spectrum of the biosample of interest to a spectral reference library obtained from pure compounds[164], [165]. Once these compounds are identified and/or quantified, the data are processed to identify the most important biomarkers or informative pathways. The underlying assumption in quantitative metabolomics is that any given sample spectrum (i.e. a mixture of metabolites) is the sum of individual spectra from each of the pure metabolites in the mixture (Figure 14). For NMR, this particular approach often requires that sample pH and/or temperature be precisely known or precisely controlled. It also requires use of sophisticated curve-fitting software and specially prepared databases of NMR spectra of pure metabolites collected at different pH values and different spectrometer frequencies (say 100–800 MHz)[165]. One reason why quantitative metabolomics works so well for NMR is because most metabolites have unique or characteristic ‘‘chemical shift’’ fingerprints. This characteristic of NMR spectra helps reduce the problem of spectral redundancy, as it is unlikely that any two compounds will have identical numbers of peaks with identical chemical shifts, peak intensities, spin couplings or line shapes. With current technology, quantitative metabolomics is capable of identifying up to 100 compounds at a time in certain biological samples[163, 166].

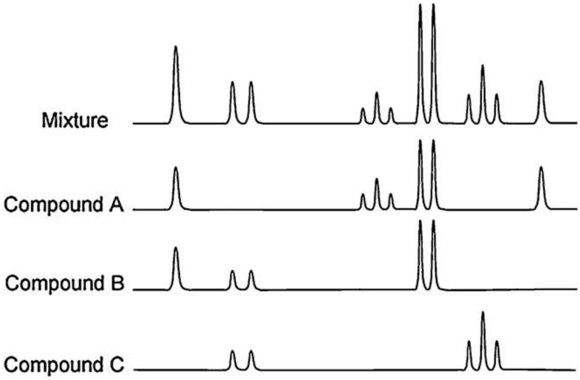
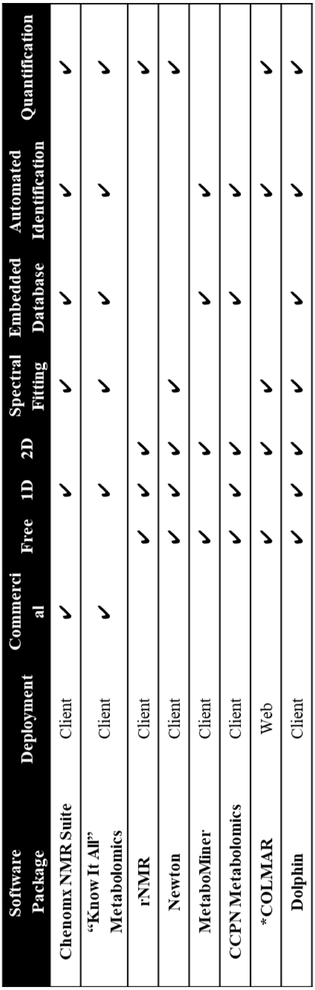


Figure 14. The central concept behind quantitative metabolomics by NMR. The NMR spectrum of a liquid mixture (top) is the sum of individual spectra for each of the pure components (Compounds A, B and C) in the mixture (adopted from [163]).

Table 3. NMR Data Processing Software for Spectral Identification & Quantification (adopted from [167])

*Specialized package for 2D covariance spectroscopy



Software Package	Deployment	Commercial	Free	1D	2D	Spectral Fitting	Embedded Database	Automated Identification	Quantification
Chenomx NMR Suite	Client	✓		✓		✓	✓	✓	✓
"Know It All" Metabolomics	Client	✓		✓		✓	✓	✓	✓
rNMR	Client		✓	✓	✓				✓
Newton	Client		✓	✓	✓	✓			✓
MetaboMiner	Client		✓		✓		✓	✓	
CCPN Metabolomics	Client		✓	✓	✓		✓	✓	
*COLMAR	Web		✓		✓	✓		✓	✓
Dolphin	Client		✓	✓	✓	✓	✓	✓	✓

1.2.3.1.5 Data analysis

There are two major approaches to analysing metabolomic data—chemometric approaches and quantitative approaches. With chemometric approaches the compounds are not initially identified—only their spectral patterns and intensities are recorded, statistically compared and used to identify the relevant spectral features that distinguish sample classes. Once these features have been identified, a variety of approaches can be used to identify the metabolites corresponding to the most important features. In contrast to chemometric approaches, quantitative metabolomics (or targeted profiling) aims to formally identify and quantify all detectable metabolites from the spectra, prior to subsequent data analysis[168]. The final goal of univariate and multivariate analysis techniques is to extract relevant information from the data with the aim of providing biological knowledge on the problem studied.

Among the wide range of statistical tests that can be applied to the metabolomic data the most frequent are the t test, analysis of variance, principal component analysis, and partial least squares discriminant analysis constitute[169].

The intrinsic nature of biological processes and metabolomic datasets is undoubtedly multivariate since it involves observation and analysis of more than one variable at a time. Consequently, the majority of metabolomics studies make use of multivariate models to report their main findings. Despite the conferred utility, powerfulness and versatility of multivariate models, their performance might be fraught by the high-dimensionality of such datasets due to the so-called 'curse of dimensionality' problem. Curse of dimensionality arises when datasets contain too much sparse data in terms of the number of input variables. This causes, in a given sample size, a maximum number of variables above which the performance of our multivariate model will degrade rather than improve. Hence, attempting to make the model conform too closely to this data (i.e., considering too many variables in our multivariate model) can introduce substantial errors and reduce its predictive power (i.e., overfitting). Therefore, using multivariate models require intensive validation work. The implementation of multivariate and univariate data analysis is not mutually exclusive and in fact, their combined use is strongly recommended to maximize the extraction of relevant information from metabolomic datasets[170].

- **Univariate methods**

Univariate methods are mainly focused on analysing one variable at a time (in omics disciplines usually one out of a panel of many measured). Tests such as t test or ANOVA[169] are the most frequently used to compare different sets of samples. The application of univariate methods in an omics context usually results in the need of significance testing for tenths to hundreds of metabolites; this makes correcting for multiple tests necessary to protect against the increasing probability of having false positives. Correcting has repercussions on the results of univariate statistical analyses since multiple testing correction in theory increases the probability of obtaining false negatives[169].

The ultimate goal of data analysis is to obtain a list of features showing both statistically significant changes and a minimum fold change (FC) ratio between experimental conditions for further MS/MS identification. Although it is not a unique way to perform data analysis there are some guidelines that can be followed to achieve the goal mentioned above[170]: Features that do not contain biological information should be removed using quality control (QC). The quality control (QC) sample should qualitatively and quantitatively represent the entire collection of samples included in the study, providing an average of all of the metabolomes analysed in the study. The QC samples are analysed intermittently for the duration of the analytical study to assess the variance observed in the data throughout the sample preparation, data acquisition and data pre-processing steps. Replicate injections should provide identical data for each injection, however in reality analytical variance will be observed and the replicate QC injections can be used to measure this variance across the analytical study. Then, compute the coefficient of variation (CV) of QC and Samples and proceed to retain only those metabolic features presenting CVQC < CVSamples. If QC samples are not available, an alternative procedure is to retain those features with CVSamples > 20%.The experimental designed must be aligned with the statistical test that will be applied, for example, the data can be paired or not. Afterwards, the final aim is that data can show normal distribution and equality of variance to make it easier finding biological differences.

Parametric test (means) will be applied to datasets that present normal distribution otherwise; non-parametric test (medians) will be performed.

Decide a false discovery rate (FDR) threshold to accept, although a general consensus is to accept 5% of FDR level but there is nothing special about this value and each researcher might justify their assumed FDR value. Plotting histograms of p-values frequency distribution to get an overview of the significant differences could help to stablish a FDR value.

Fix a cutoff FC value. It is recommendable an arbitrary 1.5-FC cutoff value meaning a minimum of 50% of variation in the two groups compared.Significant features with with higher FC values will be retained for MS/MS experiments and follow-up validation studies.

- **Multivariate methods**

On the other hand, multivariate statistics evolve the statistical analysis of more than one statistical variable simultaneously. Among different multivariate analyses, those based on projection methods represent the most efficient and useful methods for the analysis and modelling of complex data. Principal Component Analysis (PCA) is the workhorse in chemometrics. PCA allows extracting and displaying the systematic variation in the data. A PCA model provides a summary, or overview, of all observations or samples in the data table. In addition, groupings, trends, and outliers can also be found. Hence, projection based methods represent a solid basis for metabolomic analysis[171, 172, 173].

Projection methods convert multidimensional data into a low-dimensional model plane that approximates all rows (e.g., objects or observations) in X, that is, the swarm of points. The first PCA model component describes the largest variation in the swarm of points. The second component models the second largest variation and so on. All PCA components are mutually linearly orthogonal to each other (see Figure 15). The scores (T) represent a low-dimensional plane that closely approximates X, that is, the swarm of points. A scatter plot of the first two score vectors provides a summary, or overview, of all observations or samples in the data table. Groupings, trends, and outliers are revealed. The position of each object in the model plane is used to relate objects to each other. Hence, objects that are close to each other have a similar multivariate profile, given the K-descriptors. Conversely, objects that lie far from each other have dissimilar properties. Analogous to the scores, the loading vectors (p1, p2) define the relation among the measured variables, that is, the columns in the X matrix. A scatter plot, also known as the loading plot shows the influence (weight) of the individual X-variables in the model. An important feature is that directions in the score plot correspond to directions in the loading plot, for example, for identifying which variables (loadings) separate different groups of objects (the scores). This is a powerful tool for understanding the underlying patterns in the data. Hence, projection-based methods represent a solid basis for metabolomic analysis[171].

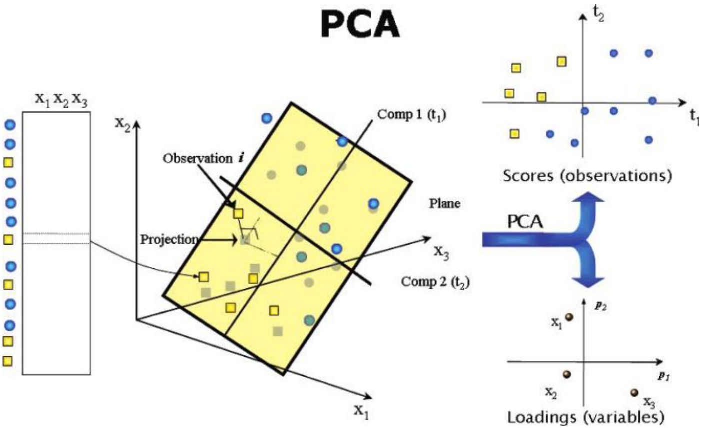


Figure 15. A principal component analysis (PCA) model approximates the variation in a data table by a low dimensional model plane. This model plane provides a score plot, where the relation among the observations or samples in the model plane is visualized, for example, if there are any groupings, trends, or outliers. The loading plot describes the influence of the variables in the model plane, and the relation among them. An important

feature is that directions in the score plot correspond to directions in the loading plot, and vice versa (adopted from [171]).

1.2.3.1.6 Biological interpretation

Biological interpretation is the final goal of metabolomics: converting a long list of metabolites showing up in a given experiment into a reduced set of meaningful biological terms, such as the pathways/biological processes enriched in them[174]. Metabolomics datasets are considered highly complex when used to relate metabolite levels to metabolic pathway activity. Despite recent developments in bioinformatics, which have improved the quality of metabolomics data, there is still no straightforward method capable of correlating metabolite level to the activity of different metabolic pathways operating within the cells. Thus, this kind of analysis still depends on extremely laborious and time-consuming processes[175].

Recently, several software tools have become available for the functional and biological interpretation of metabolomic experiments (Figure 16). They can be classified in two groups that allow complementary analysis. The first comprises tools that allow mappings and visualizations of a set of metabolites in graph representations of the metabolism (mainly metabolic pathways). The second group comprises tools for the statistical analysis of metabolite annotations, commonly known as enrichment analysis[176].

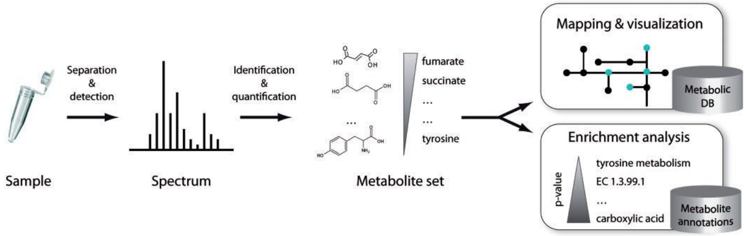
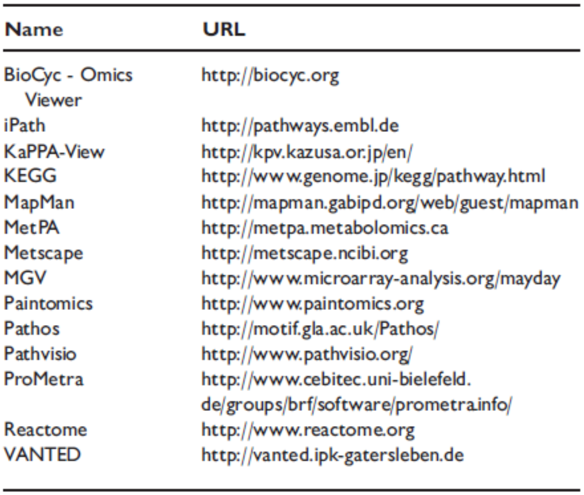


Figure 16. Metabolomics analysis workflow. Once a set of metabolites of interest have been identified, two types of tools can be used to gain biological insight into experimental results: (i) mapping and visualization of pathways and (ii) statistical enrichment analysis of metabolite annotations (adopted [176]).

- **Pathway mapping and visualization**

An obvious first step in the interpretation of metabolic experiments is to map and visualize the identified metabolites and associated experimental measurements in the context of metabolic pathways and other general biological networks. Such visualization can provide a quick overview on the metabolic context of the metabolites showing up in the experiment. For example, it makes possible to assess whether the set of identified metabolites are involved in the same biological pathway or if they are close to each other in the metabolic network. Although locating and visualizing a number of compounds in metabolic charts could look trivial at first sight, automatic and interactive tools are required due to the complexity of the metabolism and its associations with other biological phenomena. Several software applications provide this functionality (Table 4)[176].

Table 4. Pathway mapping and visualization software (adopoted from [176]).



Name	URL
BioCyc - Omics Viewer	http://biocyc.org
iPath	http://pathways.embl.de
KaPPA-View	http://kpv.kazusa.or.jp/en/
KEGG	http://www.genome.jp/kegg/pathway.html
MapMan	http://mapman.gabipd.org/web/guest/mapman
MetPA	http://metpa.metabolomics.ca
Metscape	http://metscape.ncibi.org
MGV	http://www.microarray-analysis.org/mayday
Paintomics	http://www.paintomics.org
Pathos	http://motif.gla.ac.uk/Pathos/
Pathvisio	http://www.pathvisio.org/
ProMetra	http://www.cebitec.uni-bielefeld.de/groups/brf/software/prometrainfo/
Reactome	http://www.reactome.org
VANTED	http://vanted.ipk-gatersleben.de

Among these tools, the Kyoto Encyclopedia of Gene and Genomes (KEGG) and Reactome are two of the most popular databases and are freely available:

KEGG's pathway browser includes a functionality to locate and color different entities, including metabolites. Both basic and more advance coloring functionalities are available. A list of pathways and the number of entities found in each pathway is provided, enabling the user to visualize each pathway one by one. A global view on the metabolism can also be mapped using the KEGG Atlas[177]. This functionality can also be used programmatically through KEGG's programmatic interface (Figure 17)[176], [177].

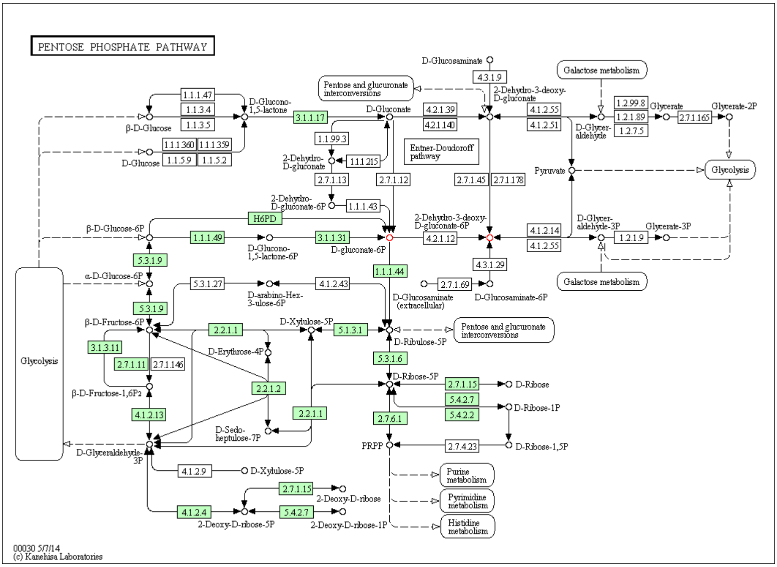


Figure 17. Kegg pentose phosphate pathway map showing dysregulated enzymes and metabolites.

Reactome is a pathway database managed by the European Bioinformatics Institute (EBI). Built around human pathway data, pathways for 20 other species are inferred by orthology. Mapping and visualization of metabolites can be performed using the 'Map IDs to Pathways' facility. In addition, Reactome calculates overrepresentation statistics of pathway annotations[176, 178].

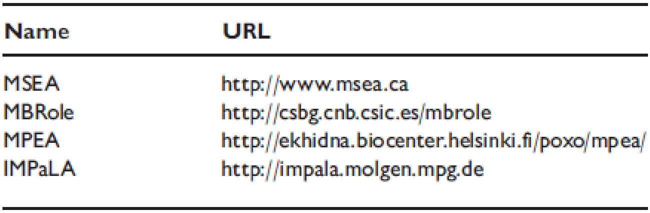
- **Enrichment analysis**

Since the first appearance of enrichment analysis methods for the functional interpretation of large gene lists in 2002, numerous tools for applying this type of analysis to transcriptomic and proteomic data are available[179]. These tools are frequently used for calculating and reporting statistical values associated to gene functional annotations provided by several data sources. In general these tools compare the annotations present in a set of genes of interest (i.e. those up or downregulated in a transcriptomics analysis) with the corresponding annotations in a reference set (i.e. all the genes in the organism or in the array) and report those annotations overrepresented in the interest set, according with a given statistical test. Consequently, these tools are very useful to 'summarize' a long list of genes and to have an idea of the biological phenomenon behind it[176].

Table 5 shows a number of publicly available software implementing similar approaches for the analysis of a list of metabolites. These tools, specifically designed for metabolomics, allow the functional interpretation of metabolomic experiments in terms of statistically significant general pathways as well as other biological annotations[176].

The tools for enrichment analysis reviewed perform two variations of this kind of analysis: overrepresentation analysis (ORA) and set enrichment analysis (SEA). For both approaches the user should provide as input a set of metabolites and select the type of annotations to examine. In the case of SEA, a numeric value associated to each metabolite (e.g. its concentration) has to be also provided. The general output is a list of annotations and their associated P-value in a tabular format.

Table 5. Metabolite enrichment analysis software (adopted from [176]).



Name	URL
MSEA	http://www.msea.ca
MBRole	http://csbg.cnb.csic.es/mbrole
MPEA	http://ekhidna.biocenter.helsinki.fi/poxo/mpea/
IMPaLA	http://impala.molgen.mpg.de

1.3 Proteomics

Collectively, proteins catalyse and control essentially all cellular processes. They form a highly structured entity known as the proteome, the constituent proteins of which carry out their functions at specific times and locations in the cell, in physical or functional association with other proteins or biomolecules. The extensive proteome network of the cell adapts dynamically to external or internal (that is, genetic) perturbations and thereby defines the cell’s functional state and determines its phenotypes. Describing and understanding the complete and quantitative proteome as well as its structure, function and dynamics is a central and fundamental challenge of biology[180].

1.3.1 Shotgun quantitative proteomics workflow

In all of the techniques, proteins are extracted from the source material then digested into peptides by a sequence specific enzyme such as trypsin. The resulting mixture of peptides is separated by reverse-phase chromatography, which is coupled online to electrospray ionization (Figure 18). The peptide ions are then transferred to the vacuum of a mass spectrometer, where they are fragmented in the gas phase to generate MS/MS (MS2) spectra that contain the information to identify and quantify specific peptides. The resulting data are analysed by mass-spectrometry-specific computational pipelines as well as general downstream systems-biology solutions that are tailored to proteomics[180, 181].

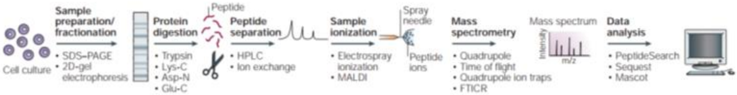


Figure 18. The mass-spectrometry/proteomic experiment. A protein population is prepared from a biological source — for example, a cell culture — and the last step in protein purification is often SDS–PAGE. The gel lane that is obtained is cut into several slices, which are then in-gel

digested. Numerous different enzymes and/or chemicals are available for this step. The generated peptide mixture is separated on- or off-line using single or multiple dimensions of peptide separation. Peptides are then ionized by electrospray ionization (depicted) or matrix-assisted laser desorption/ionization (MALDI) and can be analysed by various different mass spectrometers. Finally, the peptide-sequencing data that are obtained from the mass spectra are searched against protein databases using one of a number of database-searching programmes. Examples of the reagents or techniques that can be used at each step of this type of experiment are shown beneath each arrow. 2D, two-dimensional; FTICR, Fourier-transform ion cyclotron resonance; HPLC, high-performance liquid chromatography (adopted from [182]).

Workflow steps:

1.3.1.1 Proteins extraction

Proteins are part of a complex network of interacting biomolecules that regulate their function and localization within the cell. Extraction and isolation of proteins from chemical and physical interactions with other biomolecules from specific cellular subcompartments have become a critical step for their global analysis in a biological context. Application of traditional subcellular isolation techniques, primarily sucrose gradient sedimentation and similar methodologies, from different cell types and tissues have allowed for global analysis of proteins within subcellular compartments[183]. Detergents and amphipathic molecules disrupt hydrophobic interactions, also enabling protein extraction and solubilization. With respect to the ionic character of the hydrophilic group, they are classified into several groups: ionic (e.g. anionic sodium dodecyl sulfate (SDS)), non-ionic (uncharged, e.g. octyl glucoside, dodecyl maltoside and Triton X-100) or zwitterionic (having both positively and negatively charged groups with a net charge of zero, e.g. CHAPS, 3-[(3-Cholamidopropyl)dimethylammonio]-2-hydroxy-1-propanesulfonate (CHAPSO), tetradecanoylamidopropyl-dimethylammoniobutanesulfonate (ASB-14)). Applicable concentrations of detergents range from 1 to 4%, and the exact content of solubilization solution needs to be verified in accordance to the method of choice for protein separation (some reagents may interfere with subsequent steps)[184].

1.3.1.2 Protein depletion and denaturalization

Protein dynamic range is the largest challenge that faces proteomics technology development. Currently, all steps within an LC-MS proteomics pipeline are protein abundance-dependent. Thus, adjustment of protein concentration dynamic range has become an option for improving comprehensiveness through improved analysis of low abundance proteins[183].

Fortunately, only a few proteins are extremely abundant, such as serum albumin in the case of plasma, and thus can be specifically removed or depleted prior to LC-MS analysis. Chemical-based approaches can selectively precipitate abundant proteins, usually albumin, from plasma to improve proteomic depth and have been demonstrated with sodium chloride and ethanol[185], acetonitrile[186], the disulfide reducing agents DTT and TCEP[187] and ammonium sulfate[188]. Antibody arrays against the highest abundance proteins have also improved proteomic coverage of clinical samples[189, 183].

1.3.1.3 Proteolytic Digestion

Analysis of proteins from their proteolytic peptides circumvents some of the challenges associated with intact protein separation, ionization, and MS characterization. A protein lysate is a highly heterogeneous mixture of proteins with diverse physicochemical properties. Purposefully increasing the complexity of a sample prior to analysis is somewhat counterintuitive. However, selective protease digestion acts to normalize and compartmentalize the biochemical heterogeneity of proteins within a sample as peptides and may, in fact, create a less heterogeneous mixture when protein splice isoforms and post-translational modifications are considered. Additionally, with multiple representations of a protein as peptides the probability of sampling and identifying a peptide associated with a particular low abundance protein and/or post-translational modification increases[183].

In general, proteolytic enzymes differ by their specificity for cleaving the amide bonds between individual residues in a protein. The cleavage is carried out through hydrolysis of the amide bond before or after a specific residue, residues, or combination of residues. Trypsin has become the gold standard for protein digestion to peptides for shotgun proteomics. Trypsin is a serine protease which cleaves at the carboxyl side of arginine and lysine (Figure 19). This sequence-specific information has been used to filter identified peptides[183].

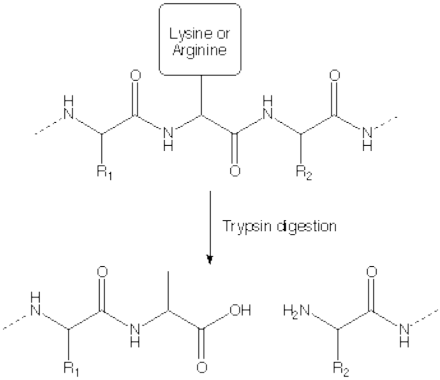


Figure 19. Trypsin digestion

1.3.1.4 Data acquisition by LC-MS

LC-MS systems consists of: 1) a chromatography column, which separates peptide mixtures based on one or more physicochemical properties prior to MS; 2) an ionization source, which converts eluting peptides into gas phase ions; 3) one or more mass analyzers, which separate ions on the basis of m/z ratios; and 4) a detector, which registers the relative abundance of ions at discrete m/z. In MS/MS, precursor ions are recorded in full-scan mode (all m/z values), followed by selective ion isolation and fragmentation for sequence identification. MS/MS instruments are operated in an automated alternating scan mode. Two main ionization technologies used are ESI and MALDI. Because ESI generates ions directly from solution, it is readily coupled to LC or capillary electrophoresis. In a standard reverse-phase HPLC setup, the column media differentially retards the migration of peptides based on selective hydrophobic interaction affinities[190].

Peptides are then eluted with a gradient of organic solvent and ionized just prior to introduction into the mass spectrometer. LC is well-suited to examining complex biological samples because: 1) peptides with the same nominal m/z are less likely to be introduced at the same time, reducing ambiguity; and 2) with fewer competing ion species, fewer artifacts arise due to ion suppression or ion-ion interference[190, 183]. Trap columns are usually used before the analytical column with the aim of enabling injections of high volumes of highly diluted samples in which the sample is concentrated and desalted[191]. Typical microcolumns for nanoLC are prepared using reversed phase materials with a 3–10 µm diameter packed into fused silica capillaries with a 12–100 µm diameter, in which sintered silica particles or silicate-polymerized ceramics have been used as frits[192, 193, 194]. The post-column connections affect peak broadening, and usually zero dead-volume unions are used, in which tubing ends are closely adjoined to one another. In this case, how the tubing is cut is critical to avoid peak broadening. In LCMS for peptides, acidic conditions with ion-pair reagents are usually used in combination with C18 stationary phases to suppress peak broadening. Trifluoroacetic acid (TFA) is one of the most popular reagents because of higher peak capacity with smaller peak width[195].

Samples for proteomic analysis are available in high amounts; however, the analytes are present in minute concentrations. Therefore, pumping systems were developed for providing flow rates in the nl/min range[196]. The quality of profiling studies is determined by the overall sensitivity, detection coverage, dynamic range, fragmentation efficiency, mass resolution, and accuracy[197]. Femptomole or better detection limits are commonly attained with LC-MS, even with mixtures exceeding several thousand components [198, 190].

The main platforms for quantitative proteomics today are orbital traps, QTOF instruments, and triple-quadrupole instruments mostly using ESI as an interface for chromatographic systems and collision-induced dissociation as the peptide fragmentation technique (with electron transfer dissociation playing a useful complementary role for large, posttranslationally modified or cross-linked peptides[199]). Together, these instruments enable the characterization of proteomes to a depth of 5,000-10,000 proteins[200]. Today, the two main approaches for the quantification of peptides and proteins from LC-MS data are (1) extracting the LC-MS intensities of peptide precursor ions in the classical data-dependent acquisition (DDA) experiment in which peptide precursors are subjected to fragmentation (sequencing) as they are eluted from the LC system and (2) extracting the LC-MS/MS intensities of peptide fragment ions (or iTRAQ/TMT reporter ions) of peptides from DDA experiments[201].

In DDA-based methods, mass spectra of all the ion species that co-elute at a specific point in the gradient elution (that is, precursor-ion spectra) are recorded at the MS1 (or full-scan) level. The instrument alternates between the acquisition of full-scan data and the acquisition of fragment-ion spectra, in which as many precursors as possible are sequentially isolated and fragmented (at the MS2 level). Of many possible instrument configurations, quadrupole–orbitrap analysers[202] dominate DDA proteomics but time-of-flight instruments also have unique promise. In typical 'top N' cycles (in which 'N' denotes the number of MS2 spectra that follow), an MS1 scan is followed by about ten fragment-ion scans[180].

1.3.1.5 Data processing: protein identification and quantification

- **Protein identification**

MS spectra contain peptide mass and intensity information, and the identity of the peptides is deduced by matching the MS/MS spectra against a sequence database. Typically, peaks are extracted from raw data, the peptide mass is estimated from the scan from which the peak was 'picked' for sequencing and the peak files are sent to a search engine. Results consist of tables of identified proteins.

The difficulty of assembling peptide identifications back to the protein level results from the same factors that made the shotgun proteomic approach so successful in the first place. Protein digestion makes peptides, and not the proteins, the currency of the method, and the connectivity between peptides and proteins is lost at the digestion stage. This loss of connectivity complicates computational analysis and biological interpretation of the data especially in the case of higher eukaryote organisms. The same peptide sequence can be present in multiple different proteins. Therefore, the identification of such shared peptides can lead to ambiguities in the determination of the identities of the sample proteins.

In general, the process of peptide assembly consists of the following steps. First, peptide assignments obtained by searching acquired MS/MS spectra against a protein sequence database using algorithms such as SEQUEST[203] or Mascot[204] are filtered using a user-specified set of criteria to remove false identifications. Second, accession numbers and annotations of protein sequence database entries corresponding to each peptide are retrieved from the sequence database. Third, peptides are grouped by their corresponding sequence database entries. Fourth, shared peptides are apportioned among all corresponding proteins, and a summary protein list is created. Ideally the apportionment of peptides to proteins should be done using a probability-based approach, i.e. taking into account the probabilities of peptide assignments[205]. This has an advantage in that it allows calculation of statistical confidence measures for protein identifications and estimation of false identification error rates resulting from filtering the data[206, 207].

When peptides are assigned to MS/MS spectra using the database search approach, the universe of all potential peptide assignments is limited to the sequences present in the searched protein sequence database. The completeness of the sequence database thus can be a decisive factor in experiments where identification of sequence polymorphisms is crucial for the biological interpretation of the data. Some of the existing protein sequence databases that are commonly used are Uniprot[208], RefSeq[209], Ensembl[210] or Entrez[211]. The choice of a particular database should be based on the goals of the experiment.

- **Protein quantification**

In addition to providing a profile of what proteins are present within a system at a given time, information on the expression levels of these proteins is required. Two approaches to quantify proteins are currently used, namely, the label free quantification and quantification based on the preliminary labelling of the protein or the corresponding peptides.

In protein-labeling approaches, different protein samples are combined together once labeling is finished and the pooled mixtures are then taken through the sample preparation step before being analyzed by a single LC-MS/MS or LC/LCMS/ MS experiment (Figure 20(a)). In contrast, with label-free quantitative methods, each sample is separately prepared, then subjected to individual LC-MS/MS runs (Figure 20(b)). Protein quantification is generally based on two categories of measurements. In the first are the measurements of ion intensity changes such as peptide peak areas or peak heights in chromatography. The second is based on the spectral counting of identified proteins after MS/MS analysis. Peptide peak intensity or spectral count is measured for individual LC-MS/MS runs and changes in protein abundance are calculated via a direct comparison between different analyses[212].

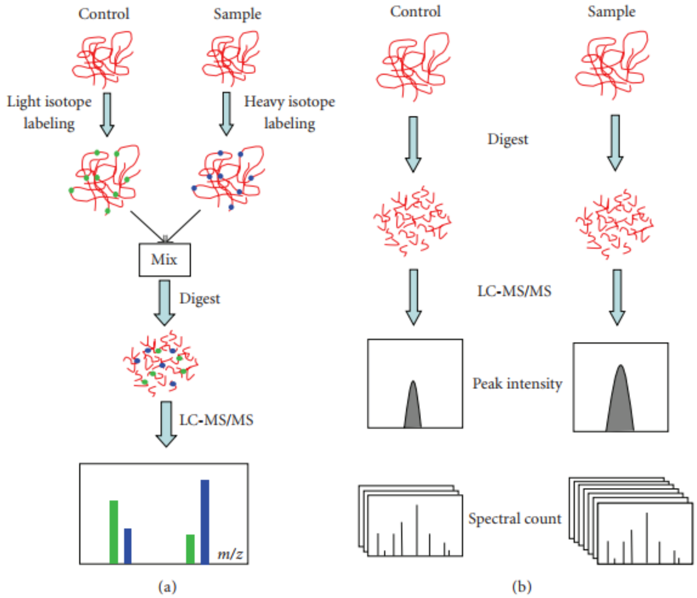


Figure 20. General approaches of quantitative proteomics. (a) Shotgun isotope labeling method. After labeling by light and heavy stable isotope, the control and sample are combined and analyzed by LC-MS/MS. The quantification is calculated based on the intensity ratio of isotope-labeled peptide pairs. (b) Label-free quantitative proteomics. Control and sample are subject to individual LC-MS/MS analysis. Quantification is based on the comparison of peak intensity of the same peptide or the spectral count of the same protein (adopted from [212]).

A number of labeling approaches can be incorporated into ‘shotgun’ type experiments due to the chemical and physical properties of isotope labeled compounds that are identical to properties of their natural counterparts except in mass, isotope labeled molecules were incorporated into mass spectrometry-based proteomics methods as internal standards or relative references. These include stable isotope labeling by amino acids in cell culture (SILAC)[213], isotope dilution[214], radiolabeled amino acid incorporation[215], chemically synthesized peptide standards[216], tandem mass tags (TMT)[217] and more recently, isobaric tags for relative and absolute quantification (iTRAQ)[218, 219]. The iTRAQ system is now commercially available with eight isobaric tags, having only initially been available with four tags, and has been widely used in proteomic studies. Most label-based quantification approaches have potential limitations: complex sample preparation, the requirement for increased sample concentration, and incomplete labeling. In order to address some of these issues the implementation of nonlabeled quantification is growing considerably[220].

The label free quantification is largely employed for its rapidity, its low cost and its simplicity of use. Currently, two widely used but fundamentally different label-free quantification strategies can be distinguished: (a) measuring and comparing the mass spectrometric signal intensity of peptide precursor ions belonging to a particular protein and (b) counting and comparing the number of fragment spectra identifying peptides of a given protein.

The first method, based on the comparison of mass spectra, the change in the signal intensity of a peptide (area or peak intensity) is correlated to the protein quantity. The ion chromatograms for every peptide are extracted from an LC-MS/MS run and their mass spectrometric peak areas are integrated over the chromatographic time scale. For low-resolution mass spectra this is typically done by creating extracted ion chromatograms (XICs) for the mass to charge ratios determined for each peptide. More recently, this concept has been extended to high-resolution data to include contributions of 13C isotopes to the overall signal intensities. The intensity value for each peptide in one experiment can then be compared to the respective signals in one or more other experiments to yield relative quantitative information. For proteomic analysis of very complex peptide mixtures, three important experimental parameters affect the analytical accuracy of quantification by ion intensities. (i) It is advantageous to employ a high mass accuracy mass spectrometer because the influence of interfering signals of similar but distinct mass can be minimized. (ii) The peptide chromatographic profile should be optimized for reproducibility to ease finding corresponding peptides between different experiments. This is not a trivial task and special software has been developed to align LC-runs prior to identifying corresponding peptides. (iii) The right balance between acquisition of survey and fragment spectra has to be found. While extensive peptide sequencing by tandem MS is required to identify as many proteins as possible in complex mixtures, a robust quantitative reading by ion intensities requires multiple sampling of the chromatographic peak by survey mass spectra. Typically, multiple fragment spectra are acquired for every survey spectrum at acquisition rates ranging from 5 spectrum/s (ion traps) to 0.2-1 spectrum/s (quadrupole-TOF instruments). 0.2 s/spectrum (ion traps) to 1–3 s/spectrum (quadrupole-TOF instruments). Given that chromatographic peak widths are in the order of 10–30 s for nano-LC separations, ion traps have an inherent advantage over QTOFs because many more MS to MS/MS cycles can be performed within the available chromatographic time. Still, even for fast sampling instruments, better quantification accuracy will inevitably mean poorer proteome coverage and vice versa[221, 222, 223].

In the second method, based on spectral counting, the number of peptides sequenced for a given protein is correlated to its quantity. The empirical observation that the more of a particular protein is present in a sample, the more tandem MS spectra are collected for peptides of that protein. Hence, relative quantification can be achieved by comparing the number of such spectra between a set of experiments. In contrast to quantification by peptide ion intensities, spectral counting benefits from extensive MS/MS data acquisition across the chromatographic time scale both for protein identification as well as protein quantification. However, the commonly employed dynamic exclusion of ions that have already been selected for fragmentation is detrimental for accurate quantification [221, 222, 223].

1.3.1.6 Data analysis and biological interpretation

Proteomics generates complex datasets prompting the use of computational tools and network analysis in order to understand biological systems as a whole. Yet, most of the published studies on differential protein expression have focused on applying a significance test for each single protein testing whether this protein is differentially expressed or not (see Figure 21)[224, 225].

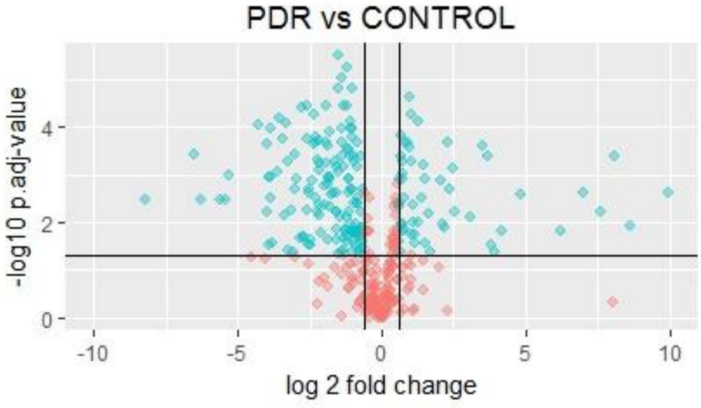


Figure 21. The volcano plot showing the estimated fold changes (x-axis) versus the log10 p-values (y-axis) for each protein.

However, studying each protein individually does not reflect the reality occurring within cells since proteins rarely act alone to perform their functions. It has been widely observed that proteins involved in the same cellular processes interact with each other since they belong to the same protein complex[226, 227].

For overcoming such important issue, the application of systems biology to proteomics allows reaching the complexity found in biological networks by taking a holistic view of the cell[2, 228]. In this context, protein-protein interaction (PPI) network where nodes represent proteins and edges physical interactions between two proteins have been widely used in systems biology. Since the organization of a network in a complex system still remains unclear, an effective approach to achieve this goal is detecting topological clusters or modules that can be involved in particular cellular functions or disease (see Figure 22)[229, 230].

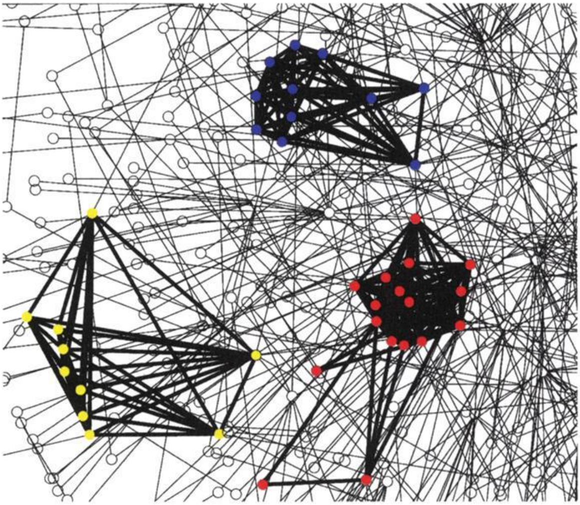


Figure 22. Fragment of the protein network. Nodes and interactions in discovered clusters are shown in bold. Nodes are colored by functional categories: red, transcription regulation; blue, cell-cycle/cell-fate control; and yellow, protein transport (adapted from [227]).

In order to extract underlying biological information from a PPI network, it is necessary to analyse it using graph-theoretic tools. Two different and complementary approaches, named topological and modular, have been developed for the study of a complex network. The topological approach studies the network as a whole by means of the analysis of the structural parameters of the graph. Instead, the modular approach divides the PPI network into modules that group nodes based on a common characteristic such as sharing the same function or belonging to the same pathway. Afterwards, each module is studied separately[231].

1.4 Diabetic retinopathy

Diabetes mellitus (DM) is a complex metabolic disorder that is associated with insulin resistance (IR), impaired insulin signalling, β-cell dysfunction, abnormal glucose levels, altered lipid metabolism, sub-clinical inflammation and increased oxidative stress. DM has become a major global health problem with almost 415 million people affected in 2015, and a projected figure of 642 million by 2040 (International Diabetes Federation, 2015)[232].

Type 2 diabetes mellitus (T2DM) leads to long-term pathogenic diseases such as cerebral vascular disease, diabetic coronary artery disease (CAD), diabetic peripheral neuropathy (DPN), diabetic retinopathy (DR), diabetic nephropathy (DN), lower extremity vascular disease and diabetic foot disease. There is no cure once the disease is diagnosed, but early treatment at the sub-clinical stage can prevent or at least halt disease progression. Thus, it is of critical importance to discover early biomarkers associated with disease progression[233].

Diabetic retinopathy (DR) is the most common complication of diabetes and the leading cause of blindness among working-aged adults around the world [234]. The global prevalence of DR among patients is approximately 35% and around one-tenth of them have vision-threatening states such as diabetic macular oedema (DMO) or proliferative diabetic retinopathy (PDR) [235]. Neovascularization due to severe hypoxia is the hallmark of PDR whereas vascular leakage due to the breakdown of the blood retinal barrier (BRB) is the main event involved in the pathogenesis of DMO [236,237]. Most of the research on the pathogenesis of DR has been focused on the impairment of the neuroretina and the breakdown of the inner BRB. However, the effects of diabetes on the retinal pigment epithelium (RPE) have received less attention.

RPE is a monolayer of pigmented cells situated between the neuroretina and choroids. RPE constitutes the outer BRB and is essential for neuroretina survival, and consequently, for visual function [238]. The specific functions of RPE are the following: i) transport of nutrients, ions, and water; ii) absorption of light and protection against photo-oxidation; iii) re-isomerization of all-trans-retinal into 11-cis-retinal, which is a key element of the visual cycle; iv) phagocytosis of shed photoreceptor membranes; and v) secretion of various essential factors for the structural integrity of the retina [238]. Therefore, the study of RPE is fundamental to gain new insights into the mechanisms that lead to DR and to identify new therapeutic targets for this devastating complication of diabetes.

1.4.1 DR stages (NPDR, PDR)

Diabetes mainly affects many cell types in the retina and thus induces vascular lesions which allows the classification and grading of retinopathy[239]. The fact that retinopathy is 'geographically disperse' across an individual retina, and between the two eyes of an individual person, makes difficult to obtain quantitative data for analysis in clinical research. Thus, it is necessary to develop functional as well as structural and biochemical biomarkers for a better characterization of diabetic retinopathy[239].

Diabetic retinopathy may be very broadly classified into two stages based on the level of microvascular degeneration and related ischemic damage: non-proliferative diabetic retinopathy (NPDR) (Figure 23) and advanced, proliferative diabetic retinopathy (PDR) (Figure 24) [239].

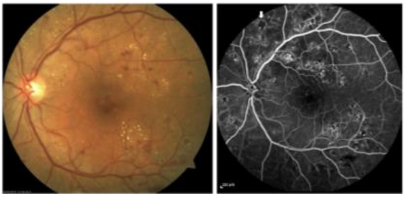


Figure 23. NPDR lesions in the diabetic retina. Colour fundus photographs (left) and fundus fluorescein angiogram (right) obtained from the left eye of a patient with nonproliferative diabetic retinopathy. Retinal haemorrhages, microaneurysms and hard exudation are seen (left). Microaneurysms are

more apparent on fluorescein angiography where areas of ischaemia are also detected (white arrow) (adopted from [239]).

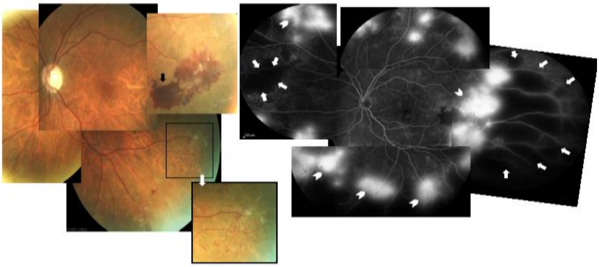


Figure 24. Clinical profile of PDR. Colour fundus photographs (left) and fundus fluorescein angiogram (right) obtained from the left eye of a patient with PDR. A large pre-retinal haemorrhage (left, black arrow) and areas of retinal neovessels (left, detailed magnified) were detected on fundus examination. Areas of retinal ischaemia (white arrows, left) and leak from retinal neovessels (white arrow heads) were seen on fluorescein angiography (right) (adopted from [239]).

NPDR can be sub-classified into i) mild NPDR (presence of microaneurysms in the retina); ii) moderate NPDR (more than mild but less than severe NPDR; and iii) severe NPDR (> 20 intraretinal hemorrhages in each of the four quadrants, venous bleeding in at least two quadrants, and intraretinal microvascular abnormalities (IRMA) in at least one quadrant in the absence of PDR)[240]. The progression of diabetic retinopathy is related to abnormalities of the vasculature including permeability of the blood retina barrier (BRB), progressive microvascular damage with vascular endothelial cell and pericyte loss, subsequent occlusion of capillaries, thickening of vascular basement membrane (BM) (Figure 25), and excessive retinal neuronal and glial abnormalities.

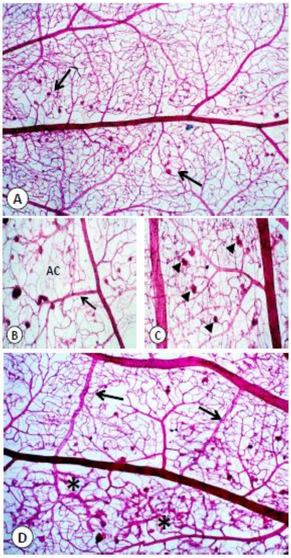


Figure 25. Vascular histology of human NPDR. Trypsin digest specimen of postmortem non-proliferative diabetic retinopathy from a type-2 diabetic patient. PAShaematoxylin staining of the retinal vasculature shows a retinal artery flanked above and below by capillary beds. Numerous microaneurysms occur a peri-arterial distribution pattern (arrows) (A). Figure 3B: Higher magnification trypsin digests show a precapillary arteriole (arrow) in which smooth muscle covering is lost and downstream of this vessel there is a number of microaneurysms and confluent areas of (non-perfused) acellular capillaries (AC). Figure 3C: Trypsin digest showing precapillary arterioles with arteriolar pathology (smooth muscle

loss) and many microaneurysms (arrowheads). Figure 3D: The central retina from a T2D patient showing IRMA (*) between pre-capillary arterioles and adjacent post-capillary venules. The IRMA occur in an area of acellular capillaries and are drained by venules (arrows) (adopted from [239]).

Screening for diabetic retinopathy

Diabetic retinopathy may develop and progress to advanced stages without producing any immediate symptoms to the patient. Screening for this disease is, thus, essential in order to establish early treatment of sight-threatening retinopathy and to avoid visual loss.. Different methods of screening for diabetic retinopathy have been used, including direct fundus examination and review of fundus photographs, obtained with or without mydriasis[241]. Annual screening for diabetic retinopathy is recommended for anyone with T1D who is aged 12 years or more and has had diabetes for five years or more, and for those with T2D, from the time of diagnosis[242, 241]. Careful follow-up with prompt intervention with laser photocoagulation and vitrectomy when necessary is the most effective method to reduce potential visual disabilities. However, despite the availability of successful treatments, a number of barriers to optimal care remain. These include a variety of financial, sociological, educational, and psychologic barriers to regular ophthalmic examinations[243]. Early detection of significant retinopathy and prompt treatment when necessary remain the fundamental goals in the effort to reduce visual disability in patients with diabetes[240, 244].

1.4.2 Metabolomics and diabetes

Metabolites are markers of biochemical, physiological, or pathological reactions and are able to show the interaction among different pathways that develop within a living cell providing an understanding of the physiology. One of the goals of metabolomics, in addition to understanding of physiologic pathways is to develop diagnostic biomarkers that could serve as tools for clinical practice, diagnosis, prognosis, and predictors of therapeutic response[245]

Some metabolomics studies[246, 247] indicate that carnitines, branched chain amino acids (BCAAs), aromatic amino acids (AAAs), and free fatty acids (FFAs) could be potential markers associated with dysglycemic states. Although there are many answers about their function in relation to diabetes, their precise role is not well defined yet. Observations on β-oxidation dysregulation have given new milestones for the understanding of the dysglycemic metabolic phenotype. Evidence shows that the most important changes have been observed on intermediary metabolism. Increased acylcarnitines suggest an overload of β-oxidation that cannot be matched by the tricarboxylic acid (TCA cycle)[248]. BCAAs and AAA have been found to be strongly related with early insulin resistance and T2D prediction independently from BMI[249]. It is possible that reduced valine and isoleucine catabolism could reduce TCA anaplerosis and flux, further decreasing oxidative capacity and fatty acid oxidation, and promoting incomplete lipid oxidation, Furthermore, acylcarnitines resulting from incomplete oxidation or other perturbations in both amino acid and fatty acid metabolism could promote oxidative stress and insulin resistance. This data supports the hypothesis that multiple steps of muscle BCAA metabolism are impaired in insulin resistant, and that these defects may not only contribute to concomitant alterations in TCA cycle activity but also interact to perturb lipid metabolism[250].

In the Framingham Heart Study, fasting blood levels of branched-chain and aromatic amino acids isoleucine, leucine, valine, tyrosine, and phenylalanine were elevated up to 12 years before the onset of clinical diabetes. Participants in the top quartile of a combination of three of these amino acids were at five-fold greater risk of developing diabetes, even after adjusting for age, fasting glucose, and other confounders[251]. Studies of branched chain amino acid supplementation in both animals and humans indicate that circulating amino acids may directly promote insulin resistance, possibly via disruption of insulin signaling in skeletal muscle. The underlying cellular mechanisms may include activation of the mTOR, JUN and IRS1 signaling pathways in skeletal muscle (Figure 26)[252]. By contrast, others have demonstrated that a diet specifically enriched in leucine can substantially decrease diet-induced obesity, hyperglycemia, and hypercholesterolemia[253].

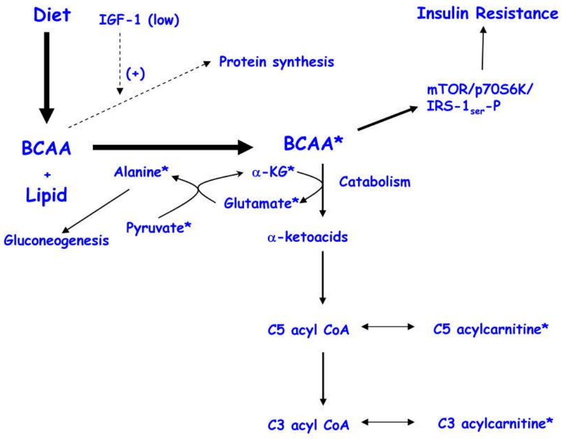


Figure 26. Schematic Summary of BCAA overload hypothesis. In the physiological context of overnutrition and low IGF-1 levels, as found in obese subjects, circulating branched-chain amino acids (BCAA) rise, leading to increased flux of these amino acids through their catabolic pathways. Changes were detected in several of the intermediary metabolites of the BCAA catabolic pathway in obese subjects, as indicated by the symbol *. A consequence of increased BCAA levels is the activation of the mTOR/S6K1 kinase pathway and phosphorylation of IRS-1 on multiple serines, contributing to insulin resistance. In addition, increased BCAA catabolic flux may contribute to increased gluconeogenesis and glucose intolerance via glutamate transamination to alanine. (adopted from [252])

Higher levels of another metabolite, 2-aminoadipate, have also been found to predict incident diabetes although are not correlated with other metabolite biomarkers of diabetes, such as branched chain amino acids and aromatic amino acids, suggesting they report on a distinct pathophysiological pathway[254]. Further work suggests that some of these branched-chain amino acids may originate from the microbiome, the serum metabolome of insulin-resistant individuals is characterized by increased levels of branched-chain amino acids (BCAAs), which correlate with a gut microbiome that has an enriched biosynthetic potential for BCAAs and is deprived of genes encoding bacterial inward transporters for these amino acids[255].

Another sub-study of the Framingham population found lipids with lower carbon number and fewer double bonds were associated with an increased risk of diabetes, whereas lipids of higher carbon number and more double bonds were associated with decreased risk[256]. The major lipid class implicated was triacylglycerols (TAGs). It has been hypothesized that a joint effect of both lipids, and amino acids may promote mitochondrial deficiency and disrupt insulin signaling particularly in skeletal muscle[257]. Further studies are needed to determine the relationship of these compounds to diabetes, and whether they play a causal role or are epiphenomenon associated with metabolic dysregulation in diabetes[245].

1.4.3 Metabolomics and Diabetic Retinopathy

DR can be investigated using different tissues; two of the most studied are vitreous samples (although the invasive nature of vitreous sampling limits study replication and the translational potential of any biomarkers identified from vitreous fluid)[258, 259, 260, 261] and plasma or sera[262, 263, 264], this last one, less invasive for the patient and more reproducible across studies.

- **Vitreous Markers**

Paris et al.[258] revealed that arginine metabolism and ammonia detoxification (urea cycle) were two of the most perturbed pathways. Elevated levels of methionine, allantoin, decanoylcarnitine, arginine, proline, citrulline, ornithine, and octanoylcarnitine were observed in patients with proliferative diabetic retinopathy. The authors speculated that these findings implicate compromised Mueller glial cell metabolism in disrupting neurovascular crosstalk within the retina, potentially promoting diabetic retinopathy progression.

Arginine is metabolized through two different pathways in the retina: the arginase pathway that produces ornithine and urea, and the nitric oxide synthase (NOS) pathway, which generates citrulline and NO. Overactivity of the enzyme arginase II causes a shortfall of arginine for the NOS pathway, resulting in reduced NO availability and uncoupling of the NOS pathway. This leads to consequent endothelial cell dysfunction and impaired vasodilation, which are characteristics of diabetic retinopathy. Lack of NO also results in increased generation of oxygen and nitrogen reactive species that accelerate diabetic retinopathy.

Barba et al. [259] reported higher lactate and lower galactitol and ascorbic acid levels in patients with proliferative diabetic retinopathy. As expected, glucose was significantly higher in samples from proliferative diabetic retinopathy patients than nondiabetic patients. The high levels of lactate likely reflect increased tissue acidosis and anaerobic glycolysis, particularly as the retina is one of the most metabolically active tissues. They also found lower levels of galactitol which they attributed to polyol pathway activation, a metabolic pathway involved in the pathogenesis of diabetic retinopathy.

- **Plasma Markers**

Chen et al. [262] reported significantly altered levels of 11 metabolites. These were decreased levels of 1,5-anhydroglucitol and increased levels of 1,5-gluconolactone, 2-deoxyribonic acid, 3,4-dihydroxybutyric acid, erythritol, gluconic acid, lactose/ cellobiose, maltose/trehalose, mannose, ribose, and urea. In the validation set, 2-deoxyribonic acid, 3,4-dihydroxybutyric acid, erythritol, gluconic acid, and ribose were found to remain elevated, while maltose was decreased. Pathway mapping of these metabolites showed significant enrichment of the pentose phosphate pathway that is responsible for the generation of NADPH to combat oxidative stress. Increased concentrations of cytosine, cytidine, and thymidine were also found associated with diabetic retinopathy compared to those without. Of these compounds, cytidine had the highest area under the curve (AUC) of 0.849 ± 0.048, and at the optimal cutoff point of 0.076 mg/L; the sensitivity and specificity of cytidine as a biomarker for diabetic retinopathy were 73.7 and 91.9%, respectively, suggesting potential value as a biomarker for diabetic retinopathy. This study had major strengths including an independent validation cohort, and adjustment for confounders such as diet and kidney disease, but had limited generalizability as only Singaporeans of South Indian ancestry were studied. Such work shows the importance of validating metabolomics studies in other populations and ethnic groups.

Another study by Li et al. [264] applied systems biology based approaches to study metabolomics of blood plasma in patients with diabetic retinopathy. This study was novel as patients were classified according to usual Western (International) classification systems of diabetic retinopathy, as well as to a Chinese Medicine classification. The authors reported that in 88 patients with type 2 diabetes, the Western classification was associated with ten metabolites (pyruvic acids, L-aspartic acid, β-hydroxybutyric acid, methymalonic acid, citric acid, glucose, stearic acid, trans-oleic acid, linoleic acid, arachidonic acid), while the Chinese classification was associated with four metabolites (pyruvic acids, L-aspartic acid, glycerol, and cholesterol). Pyruvic acid and L-aspartic acid were identified in both classification systems. What was unclear in this study was the control group and whether they were free from diabetic retinopathy. The authors acknowledged another limitation as lack of data on the presence of chronic kidney disease, which could lead to impaired renal excretion of aspartic acid and other non-essential amino acids and result in the elevated levels observed in cases. The association of lower plasma levels of two omega-6 polyunsaturated fatty acids (i.e., the PUFAs arachidonic acid and linoleic acid) with proliferative diabetic retinopathy is of interest as it has been reported that lower levels of these two PUFAs may be associated with higher levels of circulating pro-inflammatory markers such as IL-1ra and IL- 6 and lower anti-inflammatory marker TGFb[265]. Diabetic retinopathy is associated with pro-inflammatory cytokines and downregulation of omega-6 PUFAs may potentiate these effects and increase the risk of developing the condition.

1.4.4 Proteomics and Diabetic Retinopathy

The study of proteins differentially expressed in the eyes of patients with DR may provide biomarkers which can help understand the pathological mechanism of DR and to be used as potential biomarkers in the diagnosis. In this context, the qualitative and quantitative analysis of the vitreous proteome alterations has great potential in identifying better diagnostic/prognostic markers and drug target candidates for improved therapeutic interventions. Since multiple vitreous-resident proteins will contribute to the pathogenic mechanisms in human DR, a global proteome-wide analysis is of high importance.

The proteomic method for scientific analysis is an approach rapidly to survey the proteome (complete inventory of proteins expressed within a biologic sample). The application of the proteomic method for comparison of disease and control samples allows for the rapid development of a hypothesis used to understand or explain aspects of disease biology, such as disease initiation, progression, or remission.

- **Proteomics in the vitreous**

Some studies such as Feener et al.[266] or Varjosalo et al.[267] found proteins up-regulated in the PDR samples belonged to complement factors as well as regulators of coagulation. Components of the complement system are present already in the early stages of the DR, but during the pathogenesis the amount of several cascade components increases dramatically. The main consequences of the complement cascade are activation of the host defence against infection, interface between innate and adaptive immunity (e.g., augmentation of antibody responses), and disposal of cellular waste (e.g., clearance of immune complexes from tissues). Complement activation in the immediate vicinity of the capillaries could provoke pathologic alterations characterized by the presence of increased numbers of polymorphonuclear granulocytes in degenerative capillaries and by the appearance of protein deposits that possibly reflect endothelial leakage with insudation of plasma components to the Bruch membrane[268].

Varjosalo et al.[267] also detected both vascular endothelial growth factor receptor (VEGFR)-1 and -2, which regulate neovascularization and vascular leakage in the DR vitreous, in low amounts. The concept of retinal neovascularisation includes the potential involvement of activated microglia and blood borne inflammatory mediators, and systemic inflammation is an intrinsic response to diabetes. Diabetes increases the release of inflammatory mediators (such as interleukins, tumor necrosis factor (TNF), intercellular adhesion molecules, integrins and semaphorins, and angiotensin-2), and activation of microglial cells has been shown to occur in early stages of DR[269]. Altogether, their findings indicate that inflammation plays a significant role in DR conditions.

CHAPTER 2. Objectives

General objectives

Analyse and develop new strategies for converting raw MS-based metabolomics data into biological knowledge.

Study alterations in the proteome and metabolome of human retinal pigment epithelium cells exposed to hyperglycemic and/or hypoxic conditions.

To reach these two general objectives, I have developed my thesis focusing on:

Methodological aims:

(i) Analyse mass spectral databases for LC/MS-based untargeted metabolomics.
(ii) Generate and improve the characterization of LC/MS metabolomics data focusing on MS1 and MS2 annotation.

Biological aims:

- Detect and analyse changes in protein-protein interaction networks by hyperglycemic and/or hypoxic conditions.
- Predict and validate metabolite alterations due to hyperglycemic and/or hypoxic conditions integrating protein expression data in metabolic networks.

CHAPTER 3. Results

3.1 LC-MS processing: from MS1 to MS2 to metabolite ID

The analysis of biological samples in untargeted metabolomic studies using liquid chromatography coupled to electrospray mass spectrometry (LC ESI-MS) results in tens of thousands of ion signals or features. It is now well accepted that this large number of features is an overestimation of the real number of different compounds in the sample, mainly because single metabolites can be detected as multiple ions of different mass in either positive or negative ionization mode. This redundancy of features is mostly due to in-source phenomena including cation adduction, multimerization and in-source fragmentation, plus contaminants. The annotation of features, understood as their feature relationships in MS1 mode, is a challenging task and represents a serious obstacle for the real high-throughput analysis of metabolomics data. On the other hand, structural annotation of metabolites relies mainly on tandem mass spectrometry (MS/MS or MS2) analysis. Unfortunately, approximately 90% of the known metabolites reported in metabolomic databases do not have annotated spectral data from standards. This situation has fostered the development of computational tools that predict fragmentation patterns in silico and compare these to experimental MS2 spectra.

This chapter addresses some of the above issues by (i) analysing metabolomic databases characterized by well-annotated mass spectra containing the main adducts for reference substances and (ii) generating and improving the characterization of LC/MS metabolomics data focusing on MS1 and MS2 annotation.

3.1.1 NIST database for LC/MS-based metabolomics

Comprehensive and well-annotated MS-based spectral databases are essential for a good identification step. The exact mass (i.e., m/z) of a selected feature is often used to query compound-centric databases. However, database hits provide only putative assignments that must be further validated by retention time matching and/or MS/MS analysis. In the absence of a pure standard analyzed under identical analytical conditions, MS/MS data searched against a reference MS/MS database are typically the most conclusive evidence for validating and annotating a metabolite feature using MS[144].

One of these databases is the spectral database of the US National Institute of Science and Technology (NIST). Historically developed as an EI–MS database, the NIST also contains >230K ESI MS/MS spectra of small molecules, including authentic chemical standards of metabolites, lipids, biologically active peptides, and all di-peptides and tryptic tripeptides. In particular, the NIST14 provides 51,216 ion-trap spectra from 8171 compounds, and 183,068 CID spectra (Q-

TOF and triple quad) from 7692 compounds are included (Table 6). Of note, many of the MS/ MS data-containing precursor ions in NIST14 correspond to common adducts formed during ESI (in addition to [M + H]+ in positive-ion mode and [M-H]− in negative-ion mode), including [M + H-H2O]+, [M+ Na]+, [M + NH4]+, [M+ H-NH3]+, [2M + H]+, [M-H-H2O]−, [2M-H]−, and [M- 2H]2− (Figure 27). Few spectral databases such as NIST[270] offer a comparable number of MS/MS spectra from adducts. This is particularly useful because predominant adducts in the ESI spectrum vary from one metabolite to another as well as on the mobile phase used. One example is glucose-6-phosphate, which ionizes predominantly as [M + Na]+ or [M + NH4]+ using formic acid (0,1%) and ammonium enriched mobile phases, respectively (Figure 28A). Each precursor adduct, in turn, results in different MS/MS spectra (Figure 28B)[144].

Table 6. MS/MS data from electrospray ionization in NIST 14

Mass Analyzer	# spectra	# compounds	# ions
Ion trap	~39,000	~6000	~39,000
Q-TOF (CID)	~42,000	~3000	~4000
QqQ (CID)	~27,000	~1000	~3000
Orbitrap (HCD)	~69,000	~3000	~6000
Ion trap with FTMS	~5000	~2500	NA

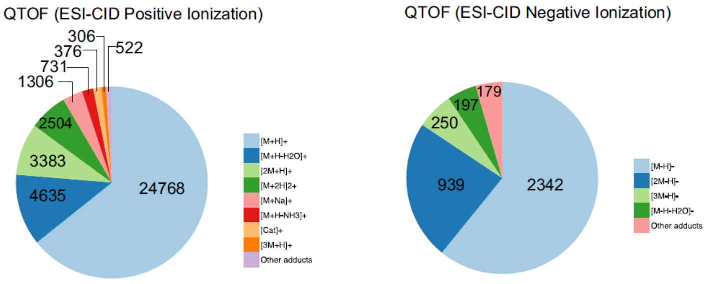


Figure 27. Pie charts representing the distribution of adducts and number of MS/MS spectra recorded in NIST 14 database using Agilent 6530 Q-TOF (ESI-CID) operating in either positive or negative ionization mode.

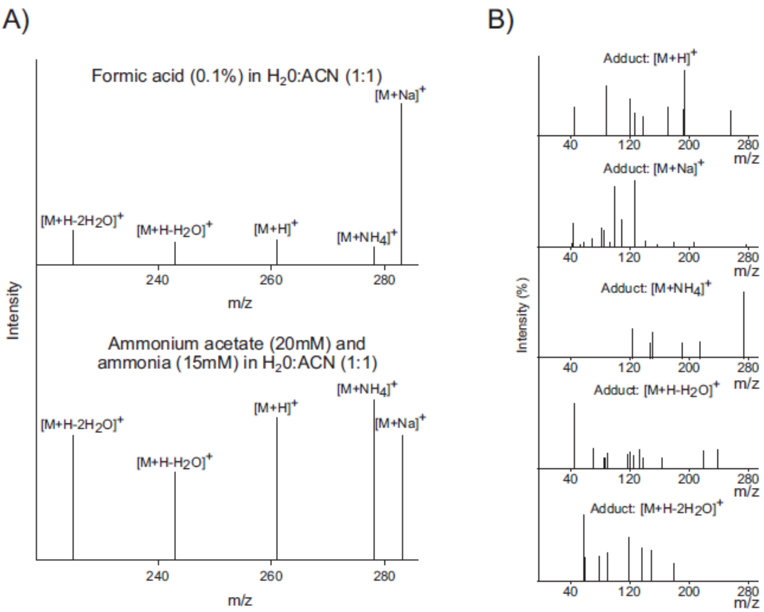


Figure 28. (A) Mass spectra of glucose 6-phosphate (KEGG cpd: C00092) under different analytical conditions (i.e., mobile phases). (B) MS/MS spectra from different precursor ion species corresponding to observed adducts.

This proves the importance of annotating features in MS1 mode before structural annotation in MS2 mode.

3.1.2 CliqueMS: A computational tool for annotating in-source MS1 untargeted metabolomics data based on a coelution similarity network

Introduction

The two main grouping principles for detecting and annotating features related to a metabolite are chromatographic peak-shape similarity (i.e., coeluting features) and peak-abundance correlation, or a combination thereof. Pairwise intensity correlation analysis across multiple samples is the basis of computational tools such as AStream[116], MSClust[117], RAMClust[118], MSFLO[119], compMS2Miner[120] or findMAIN[121]. On the other hand, peak shape similarity is used by CAMERA[122], MET-COFEA[123], ALLocator[271] or MZmine2[124]. MetAssign[125] or xMSannotator[126] have also included a probabilistic score to measure the confidence in particular assignments based on statistical clustering. However, as metabolomics data can be highly complex, because of coelution in the LC, differences in the ion source parameters and mass-resolving power of various instruments, misannotations can still prevent the correct identification of a large number of metabolites. If features are considered individually, the presence of adducts and isotopes should not be neglected and many database searches allow the specification of expected adducts when performing queries[144].

CliqueMS annotates adducts in complex LC/MS samples based on the following assumptions: i) features of the same metabolite corresponding to in-source phenomena, including adducts (e.g., Na, K) and fragments (e.g., loss of water), display similar chromatographic elution profiles; and ii) in-source phenomena (such as adducts or fragments) occur with a probability equal to the frequency with which they are observed in experiments. I have contributed to the development of CliqueMS by annotating redundant MS1 features in complex biological samples exploiting chromatographic peak-shape similarity and a calculated natural frequency of adduct formation observed in real complex biological samples and pure compounds. CliqueMS implements a novel mathematical approach to obtain the most plausible groupings of features according to a similarity network. Next, CliqueMS annotates features and ranks annotations using an estimated frequency of dominant adducts and mass fragments in complex biological samples and from all available compounds in the NIST 14 MS/MS library (Figure 29).

Here I demonstrate that CliqueMS correctly identifies and annotates a larger number of adducts, leading to more correct parental ion neutral masses than existing widely-used approaches such as CAMERA[122], for both pure and complex samples.

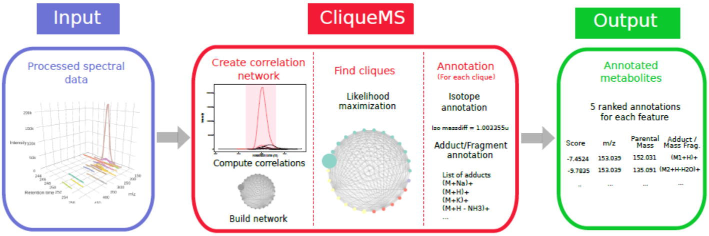


Figure 29. Schematic representation of CliqueMS. Given processed spectral data of a complex metabolite sample, CliqueMS identifies the features belonging to the same metabolite. First, CliqueMS establishes similarities between all features, and looks for cliques in the similarity network. Then, for each clique, CliqueMS proceeds to annotate each feature by establishing the parental ion neutral mass. The final output is, for each group, a scored list containing all annotated adducts with their corresponding parental masses.

Materials and methods

LC/MS grade methanol (MeOH) and acetonitrile (ACN) and analytical grade chloroform (CHCl3) were purchased from SDS (Peypin, France). Water was produced in an in-house Milli-Q purification system (Millipore, Molsheim, France). Formic acid and ammonium fluoride were purchased from Sigma-Aldrich (Steinheim, Germany).

- Standards: 9 standards (riboflavine, 1,2-distearoyl-sn-glycero-3-phosphocholine, biotin, cholic acid, deoxycholic acid, L-methionine sulfoxide, thymine, uracil and fructose) were pulled to a final concentration of 1ppm in H2O:ACN (5:95) with 0.1 % formic acid.

Retinas from Irs-2 (-/-) mice. Irs-2-deficient mice were generated initially on a C57BL6/J: SV129 background and then backcrossed to establish a pure C57BL6/J background[272]. Thus, the offspring resulting from the breeding of Irs-2(-/-) with RIP-Irs-2 line were C57BL6/J. The generation and genotyping of the Irs-2(-/-) and the RIP- Irs-2(-/-) models have been described previously[272,273]. The final set of samples consisted on 12 Irs-2 males and 14 wild type males. Mice were euthanized using CO2 and cervical dislocation. The eyes were removed, and retinas were separated from retinal pigment epithelium and kept at -80ºC in the freezer.

ARPE-19 cell culture. Cells were cultured under standard conditions in DMEM/F12 (1:1 mixture of Dulbecco's modified Eagle's medium and Ham's F12), 10% fetal calf serum (FCS) and penicillin/streptomicin. ARPE-19 cells from passage 20-23 were used and the media was changed every 3 days. Cells grown in these conditions constitute a monolayer that retains the functionality, polarity and tight junction expression of the human RPE. For our study, cells were seeded in Petri dishes (10 cm) at 0.4 x104 cells/mL and maintained in culture for 21 days with 5.5 mM D-Glucose at 37°C under 5% (v/v) CO2 in an incubator. During the last 24 hours cells were subjected to serum deprivation (1% FCS). Serum deprived media were prepared with 5.5 mM D-Glucose and cultured in normoxic or hypoxic (1% O2) conditions. Each condition was run in triplicate.

Metabolites extraction method

- Retinas: retinas were first lyophilized and metabolites were extracted adding 190 µL of MeOH and 120 µL of H2O, then vortex during 30 seconds. Afterwards, samples were frozen during 1 min in N2 liq. and thawed by cold sonication during 30 seconds. This step was applied three times. Then 380 µL of chloroform were added and vortexed during 30 seconds. Finally, samples were centrifuged (15000 rpm, 15 min a 4ºC). The supernatant was extracted and dried. The sample was suspended in 100 µL of H2O:MeOH (1:1) and stored at −80 °C until further analysis.

- ARPE-19 cells: After removing the cell medium of ARPE-19, metabolites were extracted into a extraction solvent by adding 2 mL of a cold mixture of chloroform/methanol (2:1 v/v). The resulting suspension was bath-sonicated for 3 minutes, and 2 mL of cold water was added. Then, 1 mL of chloroform/methanol (2:1 v/v) was added to the samples and bath-sonicated for 3 minutes. Cell lysates were centrifuged (5000 × g, 15 min at 4 C) and the aqueous phase was carefully transferred into a new tube. The sample was frozen, lyophilized and stored at −80 °C until further analysis.

Data acquisition

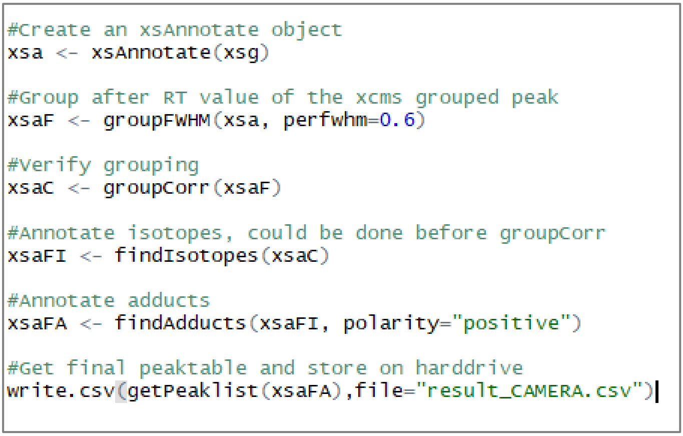
LC/MS analyses were performed using an UHPLC system (1290 series, Agilent Technologies) coupled to a 6550 ESI-QTOF MS (Agilent Technologies) operated in positive (ESI+) or negative (ESI-) electrospray ionization mode. Vials containing extracted metabolites were kept at −20 °C prior to LC/MS analysis. When the instrument was operated in positive ionization mode, metabolites were separated using an Acquity UPLC (HSS T3) C18 reverse phase (RP) column (2.1 x 150 mm, 1.8 µm) and the solvent system was A1 = 0.1% formic acid in water and B1 = 0.1% formic acid in acetonitrile. When the instrument was operated in negative ionization mode, metabolites were separated using an Acquity UPLC (BEH) C18 RP column (2.1 x 150 mm, 1.7 µm) and the solvent system was A2 = 1 mM ammonium fluoride in water and B2 = acetonitrile, as previously reported[1]. The linear gradient elution started at 100% A (time 0–2 min) and finished at 100% B (10-15 min). The injection volume was 5 µL. ESI conditions: gas temperature, 150 °C; drying gas, 13 L min–1; nebulizer, 35 psig; fragmentor, 400 V; and skimmer, 65 V. The instrument was set to acquire over the m/z range 100–1500 in full-scan mode with an acquisition rate of 4 spectra/sec. MS/MS was performed in targeted mode, and the instrument was set to acquire over the m/z range 50–1000, with a default isolation width (the width half-maximum of the quadrupole mass bandpass used during MS/MS precursor isolation) of 4 m/z. The collision energy was fixed at 20 V. Moreover, the samples were acquired in autoMSMS mode in which the precursor selection filters consisted on 3 maximum precursors ions per cycle, a threshold of 5000 counts, 250000 counts/spectrum as target, 0.20 min of active exclusion released after and 1 spectra for active exclusion excluded after.

The mixture of standards were separated using an Acquity UPLC BEH HILIC column (2.1 x 150 mm, 1.8 µm) and the solvent system was A1 = 20mM ammonium acetate and 15 mM NH4OH in water and B1 = 95% ACN and 5% H2O. Samples were operated in positive electrospray ionization mode. The linear gradient elution started at 100% B (time 0–2 min) and finished at 75% A (10-15 min). Electrospray conditions and acquisition of spectra were similar to the complex biological samples.

Data processing and statistical analysis

LC/MS (ESI+ and ESI− mode) data was processed using the XCMS software[102] to detect and align features. A feature is defined as a molecular entity with a unique m/z and a specific retention time (mzRT). XCMS analysis of these data provided a matrix containing the retention time, m/z value, and integrated peak area of each feature for every sample. Quality control samples (QCs) consisting of pooled samples from each two conditions were used in LC/MS analyses. QCs were injected at the beginning and periodically every 5 samples. Furthermore, samples entering the study were entirely randomized to reduce systematic error associated with instrumental drift. QCs were always projected in a PCA model together with the samples under study to verify that technical issues do not mask biological information. The performance of the analytical platform for each detected mzRT feature in samples was assessed by calculating the relative standard deviation of these features on pooled samples (CVQC) according to Vinaixa et al.[170]. Samples were compared using the integrated peak area of each feature via Student's t-Test and assigning a fold value to indicate the level of differential regulation due Irs2 condition. Differentially regulated metabolites that were statistically significant ($p<0.05$) detected by LC/MS were characterized by MS/MS. Data pre-processing, data analysis, and statistical calculations were performed in R (R-3.4.1).

CAMERA annotation: we run CAMERA (CAMERA_1.30.0) in R using the preprocessed data from XCMS (xcms_1.50.1). I used the recommended workflow and parameters from the manual (Figure 30).



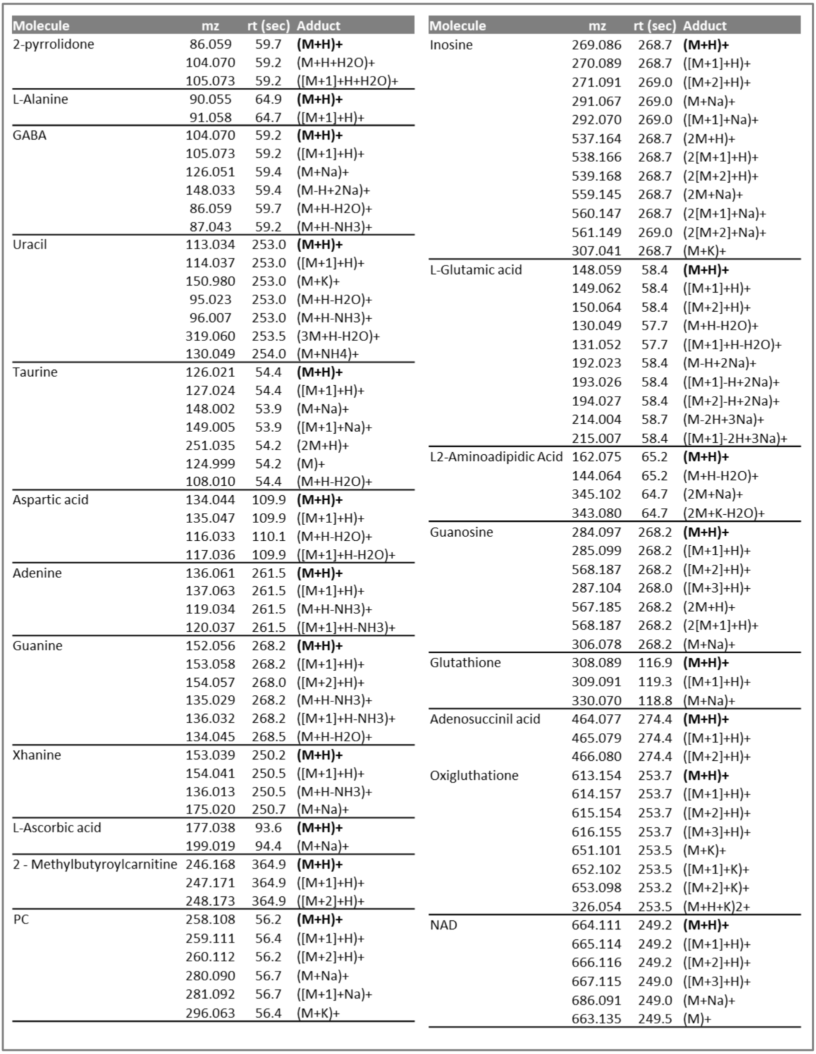
```
#Create an xsAnnotate object
xsa <- xsAnnotate(xsg)

#Group after RT value of the xcms grouped peak
xsaF <- groupFWHM(xsa, perfwhm=0.6)

#Verify grouping
xsaC <- groupCorr(xsaF)

#Annotate isotopes, could be done before groupCorr
xsaFI <- findIsotopes(xsaC)

#Annotate adducts
xsaFA <- findAdducts(xsaFI, polarity="positive")

#Get final peaktable and store on harddrive
write.csv(getPeaklist(xsaFA),file="result_CAMERA.csv")
```

Figure 30. CAMERA script and parameters.

Results

To test the accuracy of CAMERA to annotate features in MS1 mode, we identified 20 compounds ($[M+H]^+$ adduct) via MS2 fragmentation in ARPE-19 cells (Table 7). Moreover, we manually annotated all associated adducts for each compound.

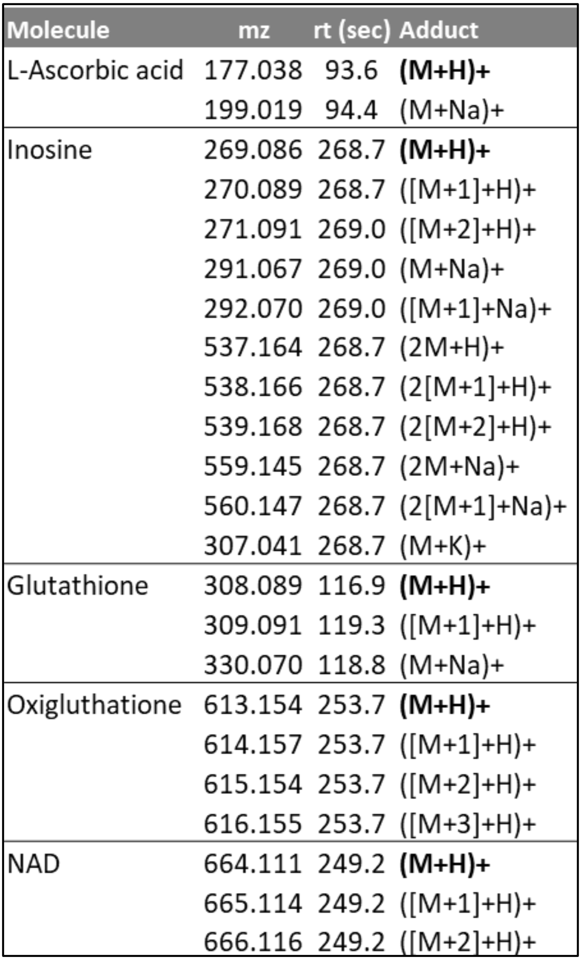
Table 7. List of identified molecules with their associated adducts in a real experimental dataset.



Molecule	mz	rt (sec)	Adduct
2-pyrrolidone	86.059	59.7	**(M+H)+**
	104.070	59.2	(M+H+H2O)+
	105.073	59.2	([M+1]+H+H2O)+
L-Alanine	90.055	64.9	**(M+H)+**
	91.058	64.7	([M+1]+H)+
GABA	104.070	59.2	**(M+H)+**
	105.073	59.2	([M+1]+H)+
	126.051	59.4	(M+Na)+
	148.033	59.4	(M-H+2Na)+
	86.059	59.7	(M+H-H2O)+
	87.043	59.2	(M+H-NH3)+
Uracil	113.034	253.0	**(M+H)+**
	114.037	253.0	([M+1]+H)+
	150.980	253.0	(M+K)+
	95.023	253.0	(M+H-H2O)+
	96.007	253.0	(M+H-NH3)+
	319.060	253.5	(3M+H-H2O)+
	130.049	254.0	(M+NH4)+
Taurine	126.021	54.4	**(M+H)+**
	127.024	54.4	([M+1]+H)+
	148.002	53.9	(M+Na)+
	149.005	53.9	([M+1]+Na)+
	251.035	54.2	(2M+H)+
	124.999	54.2	(M)+
	108.010	54.4	(M+H-H2O)+
Aspartic acid	134.044	109.9	**(M+H)+**
	135.047	109.9	([M+1]+H)+
	116.033	110.1	(M+H-H2O)+
	117.036	109.9	([M+1]+H-H2O)+
Adenine	136.061	261.5	**(M+H)+**
	137.063	261.5	([M+1]+H)+
	119.034	261.5	(M+H-NH3)+
	120.037	261.5	([M+1]+H-NH3)+
Guanine	152.056	268.2	**(M+H)+**
	153.058	268.2	([M+1]+H)+
	154.057	268.0	([M+2]+H)+
	135.029	268.2	(M+H-NH3)+
	136.032	268.2	([M+1]+H-NH3)+
	134.045	268.5	(M+H-H2O)+
Xhanine	153.039	250.2	**(M+H)+**
	154.041	250.5	([M+1]+H)+
	136.013	250.5	(M+H-NH3)+
	175.020	250.7	(M+Na)+
L-Ascorbic acid	177.038	93.6	**(M+H)+**
	199.019	94.4	(M+Na)+
2 - Methylbutyroylcarnitine	246.168	364.9	**(M+H)+**
	247.171	364.9	([M+1]+H)+
	248.173	364.9	([M+2]+H)+
PC	258.108	56.2	**(M+H)+**
	259.111	56.4	([M+1]+H)+
	260.112	56.2	([M+2]+H)+
	280.090	56.7	(M+Na)+
	281.092	56.7	([M+1]+Na)+
	296.063	56.4	(M+K)+
Inosine	269.086	268.7	**(M+H)+**
	270.089	268.7	([M+1]+H)+
	271.091	269.0	([M+2]+H)+
	291.067	269.0	(M+Na)+
	292.070	269.0	([M+1]+Na)+
	537.164	268.7	(2M+H)+
	538.166	268.7	(2[M+1]+H)+
	539.168	268.7	(2[M+2]+H)+
	559.145	268.7	(2M+Na)+
	560.147	268.7	(2[M+1]+Na)+
	561.149	269.0	(2[M+2]+Na)+
	307.041	268.7	(M+K)+
L-Glutamic acid	148.059	58.4	**(M+H)+**
	149.062	58.4	([M+1]+H)+
	150.064	58.4	([M+2]+H)+
	130.049	57.7	(M+H-H2O)+
	131.052	57.7	([M+1]+H-H2O)+
	192.023	58.4	(M-H+2Na)+
	193.026	58.4	([M+1]-H+2Na)+
	194.027	58.4	([M+2]-H+2Na)+
	214.004	58.7	(M-2H+3Na)+
	215.007	58.4	([M+1]-2H+3Na)+
L2-Aminoadipidic Acid	162.075	65.2	**(M+H)+**
	144.064	65.2	(M+H-H2O)+
	345.102	64.7	(2M+Na)+
	343.080	64.7	(2M+K-H2O)+
Guanosine	284.097	268.2	**(M+H)+**
	285.099	268.2	([M+1]+H)+
	568.187	268.2	([M+2]+H)+
	287.104	268.0	([M+3]+H)+
	567.185	268.2	(2M+H)+
	568.187	268.2	(2[M+1]+H)+
	306.078	268.2	(M+Na)+
Glutathione	308.089	116.9	**(M+H)+**
	309.091	119.3	([M+1]+H)+
	330.070	118.8	(M+Na)+
Adenosuccinil acid	464.077	274.4	**(M+H)+**
	465.079	274.4	([M+1]+H)+
	466.080	274.4	([M+2]+H)+
Oxigluthatione	613.154	253.7	**(M+H)+**
	614.157	253.7	([M+1]+H)+
	615.154	253.7	([M+2]+H)+
	616.155	253.7	([M+3]+H)+
	651.101	253.5	(M+K)+
	652.102	253.5	([M+1]+K)+
	653.098	253.2	([M+2]+K)+
	326.054	253.5	(M+H+K)2+
NAD	664.111	249.2	**(M+H)+**
	665.114	249.2	([M+1]+H)+
	666.116	249.2	([M+2]+H)+
	667.115	249.0	([M+3]+H)+
	686.091	249.0	(M+Na)+
	663.135	249.5	(M)+

Next we run CAMERA and checked if the manually identified adducts were annotated by CAMERA. Unfortunately, very few compounds were annotated correctly. Only 5 out of the 20 identified compounds were annotated correctly by CAMERA. Table 8 shows the complete list of annotated adducts by CAMERA.

Table 8. List of correctly annotated adducts by CAMERA



Molecule	mz	rt (sec)	Adduct
L-Ascorbic acid	177.038	93.6	**(M+H)+**
	199.019	94.4	(M+Na)+
Inosine	269.086	268.7	**(M+H)+**
	270.089	268.7	([M+1]+H)+
	271.091	269.0	([M+2]+H)+
	291.067	269.0	(M+Na)+
	292.070	269.0	([M+1]+Na)+
	537.164	268.7	(2M+H)+
	538.166	268.7	(2[M+1]+H)+
	539.168	268.7	(2[M+2]+H)+
	559.145	268.7	(2M+Na)+
	560.147	268.7	(2[M+1]+Na)+
	307.041	268.7	(M+K)+
Glutathione	308.089	116.9	**(M+H)+**
	309.091	119.3	([M+1]+H)+
	330.070	118.8	(M+Na)+
Oxigluthatione	613.154	253.7	**(M+H)+**
	614.157	253.7	([M+1]+H)+
	615.154	253.7	([M+2]+H)+
	616.155	253.7	([M+3]+H)+
NAD	664.111	249.2	**(M+H)+**
	665.114	249.2	([M+1]+H)+
	666.116	249.2	([M+2]+H)+

With this example we demonstrate the low performance of CAMERA since it only was able to annotate correctly 25% of the total manually identified compounds. For this reason we decided to develop a new tool called CliqueMS to annotate more efficiently the adducts, and in this way, improve our metabolomics data analysis workflow.

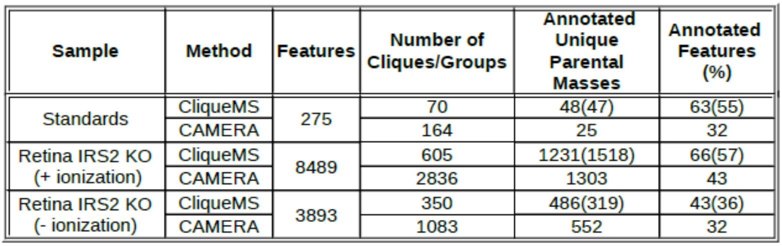
To validate the accuracy of CliqueMS we performed two kinds of experiments. First, we look at the accuracy at annotating relatively simple samples corresponding to mixture of standards for which we have a manual annotation. Second, we use CliqueMS to annotate complex samples for which we also have partial manual annotations confirmed via MS2 fragmentation patterns. We look at the accuracy of CliqueMS at correctly annotating the manually identified compounds in the samples. For reference, we compare the accuracy of CliqueMS to CAMERA's.

Mixture of standards

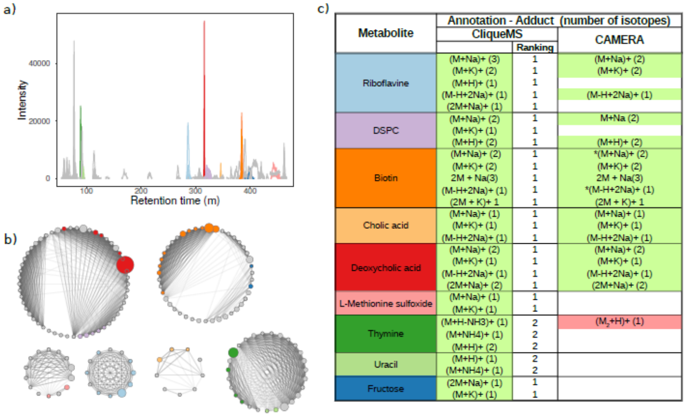
Table 9 and Figure 31 show that overall CliqueMS produces better annotations than CAMERA. CliqueMS is able to correctly identify more manually annotated metabolites, and correctly annotate more features associated to these metabolites by both correctly identifying adducts/mass fragments and their isotopes. The reason for this superior performance is two-fold. First, CliqueMS identifies a smaller number of feature groups so that features associated to the same metabolite are in the same group (Figure 31a-b). By contrast, CAMERA generates a larger number of groups which results in assigning features corresponding to the same metabolite to different groups. Furthermore, CliqueMS is better at identifying isotopes as shown in Figure 31c. Overall this results in CliqueMS being able to assign a parental mass to a larger number of features than CAMERA. Moreover, CliqueMS is able to correctly annotate all 9 metabolites within the 2 most plausible annotations for each clique (since for each metabolite, CliqueMS provides the correct annotation for at least one adduct/mass fragment and its corresponding isotopes within the two highest ranked annotations). The total number of annotated features corresponding to the standard compounds is 42 (of which 29 correspond to adducts/mass fragments and 13 to isotopes). Instead CAMERA annotates correctly 5 molecules and a total of 27 features. Note that even if we only considered the highest ranked annotation provided by CliqueMS, the number of correctly annotated metabolites (7) would be higher than for CAMERA (5).

Note that overall CliqueMS identifies a number of unique parental masses that is substantially larger than 9 (47 if we consider the best ranked annotation – see Table 9. The reason for this is two-fold: First during the process to obtain the MS1 data, metabolites can break down into smaller fragments that can also become ionized. Because the fragments that one might expect are different for each metabolite in the annotation step, CliqueMS is not considering these effects in the annotation step, therefore these fragments are assigned different parental masses. Second, we are not filtering data at short/long retention times so that some of these annotated parental masses could correspond to the solvent or additives used in the sample preparation process. Despite these facts, the difference in the percentage of features for which a parental mass is reported–63% (or 55% if we consider exclusively the annotation with the largest score) vs. 32 %– is substantial and is a direct effect of the aforementioned factors: high quality feature grouping and accurate isotope identification.

Table 9. Summary of the full set of annotations for each sample. We show the total number of features in the MS1 spectrum. For CliqueMS and CAMERA we report the total number of groups identified by the algorithm, the number of unique parental masses identified, and the percentage of features each algorithm associated to a parental mass. For the output of CliqueMS we consider the 5 annotations with the highest scores. For this reason we report: i) the average number of unique parental masses over annotations, and, in parenthesis, the number of unique parental masses in the annotation with the best score; ii) the number of features with at least one annotation within the 5 annotations with best scores, and, in parenthesis, the number of features annotated within the best ranked annotation.



Sample	**Method**	**Features**	**Number of Cliques/Groups**	**Annotated Unique Parental Masses**	**Annotated Features (%)**
Standards	CliqueMS	275	70	48(47)	63(55)
	CAMERA		164	25	32
Retina IRS2 KO (+ ionization)	CliqueMS	8489	605	1231(1518)	66(57)
	CAMERA		2836	1303	43
Retina IRS2 KO (- ionization)	CliqueMS	3893	350	486(319)	43(36)
	CAMERA		1083	552	32



Metabolite	Annotation - Adduct (number of isotopes)		
	CliqueMS		CAMERA
		Ranking	
Riboflavine	(M+Na)+ (3)	1	(M+Na)+ (2)
	(M+K)+ (2)	1	(M+K)+ (2)
	(M+H)+ (1)	1	
	(M-H+2Na)+ (1)	1	(M-H+2Na)+ (1)
	(2M+Na)+ (1)	1	
DSPC	(M+Na)+ (2)	1	M+Na (2)
	(M+K)+ (1)	1	
	(M+H)+ (2)	1	(M+H)+ (2)
Biotin	(M+Na)+ (2)	1	*(M+Na)+ (2)
	(M+K)+ (2)	1	(M+K)+ (2)
	2M + Na(3)	1	2M + Na(3)
	(M-H+2Na)+ (1)	1	*(M-H+2Na)+ (1)
	(2M + K)+ 1	1	(2M + K)+ 1
Cholic acid	(M+Na)+ (1)	1	(M+Na)+ (1)
	(M+K)+ (1)	1	(M+K)+ (1)
	(M-H+2Na)+ (1)	1	(M-H+2Na)+ (1)
Deoxycholic acid	(M+Na)+ (2)	1	(M+Na)+ (2)
	(M+K)+ (1)	1	(M+K)+ (1)
	(M-H+2Na)+ (1)	1	(M-H+2Na)+ (1)
	(2M+Na)+ (2)	1	(2M+Na)+ (2)
L-Methionine sulfoxide	(M+Na)+ (1)	1	
	(M+K)+ (1)	1	
Thymine	(M+H-NH3)+ (1)	2	$(M_2+H)+$ (1)
	(M+NH4)+ (1)	2	
	(M+H)+ (2)	2	
Uracil	(M+H)+ (1)	2	
	(M+NH4)+ (1)	2	
Fructose	(2M+Na)+ (1)	1	
	(M+K)+ (1)	1	

Figure 31. Feature annotation for a mixture of standards. a) Extracted Ion Chromatogram (EIC). The nine ionized metabolites were annotated with CliqueMS. In colors we show features that are adducts of each metabolite, as annotated by CliqueMS in (c). b) Cliques identified by CliqueMS in the same experiment, after computing cosine correlation and maximizing clique likelihood. The intensity of the link is proportional to the correlation, and the size of each node is proportional to feature intensity. The colors are the same as in (a). c) Feature annotation by CliqueMS and CAMERA. For each metabolite we show the different adducts and fragments annotated; in parenthesis we show the total number of isotopic variants of that particular adduct/mass fragment. Correctly annotated features are shown in green; incorrectly annotated features are shown in red, with M2 indicating that the associated parental mass was incorrect; non-annotated features are shown in white. For CliqueMS we also show the ranking of the feature annotation that matches manual annotation. For CAMERA the * indicates those features for which the algorithm returned two possible annotations. See Table S6 for CliqueMS annotations.

Complex biological samples

To evaluate the capacity of CliqueMS to identify adducts in complex MS1 data we analyzed real retina samples from a mouse model in which the gene irs2 had been knocked out. We analyzed spectral samples with both positive and negative ionizations. The positive ionization spectra contained 8489 features reduced to 606 cliques by CliqueMS, whereas the negative ionization spectra comprised 3893 features reduced to 354 cliques. Instead, as for the previously studied sample, CAMERA identifies a much larger number of groups: 2836 for positive ionization spectra, and 1083 for the negative ionization spectra.

CliqueMS groups the features into a smaller number of groups than CAMERA does. However, in contrast to the results for the mixture of standards, each clique is not necessarily associated to a single metabolite. In fact, because metabolite coelution is so frequent in samples with a large number of features, CliqueMS can group features corresponding to different metabolites within the same clique (see Figure 32).

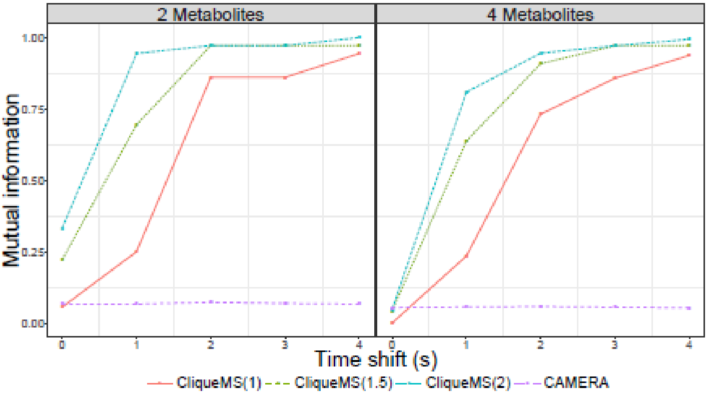
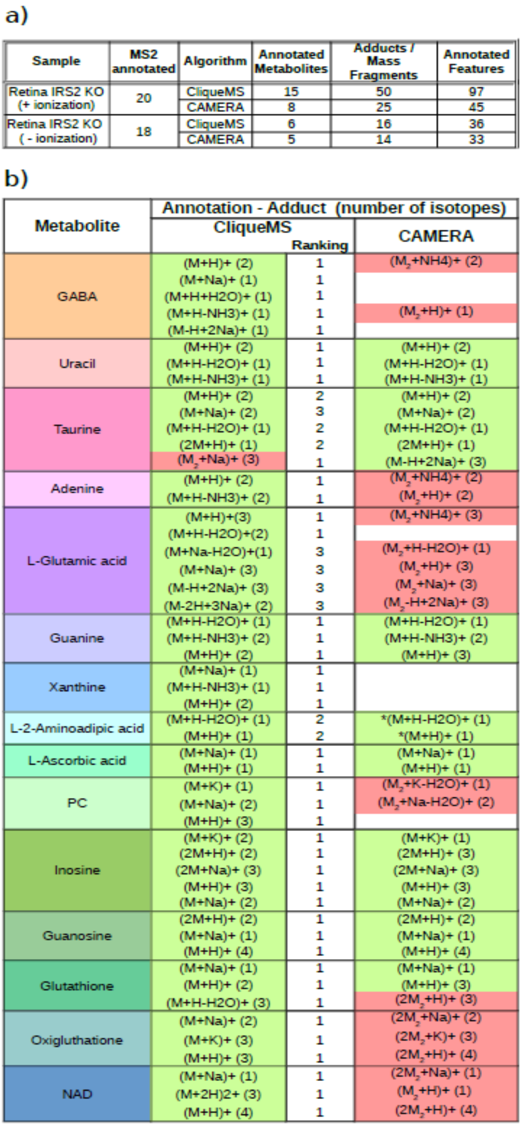


Figure 32. Identification of groups of features with similar coelution patterns. We simulate the coelution of 2 and 4 metabolites at different time shifts. The time shift is the time difference between the most intense feature of each metabolite. We run our clique identification algorithm for = 1; 1:5 and 2 (see Eq. (3)) and the group identification algorithm in CAMERA, which is also network based, for reference. To asses the accuracy of the group label assignments produced by each algorithm with respect to the known groupings, we use the adjusted mutual information (AMI). We show the AMI versus the time shift for the different algorithms we consider. The grouping algorithm within CliqueMS returns grouping closer to the nominal ones for all the time shifts considered. We find that for CliqueMS overall the best group assignments correspond to = 2.

In Figure 33 we show that, overall, CliqueMS provides a better annotation than CAMERA; specifically, CliqueMS is able to identify a larger number of the MS2 verified metabolites than CAMERA.

Furthermore, CliqueMS is able to correctly identify a larger number of adducts and mass fragments and annotates a larger number of features. We also note that while CliqueMS is able to correctly annotate metabolites that CAMERA does not identify, CAMERA does not annotate any metabolites not annotated by CliqueMS.

The differences in number of metabolites and features annotated are specially remarkable for the positive ionization sample, in which the number of features is larger and therefore more features can coelute. In this case CliqueMS is able to assign a parental mass to 66% of the features overall (and 57% if we only consider the top-ranked annotation), whereas CAMERA only assigns a parental mass to 43% of the features. In the negative ionization sample, the number features is much smaller and therefore the differences between both algorithms are not as stark.



a)

Sample	MS2 annotated	Algorithm	Annotated Metabolites	Adducts / Mass Fragments	Annotated Features
Retina IRS2 KO (+ ionization)	20	CliqueMS	15	50	97
		CAMERA	8	25	45
Retina IRS2 KO (- ionization)	18	CliqueMS	6	16	36
		CAMERA	5	14	33

b)

Metabolite	Annotation - Adduct (number of isotopes)		
	CliqueMS	Ranking	CAMERA
GABA	(M+H)+ (2) (M+Na)+ (1) (M+H+H2O)+ (1) (M+H-NH3)+ (1) (M-H+2Na)+ (1)	1 1 1 1 1	(M$_2$+NH4)+ (2) (M$_2$+H)+ (1)
Uracil	(M+H)+ (2) (M+H-H2O)+ (1) (M+H-NH3)+ (1)	1 1 1	(M+H)+ (2) (M+H-H2O)+ (1) (M+H-NH3)+ (1)
Taurine	(M+H)+ (2) (M+Na)+ (2) (M+H-H2O)+ (1) (2M+H)+ (1) (M$_2$+Na)+ (3)	2 3 2 2 1	(M+H)+ (2) (M+Na)+ (2) (M+H-H2O)+ (1) (2M+H)+ (1) (M-H+2Na)+ (3)
Adenine	(M+H)+ (2) (M+H-NH3)+ (2)	1 1	(M$_2$+NH4)+ (2) (M$_2$+H)+ (2)
L-Glutamic acid	(M+H)+(3) (M+H-H2O)+(2) (M+Na-H2O)+(1) (M+Na)+ (3) (M-H+2Na)+ (3) (M-2H+3Na)+ (2)	1 1 3 3 3 3	(M$_2$+NH4)+ (3) (M$_2$+H-H2O)+ (1) (M$_2$+H)+ (3) (M$_2$+Na)+ (3) (M$_2$-H+2Na)+ (3)
Guanine	(M+H-H2O)+ (1) (M+H-NH3)+ (2) (M+H)+ (2)	1 1 1	(M+H-H2O)+ (1) (M+H-NH3)+ (2) (M+H)+ (3)
Xanthine	(M+Na)+ (1) (M+H-NH3)+ (1) (M+H)+ (2)	1 1 1	
L-2-Aminoadipic acid	(M+H-H2O)+ (1) (M+H)+ (1)	2 2	*(M+H-H2O)+ (1) *(M+H)+ (1)
L-Ascorbic acid	(M+Na)+ (1) (M+H)+ (1)	1 1	(M+Na)+ (1) (M+H)+ (1)
PC	(M+K)+ (1) (M+Na)+ (2) (M+H)+ (3)	1 1 1	(M$_2$+K-H2O)+ (1) (M$_2$+Na-H2O)+ (2)
Inosine	(M+K)+ (2) (2M+H)+ (2) (2M+Na)+ (3) (M+H)+ (3) (M+Na)+ (2)	1 1 1 1 1	(M+K)+ (1) (2M+H)+ (3) (2M+Na)+ (3) (M+H)+ (3) (M+Na)+ (2)
Guanosine	(2M+H)+ (2) (M+Na)+ (1) (M+H)+ (4)	1 1 1	(2M+H)+ (2) (M+Na)+ (1) (M+H)+ (4)
Glutathione	(M+Na)+ (1) (M+H)+ (2) (M+H-H2O)+ (3)	1 1 1	(M+Na)+ (1) (M+H)+ (3) (2M$_2$+H)+ (3)
Oxigluthatione	(M+Na)+ (2) (M+K)+ (3) (M+H)+ (3)	1 1 1	(2M$_2$+Na)+ (2) (2M$_2$+K)+ (3) (2M$_2$+H)+ (4)
NAD	(M+Na)+ (1) (M+2H)2+ (3) (M+H)+ (4)	1 1 1	(2M$_2$+Na)+ (1) (M$_2$+H)+ (1) (2M$_2$+H)+ (4)

Figure 33. Feature annotation for complex samples. a) Summary of results of CliqueMS and CAMERA for complex samples with negative and positive ionization. We show the performance of CliqueMS and Camera relative to the metabolites annotated by MS2. The number of annotated features correspond to the number of adducts/mass fragments identified plus the number of annotated b) Detail of the adducts and mass fragments annotated by CliqueMS and CAMERA for the Retina IRS2 KO(+ ionization) sample. For each molecule we show the different adducts and mass fragments annotated; in parenthesis we show the total number of isotopic variants of that particular adduct/mass fragment. Correctly annotated features are shown in green; incorrectly annotated features are shown in red, with M2 indicating that the associated parental mass was incorrect; non-annotated features are shown in white. For CliqueMS we also show the ranking of the feature annotation that matches manual annotation. For CAMERA the * indicates those features for which the algorithm returned two possible annotations.

3.1.3 iMet: A Network-Based Computational Tool To Assist in the Annotation of Metabolites from Tandem Mass Spectra.

Introduction

Structural annotation of metabolites relies mainly on tandem mass spectrometry (MS/MS) analysis. However, approximately 90% of the known metabolites reported in metabolomic databases do not have annotated spectral data from standards. This situation has fostered the development of computational tools that predict fragmentation patterns *in silico* and compare these to experimental MS/MS spectra. However, because such methods require the molecular structure of the detected compound to be available for the algorithm, the identification of novel metabolites in organisms relevant for biotechnological and medical applications remains a challenge. Here we present iMet, a computational tool that facilitates structural annotation of metabolites not described in databases. iMet uses MS/MS spectra and the exact mass of an unknown metabolite to identify metabolites in a reference database that are structurally similar to the unknown metabolite. The algorithm also suggests the chemical transformation that converts the known metabolites into the unknown one. To validate iMet, we tested 31 metabolites proposed in the 2012-2016 CASMI challenges. iMet is freely available at http://imet.seeslab.net.

The CASMI challenge. To simulate a real scenario of metabolites not present in a database, we tested iMet using metabolites proposed in the Critical Assessment of Small Molecule Identification (CASMI) challenges from years 2012-2016. We downloaded the spectra of 31 different metabolites obtained using an ESI-QTOF mass spectrometer and that they had PubChem CID. The success of computational tools in the CASMI contest is to a large extent determined by the database used to retrieve molecular structures. Since iMet is designed to allow structural annotation of novel metabolites not present in databases, we tested the 33 metabolites in CASMI against our reference database of 5,060 molecular structures, which does not contain any of these CASMI metabolites, and compared iMet's performance to CFM-ID[149], MetFrag[274] and MS-Finder[275] using the same. iMet could not be compared to CSI:FingerID[276] because the latter is a web-based tool that only searches molecular structures in an already defined libraries (PubChem, HMDB, ChEBI and KNApSAck).

Briefly, in CFM-ID, a precursor ion is first matched against a compound database. The trained algorithm generates fragmentation spectrum from candidates and are matched against the experimental fragmentation pattern. In MetFrag candidate molecules of different databases are fragmented *in silico* and matched against mass to charge values. And finally, in MS-Finder molecular formulas of precursor ions are determined from accurate mass, isotope ratio, and product ion information. All isomer structures of the predicted formula are retrieved from metabolome databases and MS/MS fragmentations are predicted *in silico*.

Results

For the 31 metabolites in CASMI, 42% of the top candidates suggested by iMet were structurally similar to the test metabolite (Dice coefficient > 0.32), and in 48% of the results iMet was able to locate at least one structurally similar metabolite. In 48% of the cases, the top formula proposed by iMet was the correct formula of the test metabolite. The percentages of success dropped to 13-16% for the top candidates and to 19% for at least one of the top four candidates using CFM-ID, MetFrag and MS-Finder (A). Instead, when CFM-ID, MetFrag, MS-Finder and also CSI:FingerID were allowed to retrieve molecular structures from PubChem or HMDB, in general these tools performed better than iMet as expected (

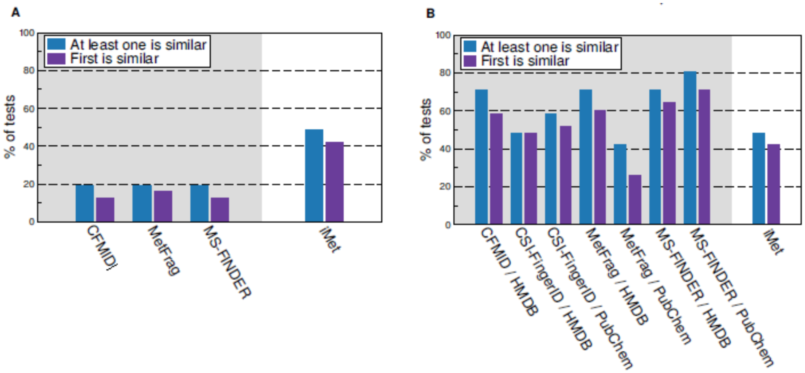


Figure 34B). These results demonstrate the potential of iMet for annotating metabolites that are note present in databases, and utility of existing methods to assist the structural annotation of known metabolites lacking MS/MS spectra in databases.

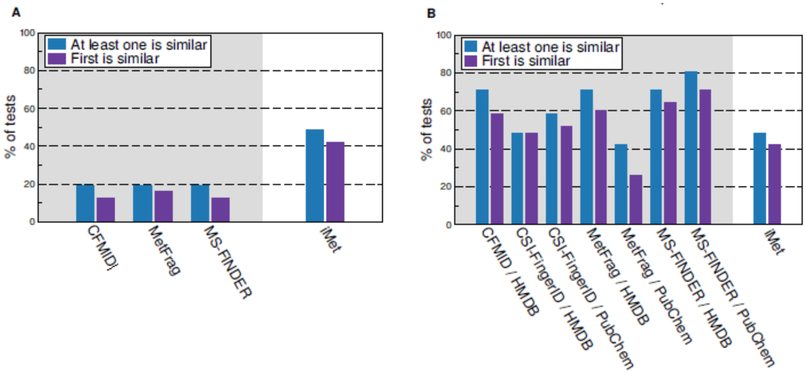


Figure 34. Comparison of methods testing 31 metabolites from the CASMI challenges 2012- 2016. Blue bars indicate the percentage of tests with at least one proposed metabolite amongst the top four candidates of each method with a Dice coefficient > 0.32 with the test metabolite. Violet bars depict the percentage of tests for which the top metabolite proposed by each method is structurally similar (Dice coefficient > 0.32) to the test metabolite. (A) With the same restricted database for all methods. (B) With different database.

Regarding iMet, it should be noted that 26 out of the 31 CASMI metabolites were obtained using other collision energies than those used in the training set (10, 20 or 40V). We tested them nevertheless to evaluate the performance of iMet when confronted with spectra obtained using inaccurate experimental data (for example, we introduced spectra obtained at 25V as if they were obtained at 20V; or 35V spectra as if they were 40V). For these 26 metabolites, 42% of the top candidates suggested by iMet were structurally similar to the test metabolite (Dice coefficient > 0.32), and in 50% of the results iMet was able to locate at least one structurally similar metabolite in the database. These results suggest that iMet does not decrease substantially its accuracy when using slightly different collision energies as inputs.

3.2 Alterations in the proteome and metabolome of human retinal pigment epithelium cells exposed to hyperglycemic and/or hypoxic conditions

3.2.1 Analysis of a protein-protein interaction network in retinal pigment epithelial cells exposed to hyperglycemic and hypoxic conditions

Introduction

Proteomics is defined as the large-scale characterization of the entire protein complement of a cell line, tissue, or organism. The proteome of a cell is dynamic and will reflect the immediate environment in which it is studied in response to internal or external stimuli, proteins can be modified by posttranslational modifications, undergo translocations within the cell, or be synthesized or degraded. Considering all the possibilities, it is likely that any given genome can potentially give rise to an infinite number of proteomes[277]. Proteomics generates complex datasets prompting the use of computational tools and network analysis in order to understand biological systems as a whole. Since 2000 the number of systems biology publications has increased linearly reaching almost 99500 manuscripts, however only 5 % of them involve proteomics studies. This fact indicates that analysing proteomics from a systems biology perspective is still growing and a lot of work needs to be done to continue exploring this field. Yet, most of the published studies on differential protein expression have focused on applying a significance test for each single protein testing whether this protein is differentially expressed or not [224,225]. However, studying each protein individually does not reflect the reality occurring within cells since proteins rarely act alone to perform their functions. It has been widely observed that proteins involved in the same cellular processes interact with each other since they belong to the same protein complex[226]. For overcoming such important issue, the application of systems biology to proteomics allows reaching the complexity found in biological networks by taking a holistic view of the cell [2, 228]. In this context, protein-protein interaction (PPI) network where nodes represent proteins and edges physical interactions between two proteins have been widely used in systems biology. Since the organization of a network in a complex system still remains unclear, an effective approach to achieve this goal is detecting topological clusters or modules that can be involved in particular cellular functions or disease [278, 229,230]. Additionally, gene expression information has been used in several studies to identify responsive PPI modules based on the significant changes of gene expression over a particular condition. In the same line, genes with correlated expression changes over many conditions are likely to be involved in similar functions or cellular processes. Thus, integrating PPI networks and gene-expression data generates a meaningful biological context in terms of functional association for differentially expressed genes [279,280,281]. However, not only studying genes but also proteins has shown that co-expression modules reflect significant enrichment for known PPI and biological function [282,283, 284,285,227].

Along these lines we present a novel approach that integrates PPI, module analysis and protein expression for detecting dysregulated groups of interacting proteins that participate in the same biological processes leading to a better description of the studied phenotype. Only in this way it has been possible to capture slight but consistent protein changes occurring in a protein module which are impossible to detect considering only individual proteins.

Here, the proposed method was applied to study the effect of hyperglycemic and hypoxic conditions on a quantitative proteomics analysis of human retinal pigment epithelial cell line (ARPE-19) in order to simulate a model of diabetic retinopathy, the most common complication of diabetes and the leading cause of blindness among working-aged adults around the world [234]. The global prevalence of DR among patients is approximately 35% and around one-tenth of them have vision-threatening states such as diabetic macular oedema (DMO) or proliferative diabetic retinopathy (PDR) [235]. Neovascularization due to severe hypoxia is the hallmark of PDR whereas vascular leakage due to the breakdown of the blood retinal barrier (BRB) is the main event involved in the pathogenesis of DMO [236,237]. Most of the research on the pathogenesis of DR has been focused on the impairment of the neuroretina and the breakdown of the inner BRB, while the effects of diabetes on the retinal pigment epithelium (RPE) have received less attention. RPE is a monolayer of pigmented cells situated between the neuroretina and choroids. RPE constitutes the outer BRB and is essential for neuroretina survival, and consequently, for visual function [238]. The specific functions of RPE are the following: i) transport of nutrients, ions, and water; ii) absorption of light and protection against photo-oxidation; iii) re-isomerization of all-trans-retinal into 11-cis-retinal, which is a key element of the visual cycle; iv) phagocytosis of shed photoreceptor membranes; and v) secretion of various essential factors for the structural integrity of the retina [238]. Therefore, the study of RPE is fundamental to gain new insights into the mechanisms that lead to DR and to identify new therapeutic targets for this devastating complication of diabetes.

The results demonstrate that our integrated method is able to efficiently capture dysregulated modules of proteins associated with specific biological processes leading to a better knowledge of the diabetic retinopathy phenotype.

Finally, untargeted proteomics data was acquired for human vitreous humor samples of patients at different stages of DR, which are 14 controls (healthy retinas), 4 non-proliferative DR (NPDR), an early stage and 12 proliferative DR (PDR). The obtained results were compared to the ones for the *in vitro* ARPE-19 cell cultures,

Materials and methods

ARPE-19 is a spontaneously immortalized human RPE cell line obtained from the American Type Culture Collection (Manassas, VA). D-Glucose or D-[U-13C]-Glucose (99 atom % 13C) were from Sigma (Madrid, Spain). Whitley H35 Hypoxystation from Nirco (Madrid, Spain). LC/MS grade methanol (MeOH) and acetonitrile (ACN) and analytical grade chloroform (CHCl3) were purchased from SDS (Peypin, France). Water was produced in an in-house Milli-Q purification system (Millipore, Molsheim, France). Formic acid, ammonium fluoride, N-methyl-N-trimethylsilyltrifluoroacetamide, methoxamine hydrochloride and pyridine were purchased from Sigma-Aldrich (Steinheim, Germany).

Myristic-d27 acid and succinic acid-2,2,3,3-d4 where from Isotec Stable Isotopes (Miamisburg, U.S.A.). A set of 13 even saturated fatty acid methyl esthers (FAMEs) from C8:0 to C30:0 were acquired from Sigma-Aldrich, NuChekPrep (Elysian, U.S.A.) and Molport (Riga, Latvia). Deuterated water (D2O) and 5-mm NMR tubes were purchased from Cortecnet (Viosins Le Bretonneux, France). DMEM/F-12 basal medium was purchased from Life Technologies. Sequencing grade modified trypsin V511A was purchased from Promega and Lys-C 125-05061 from Wako. For the quantitative proteomics experiments: the Complete Mini EDTA-free protease inhibitor and the PhosSTOP phosphatase inhibitor cocktails were from Roche (Almere, The Netherlands), the 6-plex TMT labeling kit was from Pierce (Rockford, Ilinois), and all other reagents were from Sigma (Steinheim, Germany).

Human samples. Vitreous humor samples were obtained from 28 voluntary patients that undergone eye surgery in Clínica Barraquer. The samples were extracting using vitrectomy followed by suction and then frozen at -80ºC and stored until sample preparation.

ARPE-19 cell culture

Cells were cultured under standard conditions in DMEM/F12 (1:1 mixture of Dulbecco's modified Eagle's medium and Ham's F12), 10% fetal calf serum (FCS) and penicillin/streptomicin. ARPE-19 cells from passage 20-23 were used and the media was changed every 3 days. Cells grown in these conditions constitute a monolayer that retains the functionality, polarity and tight junction expression of the human RPE. For our study, cells were seeded in Petri dishes (10 cm) at 0.4 x104 cells/mL and maintained in culture for 21 days with 5.5 mM or 25 mM of D-Glucose at 37°C under 5% (v/v) CO2 in an incubator. During the last 24 hours cells were subjected to serum deprivation (1% FCS). Serum deprived media were prepared with 5.5 mM or 25 mM of either D-Glucose or D-[U-13C]-Glucose, and cultured in normoxic or hypoxic (1% O2) conditions. Each condition was run in triplicate.

Quantitative proteomics ARPE-19 cells

Cell lysis and protein digestion. ARPE-19 cells were lysed in lysis buffer (50 mM ammonium bicarbonate, 8 M urea, 1 tablet Complete Mini EDTA-free protease inhibitor cocktail, 1 tablet PhosSTOP phosphatase inhibitor cocktail). Lysis was performed by gentle sonication on ice at 20% amplitude, with a 0.5 cycle in a Sonics Vibracell (Bioblock Scientific, France). Cell debris were removed by centrifugation at 20,000 g for 10 min at 4°C. Protein concentration was determined by an RC-DC protein assay (Bio-Rad). Proteins were reduced in 4 mM dithiothreitol (30 min at 56°C) and alkylated in 8 mM iodoacetamide (30 min at room temperature in the dark). LysC was added at an enzyme:protein ratio of 1:75 (w/w) and incubated for 4 h at 37°C. Samples were then diluted 4 times with 50 mM ammonium bicarbonate. Trypsin was added at an enzyme:protein ratio of 1:100 (w/w) and incubated overnight at 37°C. Acetic acid was added to a final concentration of 10% and samples were immediately frozen.

TMT labeling. 100 μg of each sample were desalted and concentrated using C18 solid phase extraction (Sep-Pak Vac C18 cartridge 1 cm3/200 mg, Waters), dried in vacuum and reconstituted in 120 µL of 200 mM triethylammonium bicarbonate (Sigma). Labeling was performed with the 6-plex labeling kit according to the manufacturer’s protocol. Briefly, each labeling was carried out for 1 h at room temperature and quenched with 8 µL of 5 % hydroxylamine. The four channels were mixed, dried in vacuum and resuspended in 10% formic acid.

Strong cation exchange fractionation. Peptides were fractionated by strong cation exchange (SCX) using a Zorbax BioSCX-Series II column (0.8 mm × 50 mm, 3.5 µm), as described Munoz et al. [286]. Solvent A consisted of 0.05% formic acid in 20% acetonitrile, solvent B of 0.05% formic acid, 0.5 M NaCl in 20% acetonitrile. The gradient was 0 to 2% B in 0.01 min; 2 to 3% B in 8 min; 3 to 8% B in 6 min; 8 to 20% B in 14 min; 20 to 40% B in 10 min; 40 to 90% B in 10 min; 90%B for 6 min; 90 to 0% B in 6 min. Fractions were collected once a minute and each dried in vacuum and stored at -20°C.

LC/MS analyses. Mass spectrometry. SCX fractions were analyzed on an Orbitrap Q-Exactive (Thermo Fisher Scientific) connected to an UHPLC Proxeon Easy-nLC 1000 (Thermo Scientific). Peptides were trapped on a double-fritted trap column (Dr. Maisch Reprosil C18, 3 µm, 2 cm × 100 µm) and separated on an analytical column (Agilent Zorbax SB-C18, 1.8 µm, 40 cm × 75 µm), as previously described by Cristobal et al. [287]. Solvent A consisted of 0.1 M acetic acid, solvent B of 0.1 M acetic acid in 80% acetonitrile. Samples were loaded at a pressure of 800 bar with 100% solvent A. Peptides were separated by a 110 min gradient from 10% to 40% solvent B at a flow rate of 150 nL/min. Full scan MS spectra were acquired in the Orbitrap (350-1500 m/z, resolution 35,000, AGC target 3e6, maximum injection time 250 ms). The 20 most intense precursors were selected for HCD fragmentation (isolation window 1.2 Da, resolution 17,500, AGC target 5e4, maximum injection time 120 ms, first m/z 100, NCE 33%, dynamic exclusion 60 s). *Data preprocessing*. Raw data was analyzed with MaxQuant[288] (version 1.3.0.5). MS/MS peak lists were generated and searched with Andromeda against the Swissprot human database. Trypsin/P was chosen as an enzyme, with a maximum of 2 missed cleavages. Methionine oxidation was set as variable modification. Cysteine carbamidomethylation as fixed modifications, TMT6plex (Lys) and TMT6plex (N-term) as the reporter ion quantification method. The database search was performed with a precursor tolerance of 6 ppm for the main search (20 ppm for the first search) and a fragment mass tolerance of 0.05 Da. Match between run was enabled with a time window of 2 min. Peptide and protein FDR were set at 1%, and peptide score threshold at 60. The quantification and univariate statistical processing was performed in Perseus v.1.3.8.1. Proteins were grouped and reporter ion intensities were calculated for each of the TMT channels. Ratios were calculated and normalized on median.

Classical approach: statistical analysis and functional enrichment

A significance B test was performed to determine significantly regulated proteins followed by a Benjamini false discovery rate correction.The truncation was performed using p values, with a threshold value of 0.05. Dysregulated proteins for each experimental condition (N25/N5, H5/N5 and H25/N5) were functionally enriched using g:Profiler (gProfilerR_0.6.1) [289]. The analysis was focused on biological process and reactome pathway enrichment. g:Profiler used cumulative hypergeometric P-values to identify the most significant terms corresponding to the input set of genes. Subsequently, the p-value correction was performed using g:SCS (Set Counts and Sizes), a novel method specially developed to estimate thresholds in complex and structured functional profiling data such as GO or pathways[289]. Electronic annotations were excluded to focus only on annotation with stronger evidence. Quantified experimental proteins were used as background set for more accurate statistics.

Network approach: statistical analysis and functional enrichment

- Human interactome mapping

Complex cellular systems formed by interactions among genes and gene products, or interactome networks, appear to underlie most cellular functions. In this study a systematic map of around 14,000 high-quality human binary has been used to map 2925 (89.7 %) ARPE-19 experimental proteins [290].

- Fold change (FC) behaviour across distances in the interactome network

Having the experimental proteins mapped in the interactome, we studied if the proteins showed a coordinated expression behaviour at shorter PPI network distances. To perform this aim, the absolute fold change (FC) difference was calculated for each pair of proteins at different distances (from 1 to 6). Afterwards, the mean FC difference was computed for each distance.

- Clustering method

A stochastic block models (SBM)[291,292] method has been applied ir order to find protein complexes. This method is based on clustering proteins with similar patterns of interactions with others, in other words, nodes (proteins) are assumed to belong to groups and connect to each other with probabilities that depend only on their group memberships

- Protein module functional enrichment

The functional enrichment method was performed using gprofiler following the same procedure explained above but this time studying the enrichment in each protein module. In this case, moderate hierarchical filtering was applied, in this way, for every topmost GO term in a particular group, only the sibling term with the strongest p-value was selected.

- Module statistical analysis (calculate module significance using permutation testing)

Statistical analysis has been applied to each set of proteins (a module) comparing the FC distribution of proteins belonging to each module against the distribution of all detected proteins.

In addition, statistical testing has been applied taking into account the functional enrichment result for each module to these different situations:

Module-enriched term: compare the FC distribution of proteins associated to an enriched functional term in a module against the FC distribution of all permutated detected proteins.

Overall enriched term: compare the FC distribution of all proteins associated to an enriched functional term against the FC distribution of all detected proteins.

The statistical test for all the comparisons above have been performed by counting the number of times the FC mean divided by the standard deviation (SD) of one group of proteins belonging to a module and/or to a enriched biological term is bigger or smaller than 100,000 random values of FC mean divided by the SD of detected proteins (keeping the same size). Afterwards, the p-value was calculated using the following equation:

$$p - value = \frac{100{,}000 - counts}{100{,}000}$$

Subsequently, a Bonferroni p-value correction has been applied in all comparisons, modifying the p-value threshold to correct for multiple testing in each test. In the case of module test, the new p-value threshold has been calculated dividing 0.05 by the number of detected modules. For the rest of the analysis that consider enriched functional terms, which are not completely independent between them because common proteins can be involved in different functional terms, a different calculation has been applied. To address this issue, the Jaccard Index (JI) has been used to quantify the overlapping proteins percentage for each functional term pair. All functional terms having a JI between them bigger than 50 % are considered one independent variable. Finally, the new p-value threshold has been obtained dividing 0.05 by the number of independent enriched terms.

Quantitative proteomics vitreous humor

Cell lysis and protein digestion. Vitreous humor samples had been analysed before by NMR and they were treated specially following the next method. Vitreous humor was centrifuged for 15 min at 15000 rpm and 4 °C. 300 uL of supernatant were placed in a new eppendorf tube. 250 µL of 0.73 mM TSP buffer in D2O were added to the eppendorf tubes containing the vitreous humor. Samples were then vortexed, and centrifuged for 15 min at 15000 × rpm and 4 °C. Finally, redissolved samples were placed into 5 mm NMR tubes. After the NMR analysis, each sample was split in two aliquots of 250 uL, freeze-dried and kept at -80ºC until be used for proteomics analysis. The freeze-dried vitreous humor aliquots were lysed in 1 mL of lysis buffer (8 M urea, 100 mM triethylammonium bicarbonate (TEAB)). The lysate was placed in a bench centrifuge at 4ºC, 15000 rpm and 15 minutes and was kept on ice. Protein concentration was determined by Bradford assay. 10 µg of proteins were taken from supernatant. Reduction buffer Bond Breaker was added to the extracted supernatant to make a final concentration of 10 mM of reduction buffer. The sample was placed in shaker for 30 minutes. Alkylation buffer 2-Chloroacetamide was added to achieve a final concentration of 200 mM. The solution was vortexed and left 30 minutes at room temperature in the dark. Lys.C solution was added in 1:100 (w/w) ratio of Lys-C to protein. The sample was placed in a heater-shaker at 37ºC for 4 hours. Then, a brief spin was performed to remove material from lid. The sample was diluted with 50 mM TEAB to achieve less than 2 M urea. Trypsin was added in a 1:20 (w/w) ratio. The digest solution was placed in a shaker at 37 ºC overnight. After that, the digestion was stopped by adding an equal volume of stop buffer 10% formic acid to the sample. The sample was concentrated and cleaned using C18 filters. Firstly, the filter was dysplayed over a new eppendorf and was wetted with 500 µL 100 % ACN two times. 250 µL of 0.1 % TFA were passed two times. The C18 filter was dysplayed over a new eppendorf and the sample was introduced in the filter. The sample was washed with 1 mL 0.1 % TFA two times and was eluted in a new eppendorf following the next procedure: adding 200 µL (50 % ACN, 0.1 % TFA), centrifuging, adding 200 µL (80 % ACN, 0.1 % TFA) and centrifuging. The sample elution was dried in the speed vac and resuspended in 30 µL of 5 % DMSO and 5 % formic acid.

LC/MS analyses. 2 µL of sample was analyzed on an Orbitrap Q-Exactive (Thermo Fisher Scientific) connected to an UHPLC Proxeon Easy-nLC 1000 (Thermo Scientific). Peptides were trapped on trap column (Dr. Maisch Reprosil C18, 3 µm, 2 cm × 100 µm) and separated on an analytical column (Agilent Zorbax SB-C18, 1.8 µm, 40 cm × 75 µm). Solvent A consisted of 0.1 % formic acid, solvent B of 0.1 % formic acid in 80 % ACN. The gradient started at 7 % B (0 - 5 min), then continued at 7 - 28 % B (5 - 125 min), 28 - 100 % B (125 - 126 min), 100 % B (126 - 128 min), 100 - 7 % B

(128 - 129 min) and finished at 7 % B (129 - 140 min). The flow rate was 200 nL/min. Full scan MS spectra were acquired in the Orbitrap (350 - 1500 m/z, resolution 70,000, AGC target 3e6, maximum injection time 50 ms). The 20 most intense precursors were selected for HCD fragmentation (isolation window 1.5 m/z, resolution 17,500, AGC 5e4, maximum injection time 120 ms, first m/z 180, NCE 30 %, dynamic exclusion 40 s).

Data preprocessing. Raw data was analyzed with MaxQuant (version 1.5.3.12). MS/MS peak lists were generated and searched with Andromeda against the Swissprot human database. Trypsin/P was choosen as an enzyme, with a maximum of 2 missed cleavages.

Acetyl (Protein N-term) and methionine oxidation were set as variable modification. Cysteine carbamidomethylation as fixed modifications. Match between run was enabled with a time window of 0.7 min. Peptide and protein FDR were set at 1 %. Wilcoxon comparison was performed to determine significantly regulated proteins with a threshold p value of 0.05 and absolute FC ratio bigger than 1.5. The statistical processing was performed in R (R-3.4.1).

Classical approach: statistical analysis and functional enrichment

Wilcoxon comparison was performed to determine significantly regulated proteins with a threshold p value of 0.05 and absolute FC ratio bigger than 1.5. The statistical processing was performed in R (R-3.4.1). Dysregulated proteins for each experimental condition

(NPDR/CTRL and PDR/CTRL) were functionally enriched using the same procedure than in ARPE-19 excepting no background set was used since the number of quantified experimental proteins were too low (using background set, gprofiler did not retrieve any enrichment)

Results

Quantitative proteomics ARPE-19 cells

Classical approach: statistical analysis and functional enrichment

A total of 3259 proteins were detected and quantified. In N25/N5 131 proteins were significantly regulated, 60 showed down-regulation and 71 up-regulation. In H5/N5 43 proteins showed dysregulation, 9 were down-regulated and 34 up-regulated, finally, in H25/N5, 184 proteins were significantly regulated which 96 were down-regulated and 88 up-regulated. A volcano plot can be seen in Figure S 1 from supplementary material.

Only hyperglycemic conditions: N25 and H25 showed biological process and reactome pathway enrichment. H25 condition reported 50 enriched biological processes while N25 41, finally, 37 processes were enriched in both conditions. Some of them were metabolic processes involved in ATP (mean log2 N25: -0.56; mean log2 H25: -0.69), nucleoside (mean log2 N25: -0.57; mean log2 H25: -0.68) and nucleotide (mean log2 N25: -0.46; mean log2 H25: -0.62) and oxidation-reduction process (mean log2 N25: -0.37; mean log2 H25:

-0.53), all of them showing overall down-regulation. On the other hand, there were some processes that happened only in one condition such as tricarboxylic acid cycle (mean log2 FC: -0.61) and mitochondrial translational elongation (mean log2 FC: -0.65) and termination (mean log2 FC: -0.66) that were enriched in H25 showing down-regulation. N25 showed a smaller number of unique enriched processes, between them there were positive regulation of cell adhesion (mean log2 FC: 0.11) and interferon-gamma production (mean log2 FC: 0.19).

Network approach: statistical analysis and functional enrichment

- Human interactome coverage.

89.7 % (2925) of detected and quantified proteins were mapped to the binary human interactome with a total of 7912 edges.

- FC behaviour across distances in the interactome network

The FC behaviour across distances was explored from one until five edges separating two nodes. We observed a better coordinated FC behaviour in smaller interactome distances or in other words, protein expression is more similar if they are close in the PPI network as it can be seen in Figure 35:

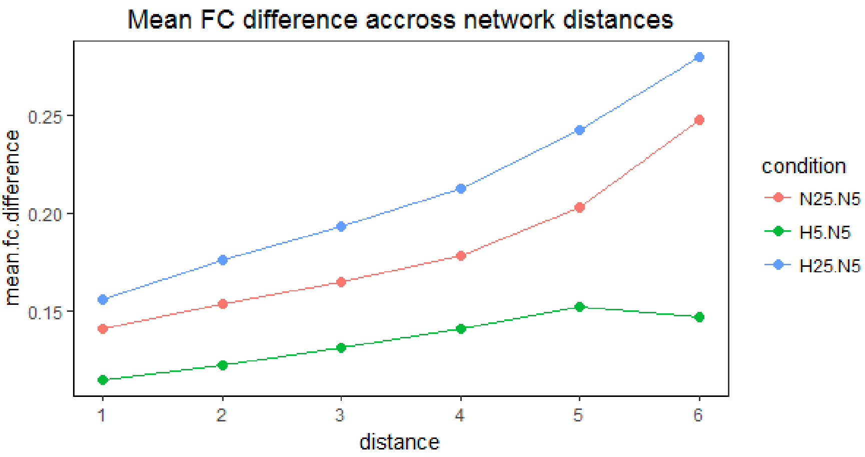


Figure 35. FC behaviour across distances in the interactome network for each pathological-like conditions (N25, H5 and H25).

Moreover, introducing the experimental condition factor, one can observe how there is a difference in the mean FC difference magnitude, in this way, H5 condition presents slight changes, followed by N25 and finally, H25 is the one encompassing the biggest FC alterations. We validated these results to prove they were not obtained by chance. For that purpose we applied the same study to randomized data. In this way, for each distance we randomized 1,000 times the proteins keeping the same number of protein pairs and calculated the mean FC difference. With the randomized data we did not observe the same effect since there was no change across distances as it can be seen in

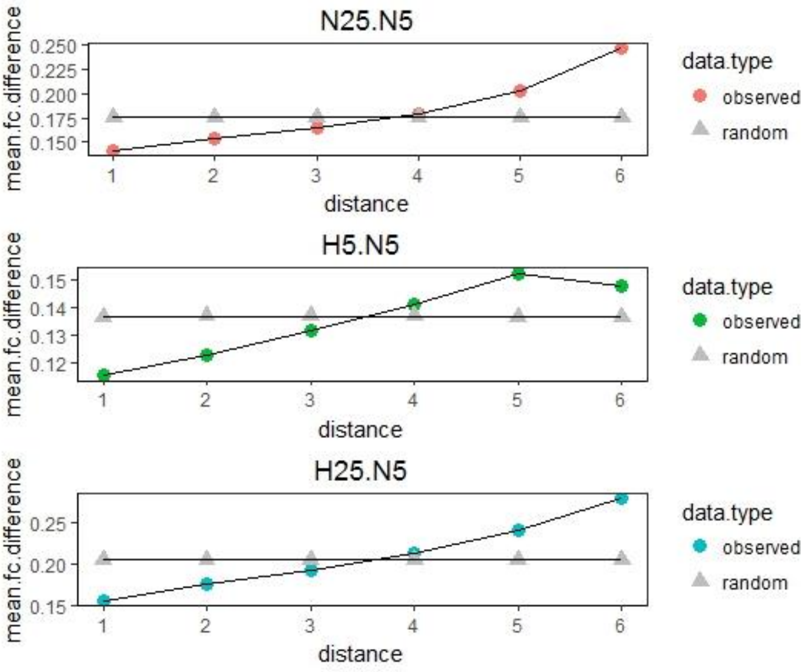


Figure S 2 from supplementary material.

- Clustering analysis: SBM

A total of 20 protein groups (modules) were retrieved from SBM analysis. The module size range covers from 6 to 129 proteins, module 1, containing 1663 proteins was discarded. Thus, for the following analysis 19 proteins groups will be considered. Mean size: 40.5, median size: 32.0, sd: 30.6.

- Protein module functional enrichment

Each module was functionally enriched in order to find which biological processes and/or pathways were associated with them, achieving in this way a better functional characterization.

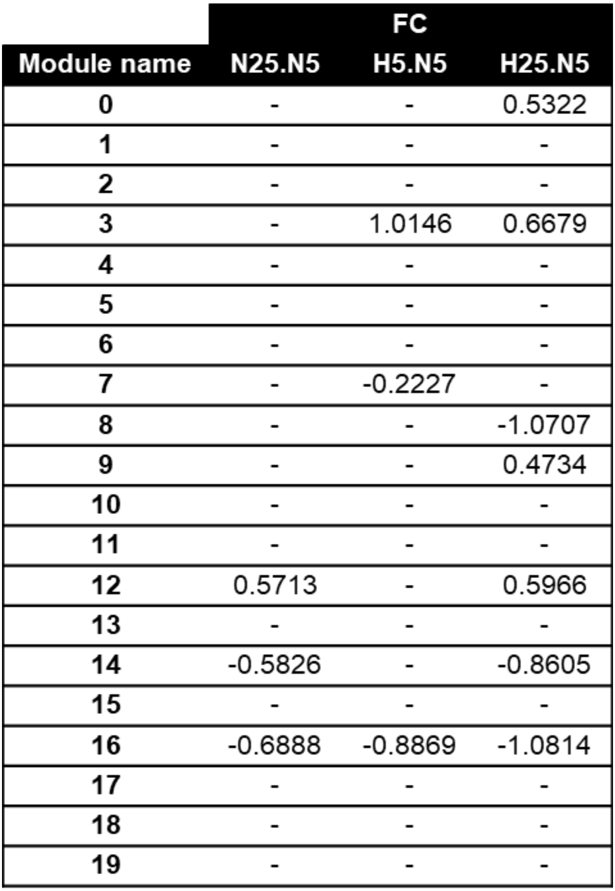
Biological process results: all modules excepting 13 and 15 are enriched with some biological process.

Pathway reactome results: fourteen of nineteen modules are enriched with some reactome pathway.

- Statistical analysis (calculate module significance using permutation testing)

Module test: the statistical test explained above was applied to each module in order to explore the dysregulation of them. From nineteen detected modules, a total number of three modules showed dysregulation in N25, three in H5 and seven in H25, as it can be seen in Table 10. More specifically, module 16 was down-regulated in the three experimental conditions, being in H25 the lowest value. Modules 12 and 14 were up and down-regulated, respectively only in hyperglycemic conditions, being H25 again, the most extreme condition in both modules. Module 3 was up-regulated in hypoxic conditions, but this time, H5 had the strongest value. Only module 7 was down-regulated in H5, while modules 0 and 9 were up-regulated and module 8 was down-regulated, only in H25.

Table 10. Statistical analysis results applied on SMB modules.



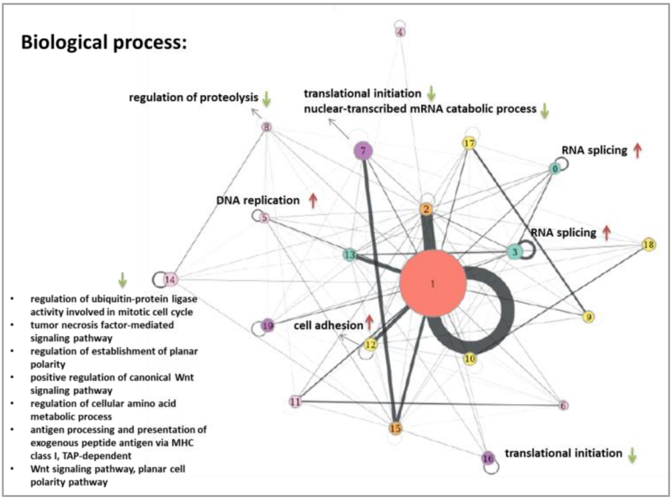
	FC		
Module name	N25.N5	H5.N5	H25.N5
0	-	-	0.5322
1	-	-	-
2	-	-	-
3	-	1.0146	0.6679
4	-	-	-
5	-	-	-
6	-	-	-
7	-	-0.2227	-
8	-	-	-1.0707
9	-	-	0.4734
10	-	-	-
11	-	-	-
12	0.5713	-	0.5966
13	-	-	-
14	-0.5826	-	-0.8605
15	-	-	-
16	-0.6888	-0.8869	-1.0814
17	-	-	-
18	-	-	-
19	-	-	-

	FC		
Module name	N25.N5	H5.N5	H25.N5
0	-	-	0.5322
1	-	-	-
2	-	-	-

3	-	1.0146	0.6679
4	-	-	-
5	-	-	-
6	-	-	-
7	-	-0.2227	-
8	-	-	-1.0707
9	-	-	0.4734
10	-	-	-
11	-	-	-
12	0.5713	-	0.5966
13	-	-	-
14	-0.5826	-	-0.8605
15	-	-	-
16	-0.6888	-0.8869	-1.0814
17	-	-	-
18	-	-	-
19	-	-	-

Considering the functional enrichment result for each module, we explored if the proteins belonging to specific enriched terms and modules were presenting dysregulation. Here, we applied the statistical test to different data as it has been explained in method section and the results were the following:

Module-enriched term: in this analysis we can observe which enriched biological processes and pathways associated with modules showed dysregulation. Focusing on biological processes, H25 condition, having fifteen dysregulated enriched terms, is the one that embraces the maximum number of dysregulated biological processes in modules (see Figure 36), followed by N25 with nine and finally, H5 with a total number of seven. Translational initiation in module 16 is the unique term that is down-regulated in the three conditions. Only happening in hyperglycemic conditions there are seven down-regulated and one up-regulated terms. Among the down-regulated ones, all of them occurring in module 14, there are some of the processes involved in the ubiquitin protein ligase activity, signal transduction, developmental process and cellular amino acid metabolic process while the up-regulated process was uniquely associated to DNA replication in module 5. Hypoxic conditions were characterized by up-regulation of RNA-splicing in modules 0 and 3. Finally, only detected in H25, there were three down-regulated enriched processes spread between the modules 7 and 8 that consist of RNA catabolic


Biological process:
regulation of proteolysis ↓
translational initiation ↓
nuclear-transcribed mRNA catabolic process ↓
RNA splicing ↑
DNA replication ↑
RNA splicing ↑
↓
• regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle
• tumor necrosis factor-mediated signaling pathway
• regulation of establishment of planar polarity
• positive regulation of canonical Wnt signaling pathway
• regulation of cellular amino acid metabolic process
• antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent
• Wnt signaling pathway, planar cell polarity pathway
cell adhesion ↑
translational initiation ↓
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

process, translational initiation and regulation of proteolysis and one up-regulated process involving cell adhesion in module 12. In the case of reactome pathways, again, H25 is the condition characterized by the maximum number of dysregulated pathways in modules, having nineteen, then, N25 with fourteen and H5 having four. Since dysregulated pathways highly overlap with biological processes results, we only have focused on explaining the new complementary results from pathways. It has been observed in hyperglycemic conditions a down-regulated of ABC-family proteins mediated transport in module 14.

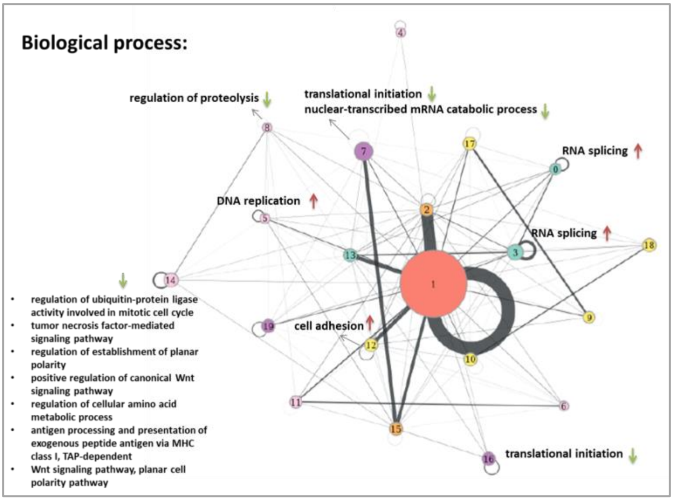

Biological process:
regulation of proteolysis
translational initiation
nuclear-transcribed mRNA catabolic process
RNA splicing
DNA replication
RNA splicing
cell adhesion
translational initiation
regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle
tumor necrosis factor-mediated signaling pathway
regulation of establishment of planar polarity
positive regulation of canonical Wnt signaling pathway
regulation of cellular amino acid metabolic process
antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent
Wnt signaling pathway, planar cell polarity pathway
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

Figure 36. Dysregulated enriched terms associated with modules in H25 versus N5 comparison.

Overall enriched term: in this analysis we studied if the dysregulated biological processes and pathways in protein modules were already altered between all the quantified proteins associated to that term, without considering the network topology. Around 53% of dysregulated enriched biological processes in modules in H25 were already dysregulated overall. In the case of N25, only 11% were altered and finally, in H5 100% were dysregulated overall.

- Comparing classical versus network approach results

In general, the results obtained from both approaches are different and they can be complementary in some cases.

In general, classical approach in hyperglycemic conditions resulted in a down-regulation of biological processes and pathways related to ATP metabolic activity while network approach was more focused on processes involved in protein synthesis.

One point in common in hyperglycemic conditions that is linked to DNA process is the down-regulation of nucleotides and nucleosides observed in classical approach and the up-regulation of DNA replication found in module 5 from network approach, however, the proteins involved in these processes from classical and network approach are completely different. Another biological processes in common but detected in different conditions is cell adhesion, in the case of classical approach positive regulation of cell adhesion appeared up-regulated in N25 while cell-adhesion appeared up-regulated in module 12 in H25 condition from network approach. Again, both methods did not share any common protein.

Quantitative proteomics vitreous humor validation

- Classical approach: statistical analysis and functional enrichment

A total of 371 proteins were detected and quantified in vitreous humor. In NPDR/CTRL 13 proteins were dysregulated, 2 of them were down-regulated and 11 were up-regulated, while in PDR/CTRL, 199 proteins showed dysregulation, which a total number of 142 and 57 were down and up-regulated respectively. A volcano plot from Figure S 3 (supplementary material) summarizes these results.

- Clinical validation in ARPE 19 in vitro model

After applying functional enrichment to the altered proteins in vitreous cell adhesion was the only common process found dysregulated between the comparison PDR/CONTROL from vitreous humor and the protein module 12 from ARPE-19. While in vitreous humor, the overall mean log2 FC protein appeared down-regulated (mean log2 FC: -1.03), in ARPE-19 was up-regulated (mean log2 FC: 0.20). Two proteins, P02751 and P19022, appeared in both approaches, although the total number of proteins defining cell adhesion was different, in vitreous humor there were forty while in arpe there were twenty one proteins.

Supplementary material

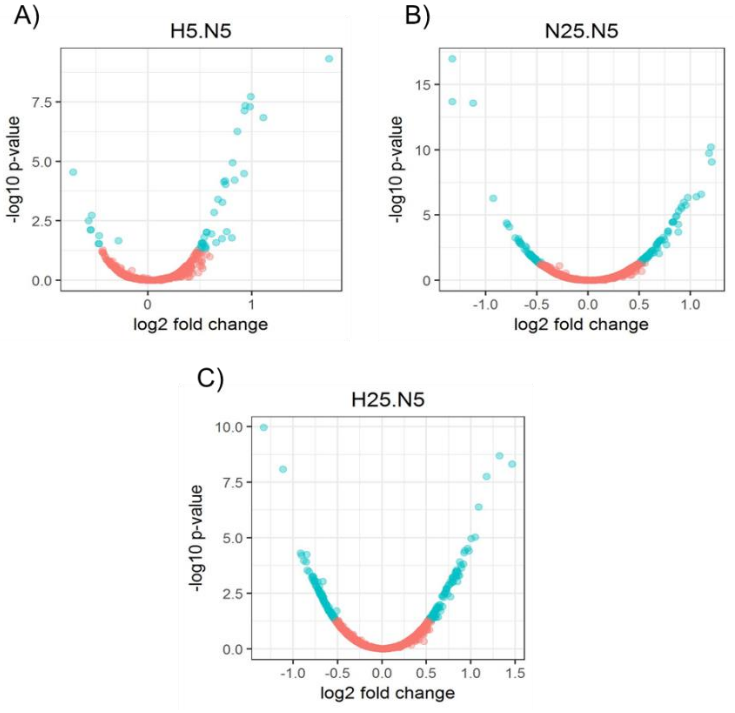


Figure S 1. Volcano plots from comparing the protein levels of physiological-like (N5) and any of the pathological-like conditions (N25, H5 and H25) in ARPE-19 cells. A) H5 versus N5, B) N25 versus N5 and C) H25 versus N5.

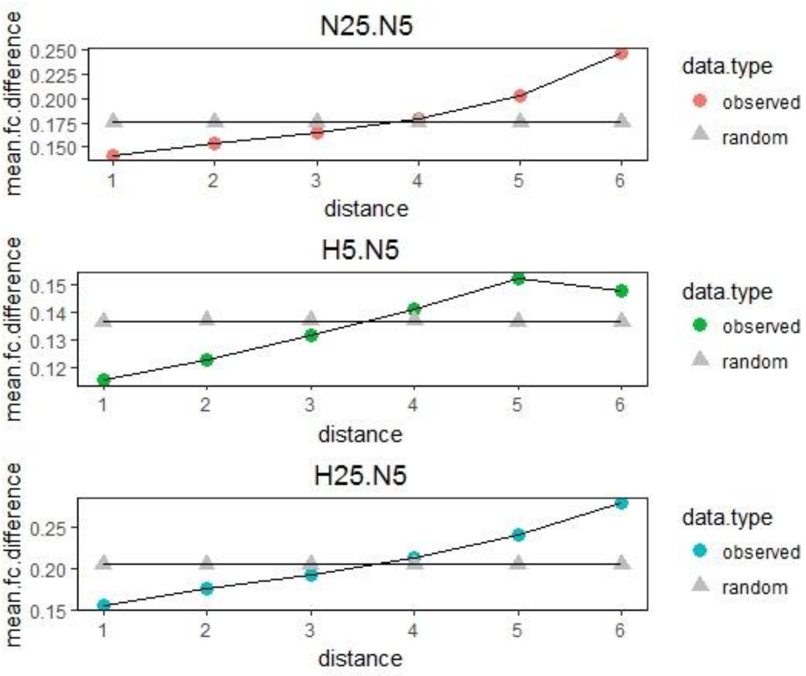


Figure S 2. FC behaviour across distances in the interactome network for each pathological-like conditions (N25, H5 and H25) compared to random data.

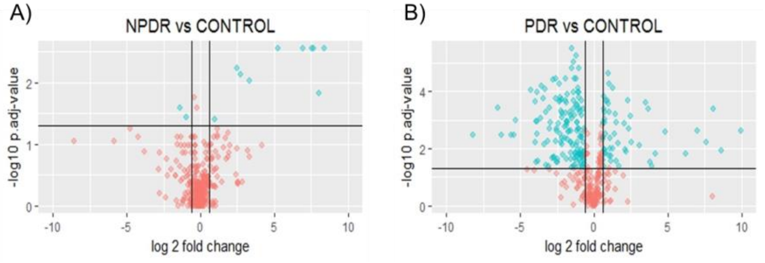


Figure S 3. Volcano plots from comparing the protein levels of control patients to A) non-proliferative diabetic retinopathy (NPDR) and B) proliferative diabetic retinopathy (PDR).

3.2.2 The connectivity of the human metabolic network reveals altered metabolites in ARPE-19 cells and vitreous humor of diabetic retinopathy patients

Introduction

Proteins are an enormous family of different structural entities encompassing wide functional diversity. Protein expression is essential for cell development and survival and it is regulated by multiple mechanisms, at the gene and protein level, including chromatin relaxation, alternative splicing or even iRNA. Proteins have different roles in metabolism, as enzymes they control metabolite concentration, and other proteins mediate their transport or regulate metabolic processes and signalling cascades. Defects on protein expression levels have been thoroughly reported as the cause of multiple diseases and have become the spotlight for many in the molecular biology fields[293,294,295,296].

Proteomics endeavours to study protein abundances, their modifications and how they interact, in order to understand cellular processes. This technology has been evolving for the last decades, improving its sensibility and developing better analysis methods, which permitted increasing the range of proteins detectable in proteomics experiments[297]. Similarly, metabolomics attempts to accurately measure the abundance of multiple metabolites in biological samples[113]. Evaluation of protein expression has been widely used to study metabolism, not only since enzymes are responsible for metabolic activity but also proteins transport metabolites and regulate metabolic processes. Together, these omics sciences can provide profound knowledge on the metabolic remodelling cells experience.

Mathematical and design models are used for simulating reality and run trial and error experiments in a quick manner, simulations require a model that contains the rules of how the elements of the system mathematically relate between them. A well-known example of a metabolic model on the biology is Recon 2[298], an expansion of Recon 1, the earliest comprehensive metabolic reconstruction built by a community of experts in the computational modelling area. Recon models were created to provide computational scientists with a peer-reviewed resource to elucidate and understand genotype–phenotype relationships in metabolism. Recon2 contains itemized information of metabolites and reactions, annotated with gene associations and organelle localization. It is possible to extend it with further information on gene-protein linkage from other databases that currently exist, a well-known example is KEGG[299].

Even if it is known that protein expression and enzyme regulation are harmonized processes in metabolic pathways, this knowledge is still not systematically implemented as a framework to interrogate metabolism, on the contrary, proteins are statistically treated as if they functioned independently. A few works has been published on this matter[300,301,302,302] representing the relations between genes or proteins as networks, or graphs, but in non-hypothesis-driven studies it is yet to be applied.

A network, or graph, is a node-link representation of relationships between a limited set of elements. Using the connectivity between metabolites based on the reactions in which they take part, a network can be built. Analyzing the topology of networks for amplifying the knowledge or enhancing the interpretation of quantification results has been used on transcriptomic and proteomic data. A number of network structures are possible, determined by how nodes and links are organized, and also a great collection of mathematical methods and algorithms can be used for exploration, seek of node clusters, etc., which depend on the input given and the desired output. In relation to metabolism, most networks are based on KEGG[299] and REACTOME[303] databases, among others, which use gene expression, of all genes or differentially expressed only, as an input to perform the network analysis.

Herein, a data analysis workflow that employs the connectivity of a metabolic network, based on Recon 2 structure in this case, is presented to unbiasedly test the metabolites within the network by evaluating the proteins associated to their transformation as a whole, instead of separately. This approach aims to better estimate the real effect of enzymatic proteins expression on cells metabolic phenotype as it contemplates both the synthesis and consumption reactions altogether. To develop and validate this novel analysis we obtained both proteomics and metabolomics data to study diabetic retinopathy, a common long-term diabetes complication occurring to type II diabetes patients that leads to vision impairment or blindness. To do so, we used two sets of samples: ARPE-19 cells, an *in vitro* cellular model of the retinal pigmentary epithelium; and human vitreous humor samples from 28 individuals at different stages of DR phenotype/disease (13 non-diabetic patients, 4 non-proliferative diabetic patients and 11 proliferative diabetic patients).

Diabetic retinopathy (DR) is the most common complication of diabetes and one of the leading causes of preventable blindness[304].

DR prevalence in the diabetic population is approximately 30%, and around one-tenth of them have vision-threatening states such as diabetic macular oedema (DMO) or proliferative diabetic retinopathy (PDR)[305]. Neovascularization due to severe hypoxia is the hallmark of PDR whereas vascular leakage due to the breakdown of the blood retinal barrier (BRB) is the main event involved in the pathogenesis of DMO[237]. Most of the research on the pathogenesis of DR has been focused on the impairment of the neuroretina and the breakdown of the inner BRB. However, the effects of diabetes on the retinal pigment epithelium (RPE) have received less attention.

RPE is a monolayer of pigmented cells situated between the neuroretina and choroids. RPE constitutes the outer BRB and is essential for neuroretina survival, and consequently, for visual function[306]. The specific functions of RPE are the following: i) transport of nutrients, ions, and water; ii) absorption of light and protection against photo-oxidation; iii) re-isomerization of all-trans-retinal into 11-cis-retinal, which is a key element of the visual cycle; iv) phagocytosis of shed photoreceptor membranes; and v) secretion of various essential factors for the structural integrity of the retina [306]. Therefore, the study of RPE is fundamental to gain new insights into the mechanisms that lead to DR and to identify new therapeutic targets for this devastating complication of diabetes.

With the exception of two proteomic studies investigating the effects of high glucose on ARPE-19 cells, at both intracellular[307] and secretory level[308], none of them have addressed the effect of hypoxia on protein expression. The few studies that investigated DR using metabolomics were exclusively based on NMR technology[259,309], and none of them examined RPE cells.

Our study, therefore, represents the first attempt to studying DR from a multi-omic and multi-platform point of view, including the implementation of a novel data analysis workflow based on the topology of a network built using the Recon 2 model.

Material and methods

Materials. ARPE-19 is a spontaneously immortalized human RPE cell line obtained from the American Type Culture Collection (Manassas, VA). D-Glucose or D-[U-13C]-Glucose (99 atom % 13C) were from Sigma (Madrid, Spain). Whitley H35 Hypoxystation from Nirco (Madrid, Spain). LC/MS grade methanol (MeOH) and acetonitrile (ACN) and analytical grade chloroform (CHCl3) were purchased from SDS (Peypin, France). Water was produced in an in-house Milli-Q purification system (Millipore, Molsheim, France). Formic acid, ammonium fluoride, N-methyl-N-trimethylsilyltrifluoroacetamide, methoxamine hydrochloride and pyridine were purchased from Sigma-Aldrich (Steinheim, Germany). Myristic-d27 acid and succinic acid-2,2,3,3-d4 where from Isotec Stable Isotopes (Miamisburg, U.S.A.). A set of 13 even saturated fatty acid methyl esthers (FAMEs) from C8:0 to C30:0 were acquired from Sigma-Aldrich, NuChekPrep (Elysian, U.S.A.) and Molport (Riga, Latvia). Deuterated water (D2O) and 5-mm NMR tubes were purchased from Cortecnet (Viosins Le Bretonneux, France). DMEM/F-12 basal medium was purchased from Life Technologies. Sequencing grade modified trypsin V511A was purchased from Promega and Lys-C 125-05061 from Wako. For the quantitative proteomics experiments: the Complete Mini EDTA-free protease inhibitor and the PhosSTOP phosphatase inhibitor cocktails were from Roche (Almere, The Netherlands), the 6-plex TMT labeling kit was from Pierce (Rockford, Ilinois), and all other reagents were from Sigma (Steinheim, Germany).

ARPE-19 Samples. Cells were cultured under standard conditions in DMEM/F12 (1:1 mixture of Dulbecco's modified Eagle's medium and Ham's F12), 10% fetal calf serum (FCS) and penicillin/streptomycin. ARPE-19 cells from passage 20-23 were used and the media was changed every 3 days. Cells grown in these conditions constitute a monolayer that retains the functionality, polarity and tight junction expression of the human RPE. For our study, cells were seeded in Petri dishes (10 cm) at 0.4 x104 cells/mL and maintained in culture for 21 days with 5.5 mM or 25 mM of D-Glucose at 37°C under 5% (v/v) CO2 in an incubator. During the last 24 hours cells were subjected to serum deprivation (1% FCS). Serum deprived media were prepared with 5.5 mM or 25 mM of either D-Glucose or D-[U-13C]-Glucose. Each condition was run in triplicate.

Human samples. Vitreous humor samples were obtained from 28 voluntary patients that undergone eye surgery in Clínica Barraquer. The samples were extracting using vitrectomy followed by suction and then frozen at -80ºC and stored until sample preparation.

ARPE-19 Quantitative proteomics

Cell lysis and protein digestion. ARPE-19 cells were lysed in lysis buffer (50 mM ammonium bicarbonate, 8 M urea, 1 tablet Complete Mini EDTA-free protease inhibitor cocktail, 1 tablet PhosSTOP phosphatase inhibitor cocktail). Lysis was performed by gentle sonication on ice at 20% amplitude, with a 0.5 cycle in a Sonics Vibracell (Bioblock Scientific, France). Cell debris were removed by centrifugation at 20,000 g for 10 min at 4°C. Protein concentration was determined by an RC-DC protein assay (Bio-Rad). Proteins were reduced in 4 mM dithiothreitol (30 min at 56°C) and alkylated in 8 mM iodoacetamide (30 min at room temperature in the dark). LysC was added at an enzyme:protein ratio of 1:75 (w/w) and incubated for 4 h at 37°C. Samples were then diluted 4 times with 50 mM ammonium bicarbonate. Trypsin was added at an enzyme:protein ratio of 1:100 (w/w) and incubated overnight at 37°C. Acetic acid was added to a final concentration of 10% and samples were immediately frozen.

TMT labeling. 100 µg of each sample were desalted and concentrated using C18 solid phase extraction (Sep-Pak Vac C18 cartridge 1 cm3/200 mg, Waters), dried in vacuum and reconstituted in 120 µL of 200 mM triethylammonium bicarbonate (Sigma). Labeling was performed with the 6-plex labeling kit according to the manufacturer's protocol. Briefly, each labeling was carried out for 1 h at room temperature and quenched with 8 µL of 5 % hydroxylamine. The four channels were mixed, dried in vacuum and resuspended in 10% formic acid.

Strong cation exchange fractionation. Peptides were fractionated by strong cation exchange (SCX) using a Zorbax BioSCX-Series II column (0.8 mm × 50 mm, 3.5 µm), as described by Munoz et al. [286]. Solvent A consisted of 0.05% formic acid in 20% acetonitrile, solvent B of 0.05% formic acid, 0.5 M NaCl in 20% acetonitrile. The gradient was 0 to 2% B in 0.01 min; 2 to 3% B in 8 min; 3 to 8% B in 6 min; 8 to 20% B in 14 min; 20 to 40% B in 10 min; 40 to 90% B in 10 min; 90%B for 6 min; 90 to 0% B in 6 min. Fractions were collected once a minute and each dried in vacuum and stored at -20°C.

LC/MS analyses. SCX fractions were analyzed on an Orbitrap Q-Exactive (Thermo Fisher Scientific) connected to an UHPLC Proxeon Easy-nLC 1000 (Thermo Scientific). Peptides were trapped on a double-fritted trap column (Dr. Maisch Reprosil C18, 3 µm, 2 cm × 100 µm) and separated on an analytical column (Agilent Zorbax SB-C18, 1.8 µm, 40 cm × 75 µm) as previously described by Cristobal et al. [287]. Solvent A consisted of 0.1 M acetic acid, solvent

B of 0.1 M acetic acid in 80% acetonitrile. Samples were loaded at a pressure of 800 bar with 100% solvent A. Peptides were separated by a 110 min gradient from 10% to 40% solvent B at a flow rate of 150 nL/min. Full scan MS spectra were acquired in the Orbitrap (350-1500 m/z, resolution 35,000, AGC target 3e6, maximum injection time 250 ms). The 20 most intense precursors were selected for HCD fragmentation (isolation window 1.2 Da, resolution 17,500, AGC target 5e4, maximum injection time 120 ms, first m/z 100, NCE 33%, dynamic exclusion 60 s).

Data analysis. Raw data was analyzed with MaxQuant[288] (version 1.3.0.5). MS/MS peak lists were generated and searched with Andromeda against the Swissprot human database. Trypsin/P was chosen as an enzyme, with a maximum of 2 missed cleavages. Methionine oxidation was set as variable modification. Cysteine carbamidomethylation as fixed modifications, TMT6plex (Lys) and TMT6plex (N-term) as the reporter ion quantification method. The database search was performed with a precursor tolerance of 6 ppm for the main search (20 ppm for the first search) and a fragment mass tolerance of 0.05 Da. Match between run was enabled with a time window of 2 min. Peptide and protein FDR were set at 1%, and peptide score threshold at 60. The quantification and statistical processing was performed in Perseus v.1.3.8.1. Proteins were grouped and reporter ion intensities were calculated for each of the TMT channels. Ratios were calculated and normalized on median. A significance B test was performed to determine significantly regulated proteins, where truncation was performed using p values, with a threshold value of 0.05 followed by a Benjamini false discovery rate correction.

ARPE Metabolomics

Metabolites extraction method. The culture medium was removed from cells and the dishes were placed on top of dry ice. Cells were scrapped immediately and metabolites extracted into the extraction solvent by adding 2 mL of a cold mixture of chloroform/methanol (2:1 v/v). The resulting suspension was bath-sonicated for 3 minutes, and 2 mL of cold water was added. Then, 1 mL of chloroform/methanol (2:1 v/v) was added to the samples and bath-sonicated for 3 minutes. Cell lysates were centrifuged (5000 × g, 15 min at 4 °C) and the aqueous phase was carefully transferred into a new tube. The aqueous and organic phases were frozen, lyophilized and stored at −80 °C until further analysis.

The organic extracts were reconstituted in 400 µl of MeOH and 200 µL of CHCl3. Samples were then sonicated, vortexed and kept at room temperature during 10 minutes. Then, 200 µL of CHCl3 were added to the sample and vortexed for 1 minute. Afterwards, 200 µL of H2O were added, vortexed and bath-sonicated for 3 minutes. The sample was kept at room temperature during 15 minutes in order to get the phases separated. The organic phase was carefully transferred into a new tube and was dried using N2. The sample was reconstituted in 300 µL of IPA:ACN:H2O (30:65:5) and centrifuged for 15 min at 15000 rpm and 4 °C. Vials containing extracted metabolites from the organic phase were kept at −80 °C prior to LC/MS analysis.

NMR analyses. The hydrophilic extracts were reconstituted in 600 µl of D2O containing 0.67 mM trisilylpropionic acid (TSP). Samples were then vortexed, and centrifuged for 15 min at 6000 × g and 4 °C. Finally, redissolved samples were placed into 5 mm NMR tubes. 1H and 13C NMR spectra were recorded at 300 K on an Avance III 600 spectrometer (Bruker, Germany) operating at a proton frequency of 600.20 MHz using a 5 mm CPTCI triple resonance (1H, 13C, 31P) gradient cryoprobe. One-dimensional 1H pulse experiments were carried out using the nuclear Overhauser effect spectroscopy (NOESY) presaturation sequence to suppress the residual water peak. The acquired spectral width was 12 kHz (20 ppm), and a total of 256 transients were collected into 64 k data points for each 1H spectrum. 13C NMR spectroscopy was performed under approximately fully relaxed conditions (repetition time 8 seconds) and broadband proton decoupling. A total of 1024 scans and 64 k data points with a spectral width of 36 KHz (240 ppm) were acquired for each 13C spectrum. Exponential line broadening of 0.3 Hz was applied before Fourier transformation and frequency domain spectra were phased and baseline-corrected using TopSpin software (version 2.1, Bruker).

LC/MS analyses. LC/MS analyses were performed using an UHPLC system (1290 series, Agilent Technologies) coupled to a 6550 ESI-QTOF MS (Agilent Technologies) operated in positive (ESI+) and negative (ESI-) electrospray ionization mode.

Aqueus phase.Fractions of 100 µL of each redissolved aqueous sample in deuterated water were placed into HPLC vials after NMR analysis with no need for solvent exchange as previously reported [310]. When the instrument was operated in positive ionization mode, metabolites were separated using an Acquity UPLC (HSS T3) C18 reverse phase (RP) column (2.1 x 150 mm, 1.8 µm) and the solvent system was A1 = 0.1% formic acid in water and B1 = 0.1% formic acid in acetonitrile. When the instrument was operated in negative ionization mode, metabolites were separated using an Acquity UPLC (BEH) C18 RP column (2.1 x 150 mm, 1.7 µm) and the solvent system was A2 = 1 mM ammonium fluoride in water and B2 = acetonitrile, as previously reported[1]. The linear gradient elution started at 100% A (time 0–2 min) and finished at 100% B (10-15 min). The injection volume was 5 µL. ESI conditions: gas temperature, 150 °C; drying gas, 13 L min–1; nebulizer, 35 psig; fragmentor, 400 V; and skimmer, 65 V. The instrument was set to acquire over the m/z range 100–1500 in full-scan mode with an acquisition rate of 4 spectra/sec. MS/MS was performed in targeted mode, and the instrument was set to acquire over the m/z range 50–1000, with a default isolation width (the width half-maximum of the quadrupole mass bandpass used during MS/MS precursor isolation) of 4 m/z. The collision energy was fixed at 20 V.

-Organic phase. Metabolites were separated using an Acquity UPLC (BEH) C8 RP column (2.1 x 150 mm, 1.7 µm) and the solvent system was A = H2O:ACN (60:40) 10 mM ammonium formate and 0.1 % formic acid and B = IPA:ACN (95:5) 10 mM ammonium formate, 0.1 % formic acid and 0.1 % H2O. The linear gradient started at 32% solvent B (time 0 -1 min) and increased to 60% solvent B within 3 min. In the following 11 min solvent B percentage was increased to 100% and was kept at this level for 0.5 min. In the following 0.5 min solvent A percentage was increased to 100% and was kept at this level for 3 min. In the next 0.5 min, solvent B percentage was increased to 100% and was kept at this level for 2 min. In the next 0.5 min, solvent A percentage was increased to 100% and was kept at this level for 3 min. In the following 0.5 min solvent B percentage was increased to 100% and was kept at this level for 3 min. Starting conditions were achieved in 0.5 min and the column was re-equilibrated for 3 min, resulting in a total UHPLC run time of 32 min. The injection volume was 1 µL. ESI conditions: gas temperature, 150 °C; drying gas, 13 L min–1; nebulizer, 35 psig; fragmentor, 400 V; and skimmer, 65 V. The instrument was set to acquire over the m/z range 50–1200 in full-scan mode with an acquisition rate of 3 spectra/sec. MS/MS was performed in targeted mode, and the instrument was set to acquire over the m/z range 50–1000, with a default isolation width (the width half-maximum of the quadrupole mass bandpass used during MS/MS precursor isolation) of 4 m/z. The collision energy was fixed at 20 V.

GC/MS analyses. Redissolved samples in deuterated water were lyophilized, dissolved in 50µL of methoxyamine hydrochloride in pyridine (30 mg/mL) and incubated with agitation during 1 hour at 65°C. Trimethylsililation was done by adding 30µl of N-methyl-N-trimethylsilyltrifluoroacetamide previously spiked with the FAMEs mix as internal standard. The samples were then shacked for 10 min and kept for 1 hour at room temperature. Derivatised samples were analysed in a 7890A Series gas chromatograph coupled to a 7200 GCqTOF MS (Agilent Technologies, Santa Clara, U.S.A.). Chromatographic column was a J&W Scientific HP5-MS (30 m x 0.25 mm i.d., 0.25 µm film) (Agilent Technologies). A volume of 1µL of sample was automatically injected into a split/splitless inlet, which was kept at a temperature of 250ºC. Helium was used as a carrier gas, at a flow rate of 1 mL/min in constant flow mode. The oven program was set at an initial temperature of 70°C for 1 min, then increased to 325°C at a rate of 10°C/min and held at 325°C for 9.5 min. Ionization was done by electronic impact, with an electron energy of 70 eV and an emission intensity of 35 A. Source temperature was of 230ºC. Mass spectra were recorded after a solvent delay of 6 minutes, after which the analyzer acquired in full-scan MS mode at a rate of 5 scan/sec, acquiring a mass range of 35-700 m/z.

Vitreous humor Metabolomics

Metabolites extraction method. Vitreous humor was centrifuged for 15 min at 15000 rpm and 4 °C. 300 µL of supernatant were placed in a new eppendorf tube. 250 µL of 0.73 mM TSP buffer in D2O were added to the eppendorf tubes containing the vitreous humor. Samples were then vortexed, and centrifuged for 15 min at 15000 × rpm and 4 °C. Finally, redissolved samples were placed into 5 mm NMR tubes. After the NMR analysis, each sample was split in two aliquots of 250 µL, freeze-dried and kept at -80ºC until be used for LC/MS and proteomics analysis. For LC/MS analysis, freeze-dried aliquots coming from NMR were resuspended in 100 µL of MeOH:H2O (8:2). Then, they were sonicated for 2 minutes, vortex and kept at -20ºC for 2 hours. After that, the samples were centrifuged for 15 min at 15000 x rpm and 4 ºC. Finally, the supernatant was placed into a LC-MS vial.

NMR analysis. 1H spectra was recorded at 300 K on an Avance III 600 spectrometer (Bruker, Germany) operating at a proton frequency of 600.20 MHz using a 5 mm CPTCI triple resonance (1H, 13C, 31P) gradient cryoprobe. One-dimensional 1H pulse experiments were carried out using the nuclear Overhauser effect spectroscopy

(NOESY) presaturation sequence to suppress the residual water peak. The acquired spectral width was 12 kHz (20 ppm), and a total of 256 transients were collected into 64 k data points for each 1H spectrum. Exponential line broadening of 0.3 Hz was applied before Fourier transformation and frequency domain spectra were phased and baseline-corrected using TopSpin software (version 2.1, Bruker).

LC/MS analyses. LC/MS analyses were performed using an UHPLC system (1290 series, Agilent Technologies) coupled to a 6550 ESI-QTOF MS (Agilent Technologies) operated in positive (ESI+) and negative (ESI-) electrospray ionization mode. Two different kind of chromatographies were applied, reverse phase and HILIC in both ionization modes, positive and negative, so, four different analytical conditions in different runs. For reverse phase and the instrument operating in positive ionization mode, metabolites were separated using an Acquity UPLC (HSS T3) C18 column (2.1 x 150 mm, 1.8 µm) and the solvent system was A1 = 0.1% formic acid in water and B1 = 0.1% formic acid in acetonitrile. For negative ionization mode, metabolites were separated using an Acquity UPLC (BEH) C18 column (2.1 x 150 mm, 1.7 µm) and the solvent system was A2 = 1 mM ammonium fluoride in water and B2 = acetonitrile, as previously reported (25 old reference). For HILIC phase and both ionization modes, metabolites were separated using an Acquity UPLC (BEH) HILIC column (2.1 x 150 mm, 1.7 µm) and the solvent system was A1 = 10 mM ammonium acetate in 5:95 (ACN/H2O) and B1 = 10 mM ammonium acetate in 95:5 (ACN:H2O). In the case of reverse phase, the linear gradient elution started at 100% A (time 0–2 min) and finished at 100% B (10-15 min). For HILIC phase, the linear gradient elution started at 99% B (time 0-1 min) and finished at 50% B (time 1-10 min).The injection volume was 2 µL. ESI conditions: gas temperature, 150 °C; drying gas, 13 L min–1; nebulizer, 35 psig; fragmentor, 150 V; and skimmer, 65 V. The instrument was set to acquire over the m/z range 80–1000 in full-scan mode with an acquisition rate of 4 spectra/sec. MS/MS was performed in targeted mode, and the instrument was set to acquire over the m/z range 50–1000, with a default isolation width (the width half-maximum of the quadrupole mass bandpass used during MS/MS precursor isolation) of 4 m/z. The collision energy was fixed at 20 V.

LC/MS and GC/MS preprocessing. Raw data files were transformed into .mzXML format using Proteowizard software and then were preprocessed using the XCMS R package, to detect and align features. The parameters used in the XCMS workflow were: xcmsSet(method="centWave", ppm=30, peakwidth=c(5,20)); retcor(method="obiwarp", profStep=0.1) and group(mzwid=0.0065, minfrac=0.5, bw= 4). A feature is defined as a molecular entity with a unique m/z and a specific retention time (mzRT). XCMS analysis of these data provided different matrix (xcmsSet object) containing the retention time, m/z value and integrated peak area of each feature for every ARPE-19 sample. A feature is defined as a molecular entity with a unique m/z and a specific retention time (mzRT). XCMS analysis of these data provided a matrix containing the retention time, m/z value, and integrated peak area of each feature for every ARPE-19 sample. Quality control samples (QCs) consisting of pooled ARPE-19 samples from each four conditions were used in LC/MS and GC/MS analyses. QCs were injected at the beginning and periodically every 5 samples. Furthermore, samples entering the study were entirely randomized to reduce systematic error associated with instrumental drift. QCs were always projected in a PCA model together with the samples under study to verify that technical issues do not mask biological information. The performance of the analytical platform for each detected mzRT feature in ARPE-19 samples was assessed by calculating the relative standard deviation of these features on pooled samples (CVQC) according to Vinaixa et al. (29). ARPE-19 samples were compared using the integrated peak area of each feature via one-way ANOVA followed by Tuckey’s multiple comparison test and assigning a fold value to indicate the level of differential regulation due hypoxic and/or hyperglycemic conditions. Differentially regulated metabolites that were statistically significant ($p<0.05$) between physiological-like (N5) and any of the pathological-like conditions (N25, H5 and H25) detected by LC/MS were characterized by MS/MS. Differentially regulated metabolites ($p<0.05$) between N5 and any of the pathological-like conditions detected by GC/MS were identified using the NIST and Fiehn mass spectral libraries. In addition, the retention time of pure standards were confirmed.

For isotopic label tracking, geoRge package in R 3.3 was used. The output table was manually explored and interesting features were identified in MS/MS analysis.

Recon 2 gene enrichment with Uniprot identifers.

Network building and analysis. All Python code described here can be found in the supporting information as HTML Notebook. The enriched Recon 2 SBML file was loaded into Python (version 3.4.4) using libsbml module, the annotation was parsed into Python dictionaries and the metabolic network was built using igraph module. Proteomics data was loaded into Python using a CSV file.

Results

For survival, cells need to control and maintain many processes; DNA synthesis, energy storage, extracellular crosstalk and metabolism stability, by keeping a pool of compounds in order to synthesize polymers and set secondary messengers for signalling cascades, to do so they need to obtain and metabolize multiple compounds. Enzymes are proteins responsible for metabolic processes and thus their regulation is closely related to the concentration of the metabolites that they transform. In metabolism, there is a complex interlinking of biochemical reactions, regulated by even more complex mechanisms, negative feedback (or feedback inhibition), protein modification or allosteric regulation.

Yet, even when ignoring enzymatic regulation, metabolism can be defined as an intricate network of metabolites and reactions, meaning that some metabolites are metabolized by many reactions. These reactions, in turn, depend on the expression of one or more proteins, meaning that the concentration of these metabolites is controlled by multiple genes.

However, when a mutation, chemical insult or perturbation occurs, regulatory mechanisms respond balancing gene and protein expression to prevent cell death. This modification on enzyme regulation may cause changes on the cell's metabolic profile, some metabolites accumulate or a pathway has an increased influx. When this takes place as a response to a disease, these changes are referred as the disease-associated phenotype. In consequence, the characterization of the phenotype-associated metabolic profile is key to understanding, early diagnosing or even treating some diseases.

Used in many areas, simulation is a quick way to gain insight of relatively understood systems. Recon 2 was designed as a resource for systems biologists to be used as a comprehensive model that includes accurate metabolic reaction annotation, by running simulations in Recon, it is possible to predict the phenotype of some gene-caused metabolic dysfunctions. Besides, Recon 2 builds into a manually curated metabolic network, which can be enriched with gene-protein annotation, these protein identifiers can be linked to the corresponding Recon reactions. Extending Recon 2 annotation with associated proteins yields a new layer of useful data for metabolism interrogation to study genotype effects on the metabolite profile. Then, using the metabolic network linkage to obtain a list of associated proteins per metabolite, namely the proteins that participate in the reactions that consume or produce that metabolite. Taking into account the statistical analysis of this protein sets, for most metabolites, the number of proteins associated, sample size, is high enough for drawing a distribution of values and statistical testing (Figure 37).

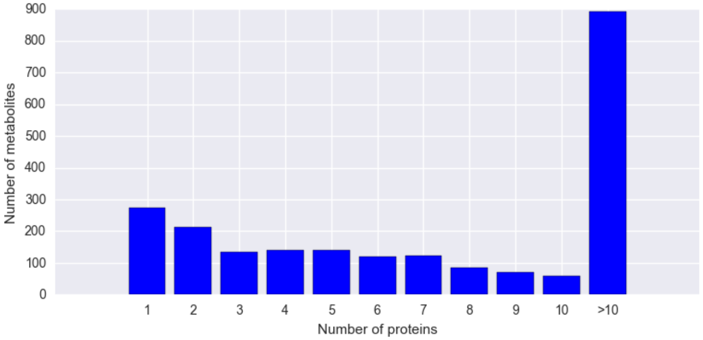


Figure 37. Bar plot representing the number of associated proteins per metabolite.

However, Recon 2 was designed to represent different cell types, so that some of genes, thus proteins, may not be expressed in the tissue sample under study.

In order to overcome that the complete network has to be refined so it contains only those proteins quantified in the proteomics experiments, reducing the primary Recon network to generate one that would be more comparable to the metabolic network occurring in the sample (from now on: 'filtered network'). Given that gene-protein-reaction relationships are available in Recon 2 annotation (when enriched with gene-protein annotation), it could be used to filter which reactions should be employed to build the filtered network. To do so, each combination of genes (reaction modifiers) of a reaction is evaluated; and a reaction is selected, if at least one protein for each gene of the modifier list is present in the proteomics dataset. In other words, if there is proof that all genes necessary to regulate a reaction are expressed, at least one protein of each gene is detected, that reaction will be used to build the filtered network and its proteins will be associated to it.

Once the filtered metabolic network is built, the list of metabolites is reduced as a consequence of reducing the number of reactions that, based on this rationale, would take place in the biological system. Next, by iterating over them, the reactions that connect each metabolite to neighbouring metabolites are obtained, represented as links in the metabolic network.

Using the associated proteins to those reactions, a list of proteins can be related to each metabolite. As in Figure 38, for most of the metabolites, the number of proteins would be still large enough to generate a distribution of values so a binomial test could be applied to calculate if that distribution is equally centered around non-change, meaning that the null hypothesis considers that the values are evenly balanced, if the concentration of the metabolite in the sample would not be affected, meaning that production and consumption are close to equilibrium.

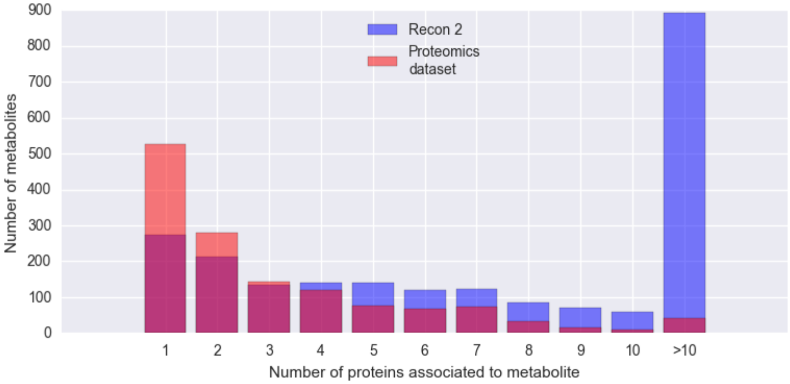


Figure 38. Bar plot representing the number of associated proteins per metabolite in Recon 2 (blue) and in our Proteomics dataset (red).

The binomial test is performed for each metabolite in the filtered metabolic network, those metabolites with a significant p-value, whose distribution of protein values is different from equilibrium, may be used for targeted metabolomics analysis. A schematic workflow representing this procedure is shown in Figure 39:

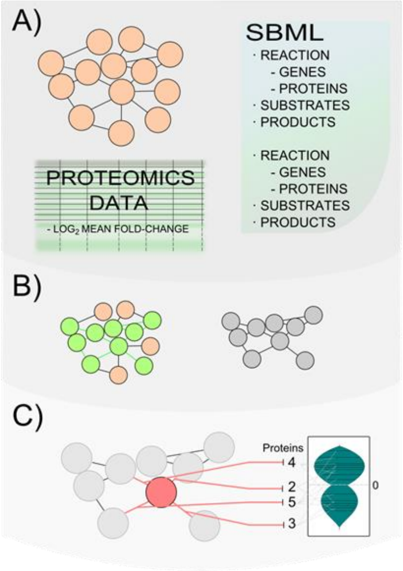


Figure 39. Scheme representing our novel data analysis workflow based on the topology of a network built using the Recon 2 model.

To challenge this novel method, we used proteomics data from ARPE-19 samples in which a total of 3260 proteins were quantified for all replicates. Additionally, these samples were metabolically profiled both at the pool and flux levels. Finally, we validated our results using human vitreous humour samples for which we obtained a global profiling for significantly altered metabolites. Interestingly, some of the alterations occurring in the *in vitro* model could be confirmed in the human samples and they could even be expanded with further similar results.

Based on the 3259 proteins, a total of 1766 reactions were filtered out of the 4320 Recon 2 non-transport reactions, resulting in mapping 1390 metabolites out of the 2000 found in Recon 2. Out of these, 162 were found to be significant after binomial testing based on the proteins associated to them.

Most of the 127 metabolites (~ 78%) are part of a lipid family, the rest are related to glycolysis, TCA, aminoacid metabolism, vitamins and glutathione metabolism, as reported by Table 11 (pathway enrichment analysis using IMPALA).

Table 11. Pathway enrichment analysis from IMPALA.

Pathway	P-value	Q-value
Fatty acid triacylglycerol and ketone body metabolism	7.53E-17	2.69E-13
Fatty Acid Beta Oxidation	5.15E-09	7.66E-07
Metabolism of amino acids and derivatives	8.84E-05	0.00631
Metabolism of nucleotides	0.000209	0.013
Pentose Phosphate Pathway	0.000381	0.0216
Methylation Pathways	0.000785	0.0301
Trans-sulfuration and one carbon metabolism	0.00139	0.034
Alpha-linolenic (omega3) and linoleic (omega6) acid metabolism	0.0016	0.0371
Trans-sulfuration pathway	0.00188	0.0371
Glucuronidation	0.00255	0.0457
Asparagine N-linked glycosylation	0.00314	0.0552
TCA Cycle	0.00331	0.0573

Nearly all presented a down regulation tendency in the protein expression profile. So as to understand the high amount of lipid-related candidates, we investigated the proteins associated to this family of metabolites, which revealed that they shared a great number of proteins, especially those related to mitochondrial fatty acid oxidation. In order to prove the validity of this result, we performed an untargeted lipidomic profiling of the ARPE-19 cell cultures, for which we found an important number of significantly down-regulated triacylglycerols in hyperglycemic and hypoxic conditions (N25 and H25) at the putative identification level (level 1 according to MS standards) as it can be seen in Figure 40.

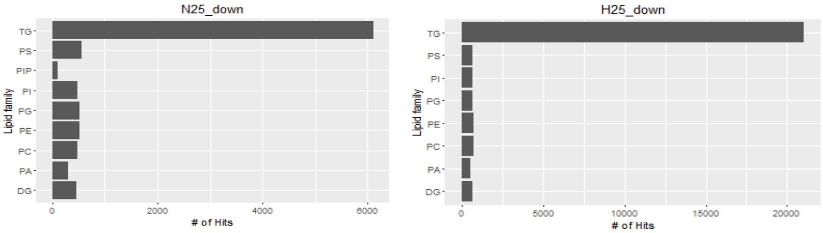


Figure 40. Down-regulated lipidomic profiling of the ARPE-19 cell cultures in hyperglycemic and hypoxic conditions (N25 and H25) at the putative identification level.

The remaining significant metabolites could be mapped into diverse metabolic pathways. With the objective of validating if these candidate metabolites had actually an altered concentration, or rate of synthesis, we performed untargeted metabolomics of the polar cell extracts and used U-13C-Glucose for isotopic label tracking for untargeted isotopologues analysis. As a result, we found four metabolites (N-acetyl-neuraminic acid, N-acetyl-glucosamine, Alpha-D-Fucose and UDP-Fucose) significantly altered from Asparagine N-linked glycosylation pathway (see Figure 41). N-acetyl-neuraminic acid was down-regulated in hyperglycemic conditions (N25 and H25), N-acetyl-glucosamine and UDP-Fucose were down-regulated only in hypoxic with hyperglycemic condition (H25) and finally, Alpha-D-

Fucose showed up regulation in hyperglycemic with normoxic condition (N25).

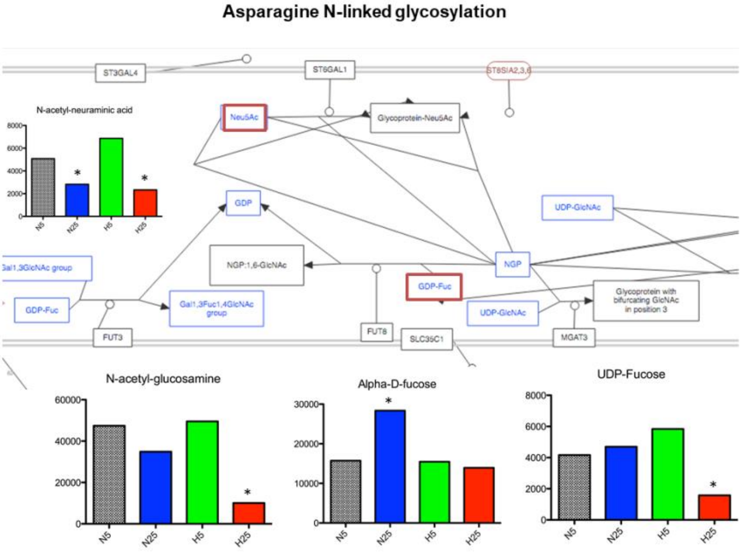


Figure 41. Significantly altered* metabolites (N-acetyl-neuraminic acid, N-acetyl-glucosamine, Alpha-D-Fucose and UDP-Fucose) present in Asparagine N-linked glycosylation pathway.

As further evidence of the clinical relevance of the results found in the *in vitro* ARPE-19 cell cultures, untargeted metabolomics data was acquired for human vitreous humor samples of patients at different stages of DR, which are 14 controls (healthy retinas), 4 non-proliferative DR (NPDR), an early stage and 12 proliferative DR (PDR), the latest stage.

Some of the metabolites reported by the metabolic network analysis and also from the untargeted metabolomics analysis were once more significantly altered in the human samples when comparing controls versus the latest stage of the disease. Moreover, some new metabolites, metabolically close to the formerly reported, were also output by metabolomics analysis. I our metabolomic study we found up-regulation in PDR condition of metabolites involved in these two pathways: trans-sulfuration and one carbon metabolism and ketone body metabolism.

10-formyl-THF, S-Adenosylhomocysteine (SAH), S-Adenosylmethionine (SAMe) and Methionine, metabolites belonging to trans-sulfuration and one carbon metabolism pathway, were up-regulated in PDR (see Figure 42).

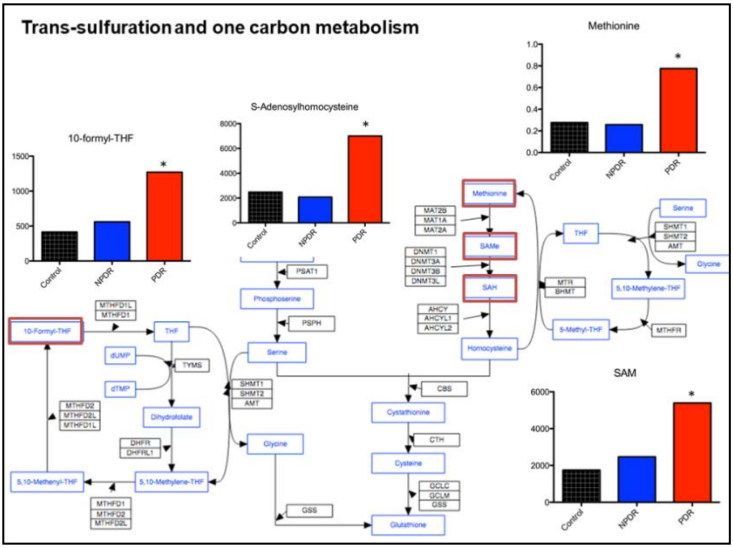


Figure 42. Trans-sulfuration and one carbon metabolism pathway showing the bar plots from altered metabolites found in vitreous humour.

And finally, metabolites present in ketone body metabolism pathway such as 3-hydroy-butyrate, Acetoacetate and Acetone were up-regulated in PDR, as it was mentioned above (see Figure 43).

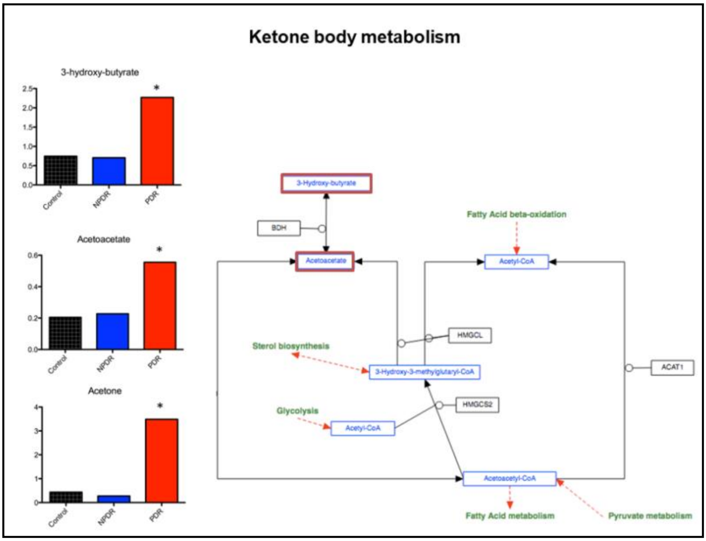


Figure 43. Ketone body metabolism pathway showing the bar plots from altered metabolites found in vitreous humour.

CHAPTER 4. Discussion

Discussion

In this thesis we have developed, analysed and validated new bioinformatic tools for converting raw MS-based metabolomics data into biological knowledge, in order to study alterations in the proteome and metabolome of human retinal pigment epithelium cells exposed to hyperglycemic and/or hypoxic conditions.

To reach these aims, this thesis has structured in two main blocks, the first one addressing a more methodological challenge to improve the metabolomics data while the second block is focused on the biology of diabetic retinopathy.

In the first block we have studied deeply the state of the art of metabolomics to identify the weak points that need to be improved. Metabolomics as a tool is evolving continuously with the aim of detecting a bigger number of metabolites more accurately. This progression finally impacts in a better characterization of the phenotype that needs to be studied. As it has been explained before, metabolomics consist in a workflow defined by a number of consecutive steps: sample preparation, data acquisition, data processing, data analysis, metabolite identification and lastly, biological interpretation. However, although the workflow is well stablished, still there is not a unique methodology that can be used for each step unlike in the other *omics* levels from the cascade (genomics, transcriptomics and proteomics). This issue is due to different reasons such as metabolomics is the newest *omic* from the cascade and thus less time has been invested in improving it but also the type of analytes (metabolites) to be studied plays an important role in defining the methodology that must be used. Compared to DNA, RNA and proteins, metabolites are much more complex since they present a broad range of different physicochemical properties making difficult to characterize the complete metabolome. Moreover, metabolites are not derived from a gene code, they result from metabolic reactions, degradation products or they can be exogenous because they come from diet. Altogether presents a big challenge that need to be resolved through improving the whole metabolomics workflow. In this thesis we have addressed these methodological issues by (i) analysing mass spectral databases for LC/MS-based untargeted metabolomics and (ii) generating and improving the characterization of LC/MS metabolomics data focusing on MS1 and MS2 annotation.

As a result of aim (i), I contributed to sutdy the NIST14 mass spectral database proving it is superior to other available databases due to its larger number of metabolites with spectral data acquired from different adducts and using a wide range of mass spectrometers. I showed the importance of adduct formation for metabolite identification, analysing MS/MS data for different adducts in my lab. This is particularly important because predominant adducts in an ESI spectrum vary from one metabolite to another as well as on the mobile phase used. Thus, the implementation of NIST14 database during the identification step in the metabolomics leads to improve and increase the whole metabolome from a biological system.

In aim (ii) we have collaborated in evaluating the performance of two new computational tools: CliqueMS and iMet, which were developed in collaboration with SEES lab.

Brievly, CliqueMS annotates in-source MS1 data based on a coelution similarity network. I have experimentally proved that CliqueMS correctly identifies and annotates a large number of adducts from pure standards and complex biological samples. In addition, I compared its performance to CAMERA's one, the most widely-used approach in metabolomics community, I have demonstrated that CliqueMS leads to a more correctly parental ion neutral masses than CAMERA. As a result, this new tool could be implemented in the metabolomics workflow to obtain a better annotation of ions increasing the number of metabolites that can be detected in biological system.

On the other hand, we evaluated the performance of iMet, a tool that facilitates structural annotation of metabolites based on MS2 data not described in mass spectral databases. I simulated a real scenario of metabolites not present in a database by testing iMet using metabolites proposed in the Critical Assessment of Small Molecule Identification (CASMI) challenges from years 2012-2016. In addition, I compared iMet's performance to other tools such as

CFM-ID, MetFrag and MS-Finder. I could demonstrate the potential of iMet for annotating metabolites that are not present in databases, as well as to compare the performance of other methods to assist the structural annotation of known metabolites lacking MS/MS spectra in databases.

The aims achieved in the methodological block can be taken as new implementations in the metabolomics workflow leading to a better characterization of the metabolome under study.

For addressing the biological aim, we have studied the biological mechanism leading to diabetic retinopathy using metabolomics and proteomics approaches. We also have taken advantage of the training and results obtained during methodological block to improve the metabolome coverage in ARPE-19 cells and vitreous humour samples. As a result (iii) we have detected and analysed changes in protein-protein interaction (PPI) networks by hyperglycemic and/or hypoxic conditions, and (iv) predicted and validated metabolite alterations due to hyperglycemic and/or hypoxic conditions integrating protein expression data in metabolic networks.

In aim (iii) I have generated proteomics data and developed a novel approach that integrates PPI, module analysis and protein expression for detecting dysregulated groups of interacting proteins involved in similar biological processes. As a result from this study we could observed a better coordinated in smaller interactome distances or in other words, protein expression was more similar if the proteins are closer in the PPI network. Having this into consideration, we applied a stockastic block model method to obtain protein modules for a further statisticall test for differente conditions experimental conditions: H5, N25 and H25. The hyperglycemic and hypoxic conditions, H25 was the one shwoing a larger number of dysregulated modules. In order to test if there was any biological significance relevance we performed a biological enrichment module analysis in which observed that experimental condition H25 was characterized by the maximum number of dysregulated biological processes and pathways in modules.

In general, the hyperglycemic conditions is characterized by a down-regulation of the processes involved in the ubiquitin protein ligase activity, signal transduction, developmental process and cellular amino acid metabolic process while the up-regulated process was associated to DNA replication. Hypoxic conditions were characterized by up-regulation of RNA-splicing. Finally, only detected in H25, there were down-regulated of RNA catabolic process, translational initiation and regulation of proteolysis and up-regulated process involving cell adhesion.

We also compared the clinical results obtained from vitreous humor during the progressive stages of the disease (control, NPDR AND PDR) and we found cell adhesion was the only common process found dysregulated between the comparison PDR/CONTROL from vitreous humor and the protein module from ARPE-19 but associated to different proteins in each case. While in vitreous humor the overall protein FC appeared down-regulated in ARPE-19 was up-regulated.

Only using this new approach it has been possible to capture slight but consistent protein changes occurring in a protein module which are impossible to detect applying the statistical classical proteomic approach which considers each protein individually.

Finally, in aim (iv) I have validated a novel proteomics data analysis workflow based on a human genome-scale metabolic network that predicted metabolic alterations in an in vitro model of DR. I have generated and analysed metabolomics data on ARPE-19 cells cultured at low and high glucose concentrations, and normoxic or hypoxic conditions, also fed with 13C-glucose for isotopic label tracking (flux analysis), to validate the predictions made by our novel data analysis workflow. In addition, I also analysed human vitreous humor from DR patients and controls for clinical validation.

I have demonstrated this approach is a reliable method to study proteomics datasets that contain a considerable amount of quantified enzymes, otherwise this analysis is not feasible since the metabolic network is unconnected.

From our validation we found altered metabolites belonging to pathways that were enriched in our metabolic network such as fatty acid triacylglycerol and ketone body metabolism, asparagine N-linked glycosylation and trans-sulfuration and one carbon metabolism.

To conclude, we believe that this methodology shows the significance of studying metabolic problems from a network point of view and will raise awareness on future proteomics studies.

CHAPTER 5. Conclusions

- Identification of the weak points that need to be improved in processing step of the metabolomic workflow.
- NIST14 mass spectral database is superior to other available databases due to its larger number of metabolites with spectral data acquired from different adducts and using a wide range of mass spectrometers.
- The adduct formation must be considered during metabolite identification because predominant adducts in an ESI spectrum vary from one metabolite to another as well as on the mobile phase used.
- CliqueMS annotation tool for MS1 data can be used as a tool to identify and annotate a large number of adducts from pure standards and complex biological samples.
- CliqueMS shows a higher performance than CAMERA, leading to a more correctly annotation of parental ion neutral masses.
- iMet, a tool that facilitates structural annotation of metabolites based on MS2 which are not present in databases.
- CFM-ID, MetFrag and MS-Finder are other tools for structural annotation that must be used in a complementary way to improve the metabolome coverage.
- Novel approach that integrates Protein-Protein Interaction (PPI), module analysis and protein expression able to detect

dysregulated groups of interacting proteins involved in similar biological processes.

- Protein expression was more similar if the proteins are closer in the PPI network.
- Hyperglycemic and hypoxyc (H25) is the explerimetnal condition responsible for inducing more dysregulation in protein levels.
- Hyperglycemic condition (H25 and N25) is characterized by a down-regulation of the processes involved in the ubiquitin protein ligase activity, signal transduction, developmental process and cellular amino acid metabolic process while the up-regulated process was associated to DNA replication.
- Hypoxic conditions (H25 and H5) were characterized by up-regulation of RNA-splicing.
- Only detected in H25, there were down-regulated of RNA catabolic process, translational initiation and regulation of proteolysis and up-regulated process involving cell adhesion.
- A further clinical validation studying the vitreous humor during the progressive stages of the disease found cell adhesion was the only common process dysregulated between the comparison PDR/CONTROL from vitreous humor and the protein module from ARPE-19 but associated to different proteins in each case.

- The novel proteomics data analysis workflow based on a human genome-scale metabolic network can predict metabolic alterations in an in vitro model of DR.
- This approach is a reliable method to study proteomics datasets that contain a considerable amount of quantified enzymes.
- Metabolites belonging to pathways that were enriched in our metabolic network such as fatty acid triacylglycerol and ketone body metabolism, asparagine N-linked glycosylation and trans-sulfuration and one carbon metabolism were also altered in the metabolomic approach.
-

CHAPTER 6. References

References

1. Yanes, O.; Tautenhahn, R.; Patti, G. J.; Siuzdak, G., Expanding coverage of the metabolome for global metabolite profiling. *Anal Chem* **2011,** *83* (6), 2152-61.

2. Ideker, T.; Galitski, T.; Hood, L., A new approach to decoding life: systems biology. *Annual review of genomics and human genetics* **2001,** *2*, 343-72.

3. A.P., D., *Opening New Horizons. In: Introduction to Fluorescence Sensing*. Springer, Cham: 2015.

4. Kitano, H., Systems biology: a brief overview. *Science* **2002,** *295* (5560), 1662-4.

5. Chuang, H. Y.; Hofree, M.; Ideker, T., A decade of systems biology. *Annual review of cell and developmental biology* **2010,** *26*, 721-44.

6. Peitsch, M. C.; de Graaf, D., A decade of Systems Biology: where are we and where are we going to? *Drug discovery today* **2014,** *19* (2), 105-7.

7. Lazebnik, Y., Can a biologist fix a radio?&#x2014;Or, what I learned while studying apoptosis. *Cancer Cell 2* (3), 179-182.

8. Joyce, A. R.; Palsson, B. O., The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol* **2006,** *7* (3), 198-210.

9. Graves, P. R.; Haystead, T. A., Molecular biologist's guide to proteomics. *Microbiology and molecular biology reviews : MMBR* **2002,** *66* (1), 39-63; table of contents.

10. Wanichthanarak, K.; Fahrmann, J. F.; Grapov, D., Genomic, Proteomic, and Metabolomic Data Integration Strategies. *Biomarker insights* **2015,** *10* (Suppl 4), 1-6.

11. Vuckovic, D., Current trends and challenges in sample preparation for global metabolomics using liquid chromatography-mass spectrometry. *Analytical and bioanalytical chemistry* **2012,** *403* (6), 1523-48.

12. Cusick, M. E.; Klitgord, N.; Vidal, M.; Hill, D. E., Interactome: gateway into systems biology. *Human molecular genetics* **2005,** *14 Spec No. 2*, R171-81.

13. Li, X.; Gianoulis, T. A.; Yip, K. Y.; Gerstein, M.; Snyder, M., Extensive In Vivo Metabolite-Protein Interactions Revealed by Large-Scale Systematic Analyses. *Cell 143* (4), 639-650.

14. Gallego, O.; Betts, M. J.; Gvozdenovic-Jeremic, J.; Maeda, K.; Matetzki, C.; Aguilar-Gurrieri, C.; Beltran-Alvarez, P.; Bonn, S.; Fernandez-Tornero, C.; Jensen, L. J.; Kuhn, M.; Trott, J.; Rybin, V.; Muller, C. W.; Bork, P.; Kaksonen, M.; Russell, R. B.; Gavin, A. C., A systematic screen for protein-lipid interactions in Saccharomyces cerevisiae. *Mol Syst Biol* **2010,** *6*, 430.

15. Functional States. In *Systems Biology: Constraint-based Reconstruction and Analysis*, Palsson, B. Ø., Ed. Cambridge University Press: Cambridge, 2015; pp 264-276.

16. Kell, D. B., Metabolomics and systems biology: making sense of the soup. *Current opinion in microbiology* **2004,** *7* (3), 296-307.

17. Weckwerth, W., Metabolomics in systems biology. *Annual review of plant biology* **2003,** *54*, 669-89.

18. Baker, M., Metabolomics: from small molecules to big ideas. *Nat Meth* **2011,** *8* (2), 117-121.

19. van der Greef, J.; van Wietmarschen, H.; van Ommen, B.; Verheij, E., Looking back into the future: 30 years of metabolomics at TNO. *Mass Spectrometry Reviews* **2013,** *32* (5), 399-415.

20. Yanes, O., La metabolómica: un *déjà vu* por la historia de la bioquímica. *Sociedad Española de Bioquímica y Biología Molecular* **2015,** *186*, 4-6.

21. Klingenberg, P., An Introduction to Metabolic Pathways, herausgegeben von S. Dagley and Donald E. Nicholson. 343 Seiten, zahlreiche Abb. Blackwell Scientific Publications, Oxford and Edinburgh 1970. Preis 75 s. *Food / Nahrung* **1971,** *15* (4), 473-473.

22. Gates, S. C.; Sweeley, C. C., Quantitative metabolic profiling based on gas chromatography. *Clinical Chemistry* **1978,** *24* (10), 1663.

23. Griffiths, J., A Brief History of Mass Spectrometry. *Analytical Chemistry* **2008,** *80* (15), 5678-5683.

24. Lerner, R. A.; Siuzdak, G.; Prospero-Garcia, O.; Henriksen, S. J.; Boger, D. L.; Cravatt, B. F., Cerebrodiene: a brain lipid isolated from sleep-deprived cats. *Proceedings of the National Academy of Sciences of the United States of America* **1994,** *91* (20), 9505-9508.

25. Oliver, S. G.; Winson, M. K.; Kell, D. B.; Baganz, F., Systematic functional analysis of the yeast genome. *Trends in Biotechnology* **1998,** *16* (9), 373-378.

26. Nicholson, J. K.; Lindon, J. C.; Holmes, E., 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* **1999,** *29* (11), 1181-1189.

27. http://www.metabolomics-msi.org/.

28. Dunn, W. B.; Ellis, D. I., Metabolomics: Current analytical platforms and methodologies. *TrAC Trends in Analytical Chemistry* **2005,** *24* (4), 285-294.

29. Yeung, P. A.-O. X., Metabolomics and Biomarkers for Drug Discovery. LID - E11 [pii] LID - 10.3390/metabo8010011 [doi]. (2218-1989 (Print)).

30. Wilcoxen, K. M.; Uehara T Fau - Myint, K. T.; Myint Kt Fau - Sato, Y.; Sato Y Fau - Oda, Y.; Oda, Y., Practical metabolomics in drug discovery. (1746-0441 (Print)).

31. Álvarez-Sánchez, B.; Priego-Capote, F.; Luque de Castro, M. D., Metabolomics analysis I. Selection of biological samples and practical aspects preceding sample preparation. *TrAC Trends in Analytical Chemistry* **2010,** *29* (2), 111-119.

32. Fernie, A. R.; Trethewey Rn Fau - Krotzky, A. J.; Krotzky Aj Fau - Willmitzer, L.; Willmitzer, L., Metabolite profiling: from diagnostics to systems biology. (1471-0072 (Print)).

33. Lindon, J. C.; Nicholson, J. K.; Holmes, E., Preface. In *The Handbook of Metabonomics and Metabolomics*, Elsevier Science B.V.: Amsterdam, 2007; pp ix-x.

34. Dunn, W. B.; Wilson, I. D.; Nicholls, A. W.; Broadhurst, D., The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. *Bioanalysis* **2012,** *4* (18), 2249-64.

35. Alonso, A.; Marsal, S.; Julia, A., Analytical methods in untargeted metabolomics: state of the art in 2015. *Frontiers in bioengineering and biotechnology* **2015,** *3*, 23.

36. Maher, A. D.; Zirah Sf Fau - Holmes, E.; Holmes E Fau - Nicholson, J. K.; Nicholson, J. K., Experimental and analytical variation in human urine in 1H NMR spectroscopy-based metabolic phenotyping studies. (0003-2700 (Print)).

37. Khamis Mona, M.; Adamko Darryl, J.; El-Aneed, A., Mass spectrometric based approaches in urine metabolomics and biomarker discovery. *Mass Spectrometry Reviews* **2015,** *36* (2), 115-134.

38. de Graaf, R. A.; Chowdhury, G. M. I.; Behar, K. L., Quantification of High-Resolution (1)H NMR Spectra from Rat Brain Extracts. *Analytical Chemistry* **2011,** *83* (1), 216-224.

39. Mapelli, V.; Olsson L Fau - Nielsen, J.; Nielsen, J., Metabolic footprinting in microbiology: methods and applications in functional genomics and biotechnology. (0167-7799 (Print)).

40. Lauridsen, M.; Hansen Sh Fau - Jaroszewski, J. W.; Jaroszewski Jw Fau - Cornett, C.; Cornett, C., Human urine as test material in 1H NMR-based metabonomics: recommendations for sample preparation and storage. (0003-2700 (Print)).

41. Theobald, U.; Mailinger W Fau - Baltes, M.; Baltes M Fau - Rizzi, M.; Rizzi M Fau - Reuss, M.; Reuss, M., In vivo analysis of metabolic dynamics in Saccharomyces cerevisiae : I. Experimental observations. (0006-3592 (Print)).

42. Ser, Z.; Liu, X.; Tang, N. N.; Locasale, J. W., Extraction parameters for metabolomics from cell extracts. *Analytical biochemistry* **2015,** *475*, 22-28.

43. Dietmair, S.; Timmins, N. E.; Gray, P. P.; Nielsen, L. K.; Kromer, J. O., Towards quantitative metabolomics of mammalian cells: development of a metabolite extraction protocol. *Analytical biochemistry* **2010,** *404* (2), 155-64.

44. El Rammouz, R.; Letisse, F.; Durand, S.; Portais, J. C.; Moussa, Z. W.; Fernandez, X., Analysis of skeletal muscle metabolome: evaluation of extraction methods for targeted metabolite quantification using liquid chromatography tandem mass spectrometry. *Analytical biochemistry* **2010,** *398* (2), 169-77.

45. Masson, P.; Alves Ac Fau - Ebbels, T. M. D.; Ebbels Tm Fau - Nicholson, J. K.; Nicholson Jk Fau - Want, E. J.; Want, E. J., Optimization and evaluation of metabolite extraction protocols for untargeted metabolic profiling of liver samples by UPLC-MS. (1520-6882 (Electronic)).

46. Halket, J. M.; Waterman, D.; Przyborowska, A. M.; Patel, R. K.; Fraser, P. D.; Bramley, P. M., Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS. *Journal of experimental botany* **2005,** *56* (410), 219-43.

47. Vinaixa, M.; Rodriguez, M. A.; Samino, S.; Díaz, M.; Beltran, A.; Mallol, R.; Bladé, C.; Ibañez, L.; Correig, X.; Yanes, O., Metabolomics Reveals Reduction of Metabolic Oxidation in Women with Polycystic Ovary Syndrome after Pioglitazone-Flutamide-Metformin Polytherapy. *PLoS ONE* **2011,** *6* (12), e29052.

48. Verwaest, K. A.; Vu Tn Fau - Laukens, K.; Laukens K Fau - Clemens, L. E.; Clemens Le Fau - Nguyen, H. P.; Nguyen Hp Fau - Van Gasse, B.; Van Gasse B Fau - Martins, J. C.; Martins Jc Fau - Van Der Linden, A.; Van Der Linden A Fau - Dommisse, R.; Dommisse, R., (1)H NMR based metabolomics of CSF and blood serum: a metabolic profile for a transgenic rat model of Huntington disease. (0006-3002 (Print)).

49. Chen, Y.; Zhou, J.; Li, J.; Feng, J.; Chen, Z.; Wang, X., Plasma metabolomic analysis of human hepatocellular carcinoma: Diagnostic and therapeutic study. (1949-2553 (Electronic)).

50. Beckonert, O.; Coen M Fau - Keun, H. C.; Keun Hc Fau - Wang, Y.; Wang Y Fau - Ebbels, T. M. D.; Ebbels Tm Fau - Holmes, E.; Holmes E Fau - Lindon, J. C.; Lindon Jc Fau - Nicholson, J. K.; Nicholson, J. K., High-resolution magic-angle-spinning NMR spectroscopy for metabolic profiling of intact tissues. (1750-2799 (Electronic)).

51. Li, M.; Song Y Fau - Cho, N.; Cho N Fau - Chang, J. M.; Chang Jm Fau - Koo, H. R.; Koo Hr Fau - Yi, A.; Yi A Fau - Kim, H.; Kim H Fau - Park, S.; Park S Fau - Moon, W. K.; Moon, W. K., An HR-MAS MR metabolomics study on breast tissues obtained with core needle biopsy. (1932-6203 (Electronic)).

52. Beltran, A.; Suarez M Fau - Rodriguez, M. A.; Rodriguez Ma Fau - Vinaixa, M.; Vinaixa M Fau - Samino, S.; Samino S Fau - Arola, L.; Arola L Fau - Correig, X.; Correig X Fau - Yanes, O.; Yanes, O., Assessment of compatibility between extraction methods for NMR- and LC/MS-based metabolomics. (1520-6882 (Electronic)).

53. Wishart, D. S.; Feunang, Y. D.; Marcu, A.; Guo, A. C.; Liang, K.; Vazquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; Sayeeda, Z.; Lo, E.; Assempour, N.; Berjanskii, M.; Singhal, S.; Arndt, D.; Liang, Y.; Badran, H.; Grant, J.; Serra-Cayuela, A.; Liu, Y.; Mandal, R.; Neveu, V.; Pon, A.; Knox, C.; Wilson, M.; Manach, C.; Scalbert, A., HMDB 4.0: the human metabolome database for 2018. (1362-4962 (Electronic)).

54. Psychogios, N.; Hau, D. D.; Peng, J.; Guo, A. C.; Mandal, R.; Bouatra, S.; Sinelnikov, I.; Krishnamurthy, R.; Eisner, R.; Gautam, B.; Young, N.; Xia, J.; Knox, C.; Dong, E.; Huang, P.; Hollander, Z.; Pedersen, T. L.; Smith, S. R.; Bamforth, F.; Greiner, R.; McManus, B.; Newman, J. W.; Goodfriend, T.; Wishart, D. S., The Human Serum Metabolome. *PLoS ONE* **2011,** *6* (2), e16957.

55. Alvaro, G.-D.; Enrique, D.-G.; Angeles, F.-R.; Alfonso Maria, L.-S.; Ana, S.; Monica, S.; Carmen, S.; Raul, G.-D., An Overview on the Importance of Combining Complementary Analytical Platforms in Metabolomic Research. *Current Topics in Medicinal Chemistry* **2017,** *17* (30), 3289-3295.

56. Moco, S.; Vervoort, J.; Bino, R. J.; De Vos, R. C. H.; Bino, R., Metabolomics technologies and metabolite identification. *TrAC Trends in Analytical Chemistry* **2007,** *26* (9), 855-866.

57. Churchwell, M. I.; Twaddle, N. C.; Meeker, L. R.; Doerge, D. R., Improving LC–MS sensitivity through increases in chromatographic performance: Comparisons of UPLC–ES/MS/MS to HPLC–ES/MS/MS. *Journal of Chromatography B* **2005,** *825* (2), 134-143.

58. Wilson, I. D.; Nicholson, J. K.; Castro-Perez, J.; Granger, J. H.; Johnson, K. A.; Smith, B. W.; Plumb, R. S., High Resolution "Ultra Performance" Liquid Chromatography Coupled to oa-TOF Mass Spectrometry as a Tool for Differential Metabolic Pathway Profiling in Functional Genomic Studies. *Journal of proteome research* **2005,** *4* (2), 591-598.

59. Dunn, W. B.; Broadhurst D Fau - Begley, P.; Begley P Fau - Zelena, E.; Zelena E Fau - Francis-McIntyre, S.; Francis-McIntyre S Fau - Anderson, N.; Anderson N Fau - Brown, M.; Brown M Fau - Knowles, J. D.; Knowles Jd Fau - Halsall, A.; Halsall A Fau - Haselden, J. N.; Haselden Jn Fau - Nicholls, A. W.; Nicholls Aw Fau - Wilson, I. D.; Wilson Id Fau - Kell, D. B.; Kell Db Fau - Goodacre,

R.; Goodacre, R., Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. (1750-2799 (Electronic)).

60. Want, E. J.; Masson P Fau - Michopoulos, F.; Michopoulos F Fau - Wilson, I. D.; Wilson Id Fau - Theodoridis, G.; Theodoridis G Fau - Plumb, R. S.; Plumb Rs Fau - Shockcor, J.; Shockcor J Fau - Loftus, N.; Loftus N Fau - Holmes, E.; Holmes E Fau - Nicholson, J. K.; Nicholson, J. K., Global metabolic profiling of animal and human tissues via UPLC-MS. (1750-2799 (Electronic)).

61. Gika, H. G.; Theodoridis, G. A.; Plumb, R. S.; Wilson, I. D., Current practice of liquid chromatography–mass spectrometry in metabolomics and metabonomics. *Journal of Pharmaceutical and Biomedical Analysis* **2014,** *87*, 12-25.

62. Gika, H. G.; Theodoridis Ga Fau - Plumb, R. S.; Plumb Rs Fau - Wilson, I. D.; Wilson, I. D., Current practice of liquid chromatography-mass spectrometry in metabolomics and metabonomics. (1873-264X (Electronic)).

63. Yin, P.; Wan D Fau - Zhao, C.; Zhao C Fau - Chen, J.; Chen J Fau - Zhao, X.; Zhao X Fau - Wang, W.; Wang W Fau - Lu, X.; Lu X Fau - Yang, S.; Yang S Fau - Gu, J.; Gu J Fau - Xu, G.; Xu, G., A metabonomic study of hepatitis B-induced liver cirrhosis and hepatocellular carcinoma by using RP-LC and HILIC coupled with mass spectrometry. (1742-2051 (Electronic)).

64. Tang, D. Q.; Zou, L.; Yin, X. X.; Ong, C. N., HILIC-MS for metabolomics: An attractive and complementary approach to RPLC-MS. (1098-2787 (Electronic)).

65. Matyska, M. T.; Pesek Jj Fau - Duley, J.; Duley J Fau - Zamzami, M.; Zamzami M Fau - Fischer, S. M.; Fischer, S. M., Aqueous normal phase retention of nucleotides on silica hydride-based columns: method development strategies for analytes relevant in clinical analysis. (1615-9314 (Electronic)).

66. Contrepois, K.; Jiang, L.; Snyder, M., Optimized Analytical Procedures for the Untargeted Metabolomic Profiling of Human Urine and Plasma by Combining Hydrophilic Interaction (HILIC) and Reverse-Phase Liquid Chromatography (RPLC)-Mass Spectrometry. (1535-9484 (Electronic)).

67. DeJongh, D. C.; Radford, T.; Hribar, J. D.; Hanessian, S.; Bieber, M.; Dawson, G.; Sweeley, C. C., Analysis of trimethylsilyl derivatives of carbohydrates by gas chromatography and mass spectrometry. *Journal of the American Chemical Society* **1969,** *91* (7), 1728-1740.

68. Gelpi, E.; Koenig, W. A.; Gibert, J.; Oró, J., Combined Gas Chromatography-Mass Spectrometry of Amino Acid Derivatives. *Journal of Chromatographic Science* **1969,** *7* (10), 604-613.

69. Brooks, C. J. W.; Horning, E. C.; Young, J. S., Characterization of sterols by gas chromatography-mass spectrometry of the trimethylsilyl ethers. *Lipids* **1968,** *3* (5), 391-402.

70. Gréen, K., Gas chromatography — Mass spectrometry of O-methyloxime derivatives of prostaglandins. *Chemistry and Physics of Lipids* **1969,** *3* (3), 254-272.

71. Niehaus, W. G.; Ryhage, R., Determination of double bond positions in polyunsaturated fatty acids by combination gas chromatography-mass spectrometry. *Analytical Chemistry* **1968,** *40* (12), 1840-1847.

72. Coward, R. F.; Smith, P., The gas chromatography of aromatic acids as their trimethylsilyl derivatives, including applications to urine analysis. *Journal of Chromatography A* **1969,** *45*, 230-243.

73. Fiehn, O., Metabolomics by Gas Chromatography-Mass Spectrometry: Combined Targeted and Untargeted Profiling. *Current protocols in molecular biology* **2016,** *114*, 30 4 1-30 4 32.

74. Krone, N.; Hughes, B. A.; Lavery, G. G.; Stewart, P. M.; Arlt, W.; Shackleton, C. H. L., Gas chromatography/mass spectrometry (GC/MS) remains a pre-eminent discovery tool in clinical steroid investigations even in the era of fast liquid chromatography tandem mass spectrometry (LC/MS/MS). *The Journal of Steroid Biochemistry and Molecular Biology* **2010,** *121* (3), 496-504.

75. Jonsson, P.; Gullberg, J.; Nordström, A.; Kusano, M.; Kowalczyk, M.; Sjöström, M.; Moritz, T., A Strategy for Identifying Differences in Large Series of Metabolomic Samples Analyzed by GC/MS. *Analytical Chemistry* **2004,** *76* (6), 1738-1745.

76. Denkert, C.; Budczies J Fau - Kind, T.; Kind T Fau - Weichert, W.; Weichert W Fau - Tablack, P.; Tablack P Fau - Sehouli, J.; Sehouli J Fau - Niesporek, S.; Niesporek S Fau - Konsgen, D.; Konsgen D Fau - Dietel, M.; Dietel M Fau - Fiehn, O.; Fiehn, O., Mass spectrometry-based metabolic profiling reveals different metabolite patterns in invasive ovarian carcinomas and ovarian borderline tumors. (0008-5472 (Print)).

77. Pasikanti, K. K.; Ho, P. C.; Chan, E. C., Gas chromatography/mass spectrometry in metabolic profiling of biological fluids. *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences* **2008,** *871* (2), 202-11.

78. Ni, Y.; Su M Fau - Qiu, Y.; Qiu Y Fau - Chen, M.; Chen M Fau - Liu, Y.; Liu Y Fau - Zhao, A.; Zhao A Fau - Jia, W.; Jia, W., Metabolic profiling using combined GC-MS and LC-MS provides a systems understanding of aristolochic acid-induced nephrotoxicity in rat. (0014-5793 (Print)).

79. Zhao, J.; Jung, Y.-H.; Jang, C.-G.; Chun, K.-H.; Kwon, S. W.; Lee, J., Metabolomic identification of biochemical changes induced by fluoxetine and imipramine in a chronic mild stress mouse model of depression. *Scientific Reports* **2015,** *5*, 8890.

80. Kind, T.; Tolstikov V Fau - Fiehn, O.; Fiehn O Fau - Weiss, R. H.; Weiss, R. H., A comprehensive urinary metabolomic approach for identifying kidney cancerr. (0003-2697 (Print)).

81. Fancy, S. A.; Beckonert, O.; Darbon, G.; Yabsley, W.; Walley, R.; Baker, D.; Perkins George, L.; Pullen Frank, S.; Rumpel, K., Gas chromatography/flame ionisation detection mass spectrometry for the detection of endogenous urine metabolites for metabonomic studies and its use as a complementary tool to nuclear magnetic resonance spectroscopy. *Rapid Communications in Mass Spectrometry* **2006,** *20* (15), 2271-2280.

82. Lisec, J.; Schauer, N.; Kopka, J.; Willmitzer, L.; Fernie, A. R., Gas chromatography mass spectrometry–based metabolite profiling in plants. *Nature Protocols* **2006,** *1*, 387.

83. Gowda, G. A.; Djukovic, D., Overview of mass spectrometry-based metabolomics: opportunities and challenges. *Methods Mol Biol* **2014,** *1198*, 3-12.

84. Banerjee, S.; Mazumdar, S., Electrospray Ionization Mass Spectrometry: A Technique to Access the Information beyond the Molecular Weight of the Analyte. *International Journal of Analytical Chemistry* **2012,** *2012*, 282574.

85. Major Hilary, J.; Williams, R.; Wilson Amy, J.; Wilson Ian, D., A metabonomic analysis of plasma from Zucker rat strains using gas chromatography/mass spectrometry and pattern recognition. *Rapid Communications in Mass Spectrometry* **2006,** *20* (22), 3295-3302.

86. Buscher, J. M.; Czernik, D.; Ewald, J. C.; Sauer, U.; Zamboni, N., Cross-platform comparison of methods for quantitative

metabolomics of primary metabolism. *Anal Chem* **2009,** *81* (6), 2135-43.

87. Lawton, K. A.; Berger, A.; Mitchell, M.; Milgram, K. E.; Evans, A. M.; Guo, L.; Hanson, R. W.; Kalhan, S. C.; Ryals, J. A.; Milburn, M. V., Analysis of the adult human plasma metabolome. *Pharmacogenomics* **2008,** *9* (4), 383-97.

88. Lindon, J. C.; Holmes, E.; Nicholson, J. K., Peer Reviewed: So What's the Deal with Metabonomics? *Analytical Chemistry* **2003,** *75* (17), 384 A-391 A.

89. Keun, H. C.; Beckonert, O.; Griffin, J. L.; Richter, C.; Moskau, D.; Lindon, J. C.; Nicholson, J. K., Cryogenic Probe 13C NMR Spectroscopy of Urine for Metabonomic Studies. *Analytical Chemistry* **2002,** *74* (17), 4588-4593.

90. Ward Jane, L.; Baker John, M.; Beale Michael, H., Recent applications of NMR spectroscopy in plant metabolomics. *The FEBS Journal* **2007,** *274* (5), 1126-1131.

91. Marion, D., An Introduction to Biological NMR Spectroscopy. *Molecular & Cellular Proteomics : MCP* **2013,** *12* (11), 3006-3025.

92. Katajamaa, M.; Oresic, M., Data processing for mass spectrometry-based metabolomics. *Journal of chromatography. A* **2007,** *1158* (1-2), 318-28.

93. Tautenhahn, R.; Cho, K.; Uritboonthai, W.; Zhu, Z.; Patti, G. J.; Siuzdak, G., An accelerated workflow for untargeted

metabolomics using the METLIN database. *Nature biotechnology* **2012,** *30* (9), 826-8.

94. Yao, Y.; Sun, T.; Wang, T.; Ruebel, O.; Northen, T.; Bowen, P. B., Analysis of Metabolomics Datasets with High-Performance Computing and Metabolite Atlases. *Metabolites* **2015,** *5* (3).

95. Pedrioli, P. G. A.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E. W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; Cheung, K.; Costello, C. E.; Hermjakob, H.; Huang, S.; Julian Jr, R. K.; Kapp, E.; McComb, M. E.; Oliver, S. G.; Omenn, G.; Paton, N. W.; Simpson, R.; Smith, R.; Taylor, C. F.; Zhu, W.; Aebersold, R., A common open representation of mass spectrometry data and its application to proteomics research. *Nature biotechnology* **2004,** *22*, 1459.

96. Hilario, M.; Kalousis, A.; Pellegrini, C.; Müller, M., Processing and classification of protein mass spectra. *Mass Spectrometry Reviews* **2006,** *25* (3), 409-449.

97. Hastings Curtis, A.; Norton Scott, M.; Roy, S.; Siuzdak, G., New algorithms for processing and peak detection in liquid chromatography/mass spectrometry data. *Rapid Communications in Mass Spectrometry* **2002,** *16* (5), 462-467.

98. Haimi, P.; Uphoff, A.; Hermansson, M.; Somerharju, P., Software Tools for Analysis of Mass Spectrometric Lipidome Data. *Analytical Chemistry* **2006,** *78* (24), 8324-8331.

99. Lange, E.; Tautenhahn, R.; Neumann, S.; Gröpl, C., Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics* **2008,** *9*, 375-375.

100. Lange, E.; Knut, R.; Groepl, C.; Kohlbacher, O.; Sturm, M.; Hildebrandt, A., *OPENMS; a generic open source framework for chromatography/MS-based proteomics*. 2005; Vol. 4, p S25-S25.

101. Weisser, H.; Nahnsen, S.; Grossmann, J.; Nilse, L.; Quandt, A.; Brauer, H.; Sturm, M.; Kenar, E.; Kohlbacher, O.; Aebersold, R.; Malmström, L., An Automated Pipeline for High-Throughput Label-Free Quantitative Proteomics. *Journal of proteome research* **2013,** *12* (4), 1628-1644.

102. Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G., XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem* **2006,** *78* (3), 779-87.

103. Benton, H. P.; Wong, D. M.; Trauger, S. A.; Siuzdak, G., XCMS2: Processing Tandem Mass Spectrometry Data for Metabolite Identification and Structural Characterization. *Analytical Chemistry* **2008,** *80* (16), 6382-6389.

104. Eilers, P. H. C., Parametric Time Warping. *Analytical Chemistry* **2004,** *76* (2), 404-411.

105. Wehrens, R.; Bloemberg, T. G.; Eilers, P. H. C., Fast parametric time warping of peak lists. *Bioinformatics* **2015,** *31* (18), 3063-3065.

106. Bloemberg, T. G.; Gerretzen, J.; Wouters, H. J. P.; Gloerich, J.; van Dael, M.; Wessels, H. J. C. T.; van den Heuvel, L. P.; Eilers, P. H. C.; Buydens, L. M. C.; Wehrens, R., Improved parametric time warping for proteomics. *Chemometrics and Intelligent Laboratory Systems* **2010,** *104* (1), 65-74.

107. Crawford, L. R.; Morrison, J. D., Computer methods in analytical mass spectrometry. Identification of an unknown compound in a catalog. *Analytical Chemistry* **1968,** *40* (10), 1464-1469.

108. Scholz, M.; Gatzek, S.; Sterling, A.; Fiehn, O.; Selbig, J., Metabolite fingerprinting: detecting biological features by independent component analysis. *Bioinformatics* **2004,** *20* (15), 2447-2454.

109. Wang, W.; Zhou, H.; Lin, H.; Roy, S.; Shaler, T. A.; Hill, L. R.; Norton, S.; Kumar, P.; Anderle, M.; Becker, C. H., Quantification of Proteins and Metabolites by Mass Spectrometry without Isotopic Labeling or Spiked Standards. *Analytical Chemistry* **2003,** *75* (18), 4818-4826.

110. Orešič, M.; Clish, C. B.; Davidov, E. J.; Verheij, E.; Vogels, J.; Havekes, L. M.; Neumann, E.; Adourian, A.; Naylor, S.; van der Greef, J.; Plasterer, T., Phenotype Characterisation Using Integrated Gene Transcript, Protein and Metabolite Profiling. *Applied Bioinformatics* **2004,** *3* (4), 205-217.

111. Hermansson, M.; Uphoff, A.; Käkelä, R.; Somerharju, P., Automated Quantitative Analysis of Complex Lipidomes by Liquid

Chromatography/Mass Spectrometry. *Analytical Chemistry* **2005,** *77* (7), 2166-2175.

112. Bijlsma, S.; Bobeldijk, I.; Verheij, E. R.; Ramaker, R.; Kochhar, S.; Macdonald, I. A.; van Ommen, B.; Smilde, A. K., Large-Scale Human Metabolomics Studies: A Strategy for Data (Pre-) Processing and Validation. *Analytical Chemistry* **2006,** *78* (2), 567-574.

113. Patti, G. J.; Yanes, O.; Siuzdak, G., Innovation: Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol* **2012,** *13* (4), 263-269.

114. Kind, T.; Fiehn, O., Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. (1471-2105 (Electronic)).

115. Domingo-Almenara, X.; Montenegro-Burke, J. R.; Benton, H. P.; Siuzdak, G., Annotation: A Computational Solution for Streamlining Metabolomics Analysis. *Analytical Chemistry* **2018,** *90* (1), 480-489.

116. Alonso, A.; Julià, A.; Beltran, A.; Vinaixa, M.; Díaz, M.; Ibañez, L.; Correig, X.; Marsal, S., AStream: an R package for annotating LC/MS metabolomic data. *Bioinformatics* **2011,** *27* (9), 1339-1340.

117. Tikunov, Y. M.; Laptenok, S.; Hall, R. D.; Bovy, A.; de Vos, R. C. H., MSClust: a tool for unsupervised mass spectra extraction of chromatography-mass spectrometry ion-wise aligned data. *Metabolomics* **2012,** *8* (4), 714-718.

118. Broeckling, C. D.; Afsar, F. A.; Neumann, S.; Ben-Hur, A.; Prenni, J. E., RAMClust: A Novel Feature Clustering Method Enables Spectral-Matching-Based Annotation for Metabolomics Data. *Analytical Chemistry* **2014,** *86* (14), 6812-6817.

119. DeFelice, B. C.; Mehta, S. S.; Samra, S.; Čajka, T.; Wancewicz, B.; Fahrmann, J. F.; Fiehn, O., Mass Spectral Feature List Optimizer (MS-FLO): A Tool To Minimize False Positive Peak Reports in Untargeted Liquid Chromatography–Mass Spectroscopy (LC-MS) Data Processing. *Analytical Chemistry* **2017,** *89* (6), 3250-3255.

120. Edmands, W. M. B.; Petrick, L.; Barupal, D. K.; Scalbert, A.; Wilson, M. J.; Wickliffe, J. K.; Rappaport, S. M., compMS2Miner: An Automatable Metabolite Identification, Visualization, and Data-Sharing R Package for High-Resolution LC–MS Data Sets. *Analytical Chemistry* **2017,** *89* (7), 3919-3928.

121. Jaeger, C.; Méret, M.; Schmitt Clemens, A.; Lisec, J., Compound annotation in liquid chromatography/high-resolution mass spectrometry based metabolomics: robust adduct ion determination as a prerequisite to structure prediction in electrospray ionization mass spectra. *Rapid Communications in Mass Spectrometry* **2017,** *31* (15), 1261-1266.

122. Kuhl, C.; Tautenhahn, R.; Bottcher, C.; Larson, T. R.; Neumann, S., CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal Chem* **2012,** *84* (1), 283-9.

123. Zhang, W.; Chang, J.; Lei, Z.; Huhman, D.; Sumner, L. W.; Zhao, P. X., MET-COFEA: A Liquid Chromatography/Mass Spectrometry Data Processing Platform for Metabolite Compound Feature Extraction and Annotation. *Analytical Chemistry* **2014,** *86* (13), 6245-6253.

124. Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M., MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **2010,** *11*, 395-395.

125. Daly, R.; Rogers, S.; Wandy, J.; Jankevics, A.; Burgess, K. E. V.; Breitling, R., MetAssign: probabilistic annotation of metabolites from LC–MS data using a Bayesian clustering approach. *Bioinformatics* **2014,** *30* (19), 2764-2771.

126. Uppal, K.; Walker, D. I.; Jones, D. P., xMSannotator: an R package for network-based annotation of high-resolution metabolomics data. *Analytical Chemistry* **2017,** *89* (2), 1063-1067.

127. Koh, Y.; Pasikanti, K. K.; Yap, C. W.; Chan, E. C. Y., Comparative evaluation of software for retention time alignment of gas chromatography/time-of-flight mass spectrometry-based metabonomic data. *Journal of chromatography. A* **2010,** *1217* (52), 8308-8316.

128. Katajamaa, M.; Miettinen, J.; Orešič, M., MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* **2006,** *22* (5), 634-636.

129. (a) Lommen, A.; Kools, H. J., MetAlign 3.0: performance enhancement by efficient use of advances in computer hardware. *Metabolomics* **2012,** *8* (4), 719-726; (b) Lommen, A., MetAlign: Interface-Driven, Versatile Metabolomics Tool for Hyphenated Full-Scan Mass Spectrometry Data Preprocessing. *Analytical Chemistry* **2009,** *81* (8), 3079-3086.

130. Wehrens, R.; Weingart, G.; Mattivi, F., metaMS: An open-source pipeline for GC–MS-based untargeted metabolomics. *Journal of Chromatography B* **2014,** *966*, 109-116.

131. Luedemann, A.; Strassburg, K.; Erban, A.; Kopka, J., TagFinder for the quantitative analysis of gas chromatography—mass spectrometry (GC-MS)-based metabolite profiling experiments. *Bioinformatics* **2008,** *24* (5), 732-737.

132. Hiller, K.; Hangebrauk, J.; Jäger, C.; Spura, J.; Schreiber, K.; Schomburg, D., MetaboliteDetector: Comprehensive Analysis Tool for Targeted and Nontargeted GC/MS Based Metabolome Analysis. *Analytical Chemistry* **2009,** *81* (9), 3429-3439.

133. O'Callaghan, S.; De Souza, D. P.; Isaac, A.; Wang, Q.; Hodkinson, L.; Olshansky, M.; Erwin, T.; Appelbe, B.; Tull, D. L.; Roessner, U.; Bacic, A.; McConville, M. J.; Likić, V. A., PyMS: a Python toolkit for processing of gas chromatography-mass spectrometry (GC-MS) data. Application and comparative study of selected tools. *BMC Bioinformatics* **2012,** *13* (1), 115.

134. Jellema, R. H.; Krishnan, S.; Hendriks, M. M. W. B.; Muilwijk, B.; Vogels, J. T. W. E., Deconvolution using signal

segmentation. *Chemometrics and Intelligent Laboratory Systems* **2010,** *104* (1), 132-139.

135. Ni, Y.; Qiu, Y.; Jiang, W.; Suttlemyre, K.; Su, M.; Zhang, W.; Jia, W.; Du, X., ADAP-GC 2.0: Deconvolution of Coeluting Metabolites from GC/TOF-MS Data for Metabolomics Studies. *Analytical Chemistry* **2012,** *84* (15), 6619-6629.

136. Stein, S. E., An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *Journal of the American Society for Mass Spectrometry* **1999,** *10* (8), 770-781.

137. Skogerson, K.; Wohlgemuth, G.; Barupal, D. K.; Fiehn, O., The volatile compound BinBase mass spectral database. *BMC Bioinformatics* **2011,** *12* (1), 321.

138. Kind, T.; Wohlgemuth, G.; Lee, D. Y.; Lu, Y.; Palazoglu, M.; Shahbaz, S.; Fiehn, O., FiehnLib: Mass Spectral and Retention Index Libraries for Metabolomics Based on Quadrupole and Time-of-Flight Gas Chromatography/Mass Spectrometry. *Analytical Chemistry* **2009,** *81* (24), 10038-10048.

139. Domingo-Almenara, X.; Brezmes, J.; Vinaixa, M.; Samino, S.; Ramirez, N.; Ramon-Krauel, M.; Lerin, C.; Díaz, M.; Ibáñez, L.; Correig, X.; Perera-Lluna, A.; Yanes, O., eRah: A Computational Tool Integrating Spectral Deconvolution and Alignment with Quantification and Identification of Metabolites in GC/MS-Based Metabolomics. *Analytical Chemistry* **2016,** *88* (19), 9821-9829.

140. Creek, D. J.; Dunn, W. B.; Fiehn, O.; Griffin, J. L.; Hall, R. D.; Lei, Z.; Mistrik, R.; Neumann, S.; Schymanski, E. L.; Sumner, L. W.; Trengove, R.; Wolfender, J.-L., Metabolite identification: are you sure? And how do your peers gauge your confidence? *Metabolomics* **2014,** *10* (3), 350-353.

141. Sumner, L. W.; Lei, Z.; Nikolau, B. J.; Saito, K.; Roessner, U.; Trengove, R., Proposed quantitative and alphanumeric metabolite identification metrics. *Metabolomics* **2014,** *10* (6), 1047-1049.

142. Salek, R. M.; Arita, M.; Dayalan, S.; Ebbels, T.; Jones, A. R.; Neumann, S.; Rocca-Serra, P.; Viant, M. R.; Vizcaíno, J.-A., Embedding standards in metabolomics: the Metabolomics Society data standards task group. *Metabolomics* **2015,** *11* (4), 782-783.

143. Goodacre, R.; Broadhurst, D.; Smilde, A. K.; Kristal, B. S.; Baker, J. D.; Beger, R.; Bessant, C.; Connor, S.; Capuani, G.; Craig, A.; Ebbels, T.; Kell, D. B.; Manetti, C.; Newton, J.; Paternostro, G.; Somorjai, R.; Sjöström, M.; Trygg, J.; Wulfert, F., Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics* **2007,** *3* (3), 231-241.

144. Vinaixa, M.; Schymanski, E. L.; Neumann, S.; Navarro, M.; Salek, R. M.; Yanes, O., Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects. *TrAC Trends in Analytical Chemistry* **2016,** *78*, 23-35.

145. Dunn, W. B.; Erban, A.; Weber, R. J. M.; Creek, D. J.; Brown, M.; Breitling, R.; Hankemeier, T.; Goodacre, R.; Neumann, S.; Kopka, J.; Viant, M. R., Mass appeal: metabolite identification in

mass spectrometry-focused untargeted metabolomics. *Metabolomics* **2013,** *9* (1), 44-66.

146. Kind, T.; Scholz, M.; Fiehn, O., How Large Is the Metabolome? A Critical Analysis of Data Exchange Practices in Chemistry. *PLoS ONE* **2009,** *4* (5), e5440.

147. Heinonen, M.; Shen, H.; Zamboni, N.; Rousu, J., Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics* **2012,** *28* (18), 2333-2341.

148. Gerlich, M.; Neumann, S., MetFusion: integration of compound identification strategies. *Journal of Mass Spectrometry* **2013,** *48* (3), 291-298.

149. Allen, F.; Pon, A.; Wilson, M.; Greiner, R.; Wishart, D., CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Research* **2014,** *42* (W1), W94-W99.

150. Ridder, L.; van der Hooft, J. J. J.; Verhoeven, S.; de Vos, R. C. H.; Bino, R. J.; Vervoort, J., Automatic Chemical Structure Annotation of an LC–MSn Based Metabolic Profile from Green Tea. *Analytical Chemistry* **2013,** *85* (12), 6033-6040.

151. Grimme, S., Towards First Principles Calculation of Electron Impact Mass Spectra of Molecules. *Angewandte Chemie International Edition* **2013,** *52* (24), 6306-6312.

152. Fiehn, O.; Barupal, D. K.; Kind, T., Extending Biochemical Databases by Metabolomic Surveys. *The Journal of Biological Chemistry* **2011,** *286* (27), 23637-23643.

153. Ogata, H.; Goto, S.; Sato, K.; Fujibuchi, W.; Bono, H.; Kanehisa, M., KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **1999,** *27* (1), 29-34.

154. Joshi-Tope, G.; Gillespie, M.; Vastrik, I.; D'Eustachio, P.; Schmidt, E.; de Bono, B.; Jassal, B.; Gopinath, G. R.; Wu, G. R.; Matthews, L.; Lewis, S.; Birney, E.; Stein, L., Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research* **2005,** *33* (Database Issue), D428-D432.

155. Kelder, T.; van Iersel, M. P.; Hanspers, K.; Kutmon, M.; Conklin, B. R.; Evelo, C. T.; Pico, A. R., WikiPathways: building research communities on biological pathways. *Nucleic Acids Research* **2012,** *40* (D1), D1301-D1307.

156. Karp, P. D.; Ouzounis, C. A.; Moore-Kochlacs, C.; Goldovsky, L.; Kaipa, P.; Ahrén, D.; Tsoka, S.; Darzentas, N.; Kunin, V.; López-Bigas, N., Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Research* **2005,** *33* (19), 6083-6089.

157. Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H., PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research* **2009,** *37* (suppl_2), W623-W633.

158. Pence, H. E.; Williams, A., ChemSpider: An Online Chemical Information Resource. *Journal of Chemical Education* **2010,** *87* (11), 1123-1124.

159. Schultz, A. W.; Wang, J.; Zhu, Z.-J.; Johnson, C. H.; Patti, G. J.; Siuzdak, G., Liquid Chromatography Quadrupole Time-of-Flight Characterization of Metabolites Guided by the METLIN Database. *Nature Protocols* **2013,** *8* (3), 451-460.

160. Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai Masami, Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T., MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry* **2010,** *45* (7), 703-714.

161. Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; Bouatra, S.; Sinelnikov, I.; Arndt, D.; Xia, J.; Liu, P.; Yallou, F.; Bjorndahl, T.; Perez-Pineiro, R.; Eisner, R.; Allen, F.; Neveu, V.; Greiner, R.; Scalbert, A., HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Research* **2013,** *41* (Database issue), D801-D807.

162. Nicholson, J. K.; Wilson, I. D., Understanding &#39;Global&#39; Systems Biology: Metabonomics and the Continuum of Metabolism. *Nature Reviews Drug Discovery* **2003,** *2,* 668.

163. Wishart, D. S., Quantitative metabolomics using NMR. *TrAC Trends in Analytical Chemistry* **2008,** *27* (3), 228-237.

164. (a) Vinaixa, M.; Ángel Rodríguez, M.; Rull, A.; Beltrán, R.; Bladé, C.; Brezmes, J.; Cañellas, N.; Joven, J.; Correig, X., Metabolomic Assessment of the Effect of Dietary Cholesterol in the Progressive Development of Fatty Liver Disease. *Journal of proteome research* **2010,** *9* (5), 2527-2538; (b) Serkova, N. J.; Niemann, C. U., Pattern recognition and biomarker validation using quantitative 1H-NMR-based metabolomics. *Expert Review of Molecular Diagnostics* **2006,** *6* (5), 717-731.

165. Weljie, A. M.; Newton, J.; Mercier, P.; Carlson, E.; Slupsky, C. M., Targeted Profiling: Quantitative Analysis of 1H NMR Metabolomics Data. *Analytical Chemistry* **2006,** *78* (13), 4430-4442.

166. Serkova, N. J.; Zhang, Y.; Coatney, J. L.; Hunter, L.; Wachs, M. E.; Niemann, C. U.; Mandell, M. S., Early Detection of Graft Failure Using the Blood Metabolic Profile of a Liver Recipient. *Transplantation* **2007,** *83* (4), 517-521.

167. Ellinger, J. J.; Chylla, R. A.; Ulrich, E. L.; Markley, J. L., Databases and Software for NMR-Based Metabolomics. *Current Metabolomics* **2013,** *1* (1), 10.2174/2213235X11301010028.

168. Xia, J.; Psychogios, N.; Young, N.; Wishart, D. S., MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res* **2009,** *37* (Web Server issue), W652-60.

169. Saccenti, E.; Hoefsloot, H. C. J.; Smilde, A. K.; Westerhuis, J. A.; Hendriks, M. M. W. B., Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics* **2014,** *10* (3), 361-374.

170. Vinaixa, M.; Samino, S.; Saez, I.; Duran, J.; Guinovart, J. J.; Yanes, O., A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data. *Metabolites* **2012,** *2* (4), 775-95.

171. Trygg, J.; Holmes, E.; Lundstedt, T., Chemometrics in metabonomics. *Journal of proteome research* **2007,** *6* (2), 469-79.

172. Bjerrum, J. T., Metabonomics: analytical techniques and associated chemometrics at a glance. *Methods Mol Biol* **2015,** *1277*, 1-14.

173. Liland, K. H., Multivariate methods in metabolomics – from pre-processing to dimension reduction and statistical analysis. *TrAC Trends in Analytical Chemistry* **2011,** *30* (6), 827-841.

174. Chagoyen, M.; Lopez-Ibanez, J.; Pazos, F., Functional Analysis of Metabolomics Data. *Methods Mol Biol* **2016,** *1415*, 399-406.

175. Aggio, R. B.; Ruggiero, K.; Villas-Boas, S. G., Pathway Activity Profiling (PAPi): from the metabolite profile to the metabolic pathway activity. *Bioinformatics* **2010,** *26* (23), 2969-76.

176. Chagoyen, M.; Pazos, F., Tools for the functional interpretation of metabolomic experiments. *Briefings in bioinformatics* **2013,** *14* (6), 737-44.

177. Okuda, S.; Yamada, T.; Hamajima, M.; Itoh, M.; Katayama, T.; Bork, P.; Goto, S.; Kanehisa, M., KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res* **2008,** *36* (Web Server issue), W423-6.

178. D'Eustachio, P., Reactome knowledgebase of human biological pathways and processes. *Methods Mol Biol* **2011,** *694*, 49-61.

179. Huang, D. W.; Sherman, B. T.; Lempicki, R. A., Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research* **2009,** *37* (1), 1-13.

180. Aebersold, R.; Mann, M., Mass-spectrometric exploration of proteome structure and function. *Nature* **2016,** *537* (7620), 347-55.

181. Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M.-C.; Yates, J. R., Protein Analysis by Shotgun/Bottom-up Proteomics. *Chemical reviews* **2013,** *113* (4), 2343-2394.

182. Steen, H.; Mann, M., The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol* **2004,** *5* (9), 699-711.

183. Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M. C.; Yates, J. R., 3rd, Protein analysis by shotgun/bottom-up proteomics. *Chemical reviews* **2013,** *113* (4), 2343-94.

184. Bodzon-Kulakowska, A.; Bierczynska-Krzysik, A.; Dylag, T.; Drabik, A.; Suder, P.; Noga, M.; Jarzebinska, J.; Silberring, J., Methods for samples preparation in proteomic research. *Journal of Chromatography B* **2007,** *849* (1), 1-31.

185. Colantonio David, A.; Dunkinson, C.; Bovenkamp Diane, E.; Van Eyk Jennifer, E., Effective removal of albumin from serum. *Proteomics* **2005,** *5* (15), 3831-3835.

186. Kay, R.; Barton, C.; Ratcliffe, L.; Matharoo-Ball, B.; Brown, P.; Roberts, J.; Teale, P.; Creaser, C., Enrichment of low molecular weight serum proteins using acetonitrile precipitation for mass spectrometry based proteomic analysis. *Rapid Communications in Mass Spectrometry* **2008,** *22* (20), 3255-3260.

187. Warder, S. E.; Tucker, L. A.; Strelitzer, T. J.; McKeegan, E. M.; Meuth, J. L.; Jung, P. M.; Saraf, A.; Singh, B.; Lai-Zhang, J.; Gagne, G.; Rogers, J. C., Reducing agent-mediated precipitation of high-abundance plasma proteins. *Analytical biochemistry* **2009,** *387* (2), 184-193.

188. Mahn, A.; Ismail, M., Depletion of highly abundant proteins in blood plasma by ammonium sulfate precipitation for 2D-PAGE analysis. *Journal of Chromatography B* **2011,** *879* (30), 3645-3648.

189. Pieper, R.; Su, Q.; Gatlin Christine, L.; Huang, S. T.; Anderson, N. L.; Steiner, S., Multi-component immunoaffinity subtraction chromatography: An innovative step towards a comprehensive survey of the human plasma proteome. *Proteomics* **2003,** *3* (4), 422-432.

190. Listgarten, J.; Emili, A., Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol Cell Proteomics* **2005,** *4* (4), 419-34.

191. Peng, J.; Gygi, S. P., Proteomics: the move to mixtures. *Journal of mass spectrometry : JMS* **2001,** *36* (10), 1083-91.

192. Kennedy, R. T.; Jorgenson, J. W., Preparation and evaluation of packed capillary liquid chromatography columns with inner diameters from 20 to 50 micrometers. *Analytical Chemistry* **1989,** *61* (10), 1128-1135.

193. Emmett, M. R.; Caprioli, R. M., Micro-electrospray mass spectrometry: Ultra-high-sensitivity analysis of peptides and proteins. *Journal of the American Society for Mass Spectrometry* **1994,** *5* (7), 605-613.

194. Hsieh, S.; Jorgenson, J. W., Preparation and Evaluation of Slurry-Packed Liquid Chromatography Microcolumns with Inner Diameters from 12 to 33 μm. *Analytical Chemistry* **1996,** *68* (7), 1212-1217.

195. Ishihama, Y., Proteomic LC–MS systems using nanoscale liquid chromatography with tandem mass spectrometry. *Journal of Chromatography A* **2005,** *1067* (1), 73-83.

196. Mitulovic, G.; Mechtler, K., HPLC techniques for proteomics analysis--a short overview of latest developments. *Briefings in functional genomics & proteomics* **2006,** *5* (4), 249-60.

197. Aebersold, R.; Mann, M., Mass spectrometry-based proteomics. *Nature* **2003,** *422*, 198.

198. Kislinger, T.; Emili, A., Going global: protein expression profiling using shotgun mass spectrometry. *Curr Opin Mol Ther* **2003,** *5* (3), 285-293.

199. Kim, M.-S.; Pandey, A., Electron Transfer Dissociation Mass Spectrometry in Proteomics. *Proteomics* **2012,** *12* (0), 530-542.

200. Beck, M.; Schmidt, A.; Malmstroem, J.; Claassen, M.; Ori, A.; Szymborska, A.; Herzog, F.; Rinner, O.; Ellenberg, J.; Aebersold, R., The quantitative proteome of a human cell line. *Molecular Systems Biology* **2011,** *7*, 549-549.

201. Bantscheff, M.; Lemeer, S.; Savitski, M. M.; Kuster, B., Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Analytical and bioanalytical chemistry* **2012,** *404* (4), 939-65.

202. Zubarev, R. A.; Makarov, A., Orbitrap Mass Spectrometry. *Analytical Chemistry* **2013,** *85* (11), 5288-5296.

203. Eng, J. K.; McCormack, A. L.; Yates, J. R., An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* **1994,** *5* (11), 976-989.

204. Perkins David, N.; Pappin Darryl, J. C.; Creasy David, M.; Cottrell John, S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *ELECTROPHORESIS* **1999,** *20* (18), 3551-3567.

205. Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R., A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry. *Analytical Chemistry* **2003,** *75* (17), 4646-4658.

206. Nesvizhskii, A. I.; Aebersold, R., Analysis, statistical validation and dissemination of large-scale proteomics datasets

generated by tandem MS. *Drug discovery today* **2004,** *9* (4), 173-181.

207. Nesvizhskii, A. I.; Aebersold, R., Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* **2005,** *4* (10), 1419-40.

208. The UniProt Consortium, UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **2017,** *45* (D1), D158-D169.

209. Pruitt, K. D.; Tatusova, T.; Maglott, D. R., NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **2007,** *35* (Database issue), D61-D65.

210. Cunningham, F.; Amode, M. R.; Barrell, D.; Beal, K.; Billis, K.; Brent, S.; Carvalho-Silva, D.; Clapham, P.; Coates, G.; Fitzgerald, S.; Gil, L.; Girón, C. G.; Gordon, L.; Hourlier, T.; Hunt, S. E.; Janacek, S. H.; Johnson, N.; Juettemann, T.; Kähäri, A. K.; Keenan, S.; Martin, F. J.; Maurel, T.; McLaren, W.; Murphy, D. N.; Nag, R.; Overduin, B.; Parker, A.; Patricio, M.; Perry, E.; Pignatelli, M.; Riat, H. S.; Sheppard, D.; Taylor, K.; Thormann, A.; Vullo, A.; Wilder, S. P.; Zadissa, A.; Aken, B. L.; Birney, E.; Harrow, J.; Kinsella, R.; Muffato, M.; Ruffier, M.; Searle, S. M. J.; Spudich, G.; Trevanion, S. J.; Yates, A.; Zerbino, D. R.; Flicek, P., Ensembl 2015. *Nucleic Acids Research* **2015,** *43* (Database issue), D662-D669.

211. Maglott, D.; Ostell, J.; Pruitt, K. D.; Tatusova, T., Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research* **2005,** *33* (Database Issue), D54-D58.

212. Zhu, W.; Smith, J. W.; Huang, C. M., Mass spectrometry-based label-free quantitative proteomics. *Journal of biomedicine & biotechnology* **2010,** *2010*, 840518.

213. Ong, S.-E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M., Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Molecular & Cellular Proteomics* **2002,** *1* (5), 376-386.

214. Fairwell, T.; Barnes, W. T.; Richards, F. F.; Lovins, R. E., Sequence analysis of complex protein mixtures by isotope dilution and mass spectrometry. *Biochemistry* **1970,** *9* (11), 2260-2267.

215. Sirlin, J. L., ON THE INCORPORATION OF METHIONINE 35S INTO PROTEINS DETECTABLE BY AUTORADIOGRAPHY. *Journal of Histochemistry & Cytochemistry* **1958,** *6* (3), 185-190.

216. Gerber, S. A.; Rush, J.; Stemman, O.; Kirschner, M. W.; Gygi, S. P., Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proceedings of the National Academy of Sciences* **2003,** *100* (12), 6940.

217. Thompson, A.; Schäfer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Hamon, C., Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. *Analytical Chemistry* **2003,** *75* (8), 1895-1904.

218. Wiese, S.; Reidegeld Kai, A.; Meyer Helmut, E.; Warscheid, B., Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research. *Proteomics* **2007,** *7* (3), 340-350.

219. Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J., Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents. *Molecular & Cellular Proteomics* **2004,** *3* (12), 1154-1169.

220. Patel, V. J.; Thalassinos, K.; Slade, S. E.; Connolly, J. B.; Crombie, A.; Murrell, J. C.; Scrivens, J. H., A comparison of labeling and label-free mass spectrometry-based proteomics approaches. *Journal of proteome research* **2009,** *8* (7), 3752-9.

221. Bantscheff, M.; Schirle, M.; Sweetman, G.; Rick, J.; Kuster, B., Quantitative mass spectrometry in proteomics: a critical review. *Analytical and bioanalytical chemistry* **2007,** *389* (4), 1017-1031.

222. Wang, M.; You, J.; Bemis, K. G.; Tegeler, T. J.; Brown, D. P. G., Label-free mass spectrometry-based protein quantification technologies in proteomic analysis. *Briefings in Functional Genomics* **2008,** *7* (5), 329-339.

223. Deracinois, B.; Flahaut, C.; Duban-Deweer, S.; Karamanos, Y., Comparative and Quantitative Global Proteomics Approaches: An Overview. *Proteomes* **2013,** *1* (3).

224. Urfer, W.; Grzegorczyk, M.; Jung, K., Statistics for proteomics: a review of tools for analyzing experimental data. *Proteomics* **2006,** *6 Suppl 2*, 48-55.

225. Kammers, K.; Cole, R. N.; Tiengwe, C.; Ruczinski, I., Detecting significant changes in protein abundance. *EuPA Open Proteomics* **2015,** *7*, 11-19.

226. von Mering, C.; Krause, R.; Snel, B.; Cornell, M.; Oliver, S. G.; Fields, S.; Bork, P., Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **2002,** *417* (6887), 399-403.

227. Spirin, V.; Mirny, L. A., Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences* **2003,** *100* (21), 12123.

228. Ma, X.; Gao, L., Biological network analysis: insights into structure and functions. *Briefings in Functional Genomics* **2012,** *11* (6), 434-442.

229. Brohée, S.; van Helden, J., Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* **2006,** *7* (1), 488.

230. Wang, J.; Li, M.; Deng, Y.; Pan, Y., Recent advances in clustering methods for protein interaction networks. *BMC Genomics* **2010,** *11* (3), S10.

231. Sanz-Pamplona, R.; Berenguer, A.; Sole, X.; Cordero, D.; Crous-Bou, M.; Serra-Musach, J.; Guinó, E.; Pujana, M. Á.; Moreno,

V., Tools for protein-protein interaction network analysis in cancer research. *Clinical and Translational Oncology* **2012,** *14* (1), 3-14.

232. Wu, T.; Qiao, S.; Shi, C.; Wang, S.; Ji, G., Metabolomics window into diabetic complications. *Journal of Diabetes Investigation* **2018,** *9* (2), 244-255.

233. Wu, T.; Qiao, S.; Shi, C.; Wang, S.; Ji, G., The metabolomics window into diabetic complications. *Journal of Diabetes Investigation,* n/a-n/a.

234. Kobrin Klein, B. E., Overview of Epidemiologic Studies of Diabetic Retinopathy. *Ophthalmic Epidemiology* **2007,** *14* (4), 179-183.

235. Yau, J. W. Y.; Rogers, S. L.; Kawasaki, R.; Lamoureux, E. L.; Kowalski, J. W.; Bek, T.; Chen, S.-J.; Dekker, J. M.; Fletcher, A.; Grauslund, J.; Haffner, S.; Hamman, R. F.; Ikram, M. K.; Kayama, T.; Klein, B. E. K.; Klein, R.; Krishnaiah, S.; Mayurasakorn, K.; O'Hare, J. P.; Orchard, T. J.; Porta, M.; Rema, M.; Roy, M. S.; Sharma, T.; Shaw, J.; Taylor, H.; Tielsch, J. M.; Varma, R.; Wang, J. J.; Wang, N.; West, S.; Xu, L.; Yasuda, M.; Zhang, X.; Mitchell, P.; Wong, T. Y., Global Prevalence and Major Risk Factors of Diabetic Retinopathy. *Diabetes Care* **2012,** *35* (3), 556.

236. Joussen, A. M.; Smyth, N.; Niessen, C., Pathophysiology of diabetic macular edema. *Developments in ophthalmology* **2007,** *39,* 1-12.

237. Simo, R.; Carrasco, E.; Garcia-Ramirez, M.; Hernandez, C., Angiogenic and antiangiogenic factors in proliferative diabetic retinopathy. *Current diabetes reviews* **2006,** *2* (1), 71-98.

238. Simó, R.; Villarroel, M.; Corraliza, L.; Hernández, C.; Garcia-Ramírez, M., The Retinal Pigment Epithelium: Something More than a Constituent of the Blood-Retinal Barrier—Implications for the Pathogenesis of Diabetic Retinopathy. *Journal of Biomedicine and Biotechnology* **2010,** *2010*, 190724.

239. Stitt, A. W.; Curtis, T. M.; Chen, M.; Medina, R. J.; McKay, G. J.; Jenkins, A.; Gardiner, T. A.; Lyons, T. J.; Hammes, H.-P.; Simó, R.; Lois, N., The progress in understanding and treatment of diabetic retinopathy. *Progress in Retinal and Eye Research* **2016,** *51*, 156-186.

240. Wilkinson, C. P.; Ferris, F. L., III; Klein, R. E.; Lee, P. P.; Agardh, C. D.; Davis, M.; Dills, D.; Kampik, A.; Pararajasegaram, R.; Verdaguer, J. T., Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology* **2003,** *110* (9), 1677-1682.

241. Stefánsson, E.; Bek, T.; Porta, M.; Larsen, N.; Kristinsson Jóhannes, K.; Agardh, E., Screening and prevention of diabetic blindness. *Acta Ophthalmologica Scandinavica* **2001,** *78* (4), 374-385.

242. Scanlon, P. H.; Foy, C.; Chen, F. K., Visual acuity measurement and ocular co-morbidity in diabetic retinopathy screening. *British Journal of Ophthalmology* **2008,** *92* (6), 775.

243. Klein, R., Barriers to prevention of vision loss caused by diabetic retinopathy. *Archives of Ophthalmology* **1997,** *115* (8), 1073-1075.

244. Subedi, S.; Subedi, K. U.; Badhu, B. P., Doctor's role in early detection of diabetic retinopathy and prevention of blindness from its complications. *JNMA J Nepal Med Assoc* **2005,** *44* (157), 26-30.

245. Liew, G.; Lei, Z.; Tan, G.; Joachim, N.; Ho, I. V.; Wong, T. Y.; Mitchell, P.; Gopinath, B.; Crossett, B., Metabolomics of Diabetic Retinopathy. *Current Diabetes Reports* **2017,** *17* (11), 102.

246. Pallares-Méndez, R.; Aguilar-Salinas, C. A.; Cruz-Bautista, I.; del Bosque-Plata, L., Metabolomics in diabetes, a review. *Annals of Medicine* **2016,** *48* (1-2), 89-102.

247. Zhao, X.; Fritsche, J.; Wang, J.; Chen, J.; Rittig, K.; Schmitt-Kopplin, P.; Fritsche, A.; Häring, H.-U.; Schleicher, E. D.; Xu, G.; Lehmann, R., Metabonomic fingerprints of fasting plasma and spot urine reveal human pre-diabetic metabolic traits. *Metabolomics* **2010,** *6* (3), 362-374.

248. Koves, T. R.; Ussher, J. R.; Noland, R. C.; Slentz, D.; Mosedale, M.; Ilkayeva, O.; Bain, J.; Stevens, R.; Dyck, J. R. B.; Newgard, C. B.; Lopaschuk, G. D.; Muoio, D. M., Mitochondrial Overload and Incomplete Fatty Acid Oxidation Contribute to Skeletal Muscle Insulin Resistance. *Cell metabolism* **2008,** *7* (1), 45-56.

249. Batch, B. C.; Shah, S. H.; Newgard, C. B.; Turer, C. B.; Haynes, C.; Bain, J. R.; Muehlbauer, M.; Patel, M. J.; Stevens, R. D.; Appel, L. J.; Newby, L. K.; Svetkey, L. P., Branched chain amino acids are novel biomarkers for discrimination of metabolic wellness. *Metabolism: clinical and experimental* **2013,** *62* (7), 961-969.

250. Lerin, C.; Goldfine, A. B.; Boes, T.; Liu, M.; Kasif, S.; Dreyfuss, J. M.; De Sousa-Coelho, A. L.; Daher, G.; Manoli, I.; Sysol, J. R.; Isganaitis, E.; Jessen, N.; Goodyear, L. J.; Beebe, K.; Gall, W.; Venditti, C. P.; Patti, M.-E., Defects in muscle branched-chain amino acid oxidation contribute to impaired lipid metabolism. *Molecular Metabolism* **2016,** *5* (10), 926-936.

251. Wang, T. J.; Larson, M. G.; Vasan, R. S.; Cheng, S.; Rhee, E. P.; McCabe, E.; Lewis, G. D.; Fox, C. S.; Jacques, P. F.; Fernandez, C.; O’Donnell, C. J.; Carr, S. A.; Mootha, V. K.; Florez, J. C.; Souza, A.; Melander, O.; Clish, C. B.; Gerszten, R. E., Metabolite Profiles and the Risk of Developing Diabetes. *Nature medicine* **2011,** *17* (4), 448-453.

252. Newgard, C. B.; An, J.; Bain, J. R.; Muehlbauer, M. J.; Stevens, R. D.; Lien, L. F.; Haqq, A. M.; Shah, S. H.; Arlotto, M.; Slentz, C. A.; Rochon, J.; Gallup, D.; Ilkayeva, O.; Wenner, B. R.; Yancy, W. E.; Eisenson, H.; Musante, G.; Surwit, R.; Millington, D. S.; Butler, M. D.; Svetkey, L. P., A Branched-Chain Amino Acid-Related Metabolic Signature that Differentiates Obese and Lean Humans and Contributes to Insulin Resistance. *Cell metabolism* **2009,** *9* (4), 311-326.

253. Zhang, Y.; Guo, K.; LeBlanc, R. E.; Loh, D.; Schwartz, G. J.; Yu, Y.-H., Increasing Dietary Leucine Intake Reduces Diet-Induced Obesity and Improves Glucose and Cholesterol Metabolism in Mice via Multimechanisms. *Diabetes* **2007,** *56* (6), 1647.

254. Wang, T. J.; Ngo, D.; Psychogios, N.; Dejam, A.; Larson, M. G.; Vasan, R. S.; Ghorbani, A.; O'Sullivan, J.; Cheng, S.; Rhee, E. P.; Sinha, S.; McCabe, E.; Fox, C. S.; O'Donnell, C. J.; Ho, J. E.; Florez, J. C.; Magnusson, M.; Pierce, K. A.; Souza, A. L.; Yu, Y.; Carter, C.; Light, P. E.; Melander, O.; Clish, C. B.; Gerszten, R. E., 2-Aminoadipic acid is a biomarker for diabetes risk. *The Journal of clinical investigation* **2013,** *123* (10), 4309-4317.

255. Pedersen, H. K.; Gudmundsdottir, V.; Nielsen, H. B.; Hyotylainen, T.; Nielsen, T.; Jensen, B. A. H.; Forslund, K.; Hildebrand, F.; Prifti, E.; Falony, G.; Le Chatelier, E.; Levenez, F.; Doré, J.; Mattila, I.; Plichta, D. R.; Pöhö, P.; Hellgren, L. I.; Arumugam, M.; Sunagawa, S.; Vieira-Silva, S.; Jørgensen, T.; Holm, J. B.; Trošt, K.; Consortium, M.; Kristiansen, K.; Brix, S.; Raes, J.; Wang, J.; Hansen, T.; Bork, P.; Brunak, S.; Oresic, M.; Ehrlich, S. D.; Pedersen, O., Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature* **2016,** *535*, 376.

256. Rhee, E. P.; Cheng, S.; Larson, M. G.; Walford, G. A.; Lewis, G. D.; McCabe, E.; Yang, E.; Farrell, L.; Fox, C. S.; O'Donnell, C. J.; Carr, S. A.; Vasan, R. S.; Florez, J. C.; Clish, C. B.; Wang, T. J.; Gerszten, R. E., Lipid profiling identifies a triacylglycerol signature of insulin resistance and improves diabetes

prediction in humans. *The Journal of clinical investigation* **2011,** *121* (4), 1402-1411.

257. Newgard, Christopher B., Interplay between Lipids and Branched-Chain Amino Acids in Development of Insulin Resistance. *Cell metabolism* **2012,** *15* (5), 606-614.

258. Paris, L. P.; Johnson, C. H.; Aguilar, E.; Usui, Y.; Cho, K.; Hoang, L. T.; Feitelberg, D.; Benton, H. P.; Westenskow, P. D.; Kurihara, T.; Trombley, J.; Tsubota, K.; Ueda, S.; Wakabayashi, Y.; Patti, G. J.; Ivanisevic, J.; Siuzdak, G.; Friedlander, M., Global metabolomics reveals metabolic dysregulation in ischemic retinopathy. *Metabolomics* **2015,** *12* (1), 15.

259. Barba, I.; Garcia-Ramírez, M.; Hernández, C.; Alonso, M. A.; Masmiquel, L.; García-Dorado, D.; Simó, R., Metabolic Fingerprints of Proliferative Diabetic Retinopathy: An 1H-NMR–Based Metabonomic Approach Using Vitreous Humor. *Investigative ophthalmology & visual science* **2010,** *51* (9), 4416-4421.

260. Young, S. P.; Nessim, M.; Falciani, F.; Trevino, V.; Banerjee, S. P.; Scott, R. A. H.; Murray, P. I.; Wallace, G. R., Metabolomic analysis of human vitreous humor differentiates ocular inflammatory disease. *Molecular Vision* **2009,** *15*, 1210-1217.

261. Locci, E.; Scano, P.; Rosa, M. F.; Nioi, M.; Noto, A.; Atzori, L.; Demontis, R.; De-Giorgio, F.; d'Aloja, E., A Metabolomic Approach to Animal Vitreous Humor Topographical Composition: A Pilot Study. *PLoS ONE* **2014,** *9* (5), e97773.

262. Chen, L.; Cheng, C.-Y.; Choi, H.; Ikram, M. K.; Sabanayagam, C.; Tan, G. S. W.; Tian, D.; Zhang, L.; Venkatesan, G.; Tai, E. S.; Wang, J. J.; Mitchell, P.; Cheung, C. M. G.; Beuerman, R. W.; Zhou, L.; Chan, E. C. Y.; Wong, T. Y., Plasma Metabonomic Profiling of Diabetic Retinopathy. *Diabetes* **2016,** *65* (4), 1099.

263. Lehmann, M.; Yanes, O.; Krohne, T. U.; Dorsey, A. L.; Aguilar, E.; Marchetti, V.; Moreno, S. K.; Trombley, J.; Siuzdak, G.; Friedlander, M., Metabolomic Analysis of Serum from Diabetic Patients With and Without Retinopathy. *Investigative ophthalmology & visual science* **2011,** *52* (14), 3563-3563.

264. Li, X.; Luo, X.; Lu, X.; Duan, J.; Xu, G., Metabolomics study of diabetic retinopathy using gas chromatography-mass spectrometry: a comparison of stages and subtypes diagnosed by Western and Chinese medicine. *Molecular bioSystems* **2011,** *7* (7), 2228-2237.

265. Ferrucci, L.; Cherubini, A.; Bandinelli, S.; Bartali, B.; Corsi, A.; Lauretani, F.; Martin, A.; Andres-Lacueva, C.; Senin, U.; Guralnik, J. M., Relationship of Plasma Polyunsaturated Fatty Acids to Circulating Inflammatory Markers. *The Journal of Clinical Endocrinology & Metabolism* **2006,** *91* (2), 439-446.

266. Gao, B.-B.; Chen, X.; Timothy, N.; Aiello, L. P.; Feener, E. P., Characterization of the Vitreous Proteome in Diabetes without Diabetic Retinopathy and Diabetes with Proliferative Diabetic Retinopathy. *Journal of proteome research* **2008,** *7* (6), 2516-2525.

267. Loukovaara, S.; Nurkkala, H.; Tamene, F.; Gucciardo, E.; Liu, X.; Repo, P.; Lehti, K.; Varjosalo, M., Quantitative Proteomics Analysis of Vitreous Humor from Diabetic Retinopathy Patients. *Journal of proteome research* **2015,** *14* (12), 5131-5143.

268. Gerl, V. B.; Bohl, J. r.; Pitz, S.; Stoffelns, B.; Pfeiffer, N.; Bhakdi, S., Extensive Deposits of Complement C3d and C5b-9 in the Choriocapillaris of Eyes of Patients with Diabetic Retinopathy. *Investigative ophthalmology & visual science* **2002,** *43* (4), 1104-1108.

269. Grigsby, J. G.; Cardona, S. M.; Pouw, C. E.; Muniz, A.; Mendiola, A. S.; Tsin, A. T. C.; Allen, D. M.; Cardona, A. E., The Role of Microglia in Diabetic Retinopathy. *Journal of Ophthalmology* **2014,** *2014,* 15.

270. National Institute of Standards and Technology, NIST/EPA/NIH Mass Spectral Library v2014, US Secretary of Commerce, Gaithersburg, Maryland, USA. 2014.

271. Kessler, N.; Walter, F.; Persicke, M.; Albaum, S. P.; Kalinowski, J.; Goesmann, A.; Niehaus, K.; Nattkemper, T. W., ALLocator: An Interactive Web Platform for the Analysis of Metabolomic LC-ESI-MS Datasets, Enabling Semi-Automated, User-Revised Compound Annotation and Mass Isotopomer Ratio Analysis. *PLoS ONE* **2014,** *9* (11), e113909.

272. Hennige, A. M.; Burks, D. J.; Ozcan, U.; Kulkarni, R. N.; Ye, J.; Park, S.; Schubert, M.; Fisher, T. L.; Dow, M. A.; Leshan, R.; Zakaria, M.; Mossa-Basha, M.; White, M. F., Upregulation of insulin

receptor substrate-2 in pancreatic beta cells prevents diabetes. *The Journal of clinical investigation* **2003,** *112* (10), 1521-32.

273. Withers, D. J.; Gutierrez, J. S.; Towery, H.; Burks, D. J.; Ren, J. M.; Previs, S.; Zhang, Y.; Bernal, D.; Pons, S.; Shulman, G. I.; Bonner-Weir, S.; White, M. F., Disruption of IRS-2 causes type 2 diabetes in mice. *Nature* **1998,** *391* (6670), 900-4.

274. Ruttkies, C.; Schymanski, E. L.; Wolf, S.; Hollender, J.; Neumann, S., MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *Journal of Cheminformatics* **2016,** *8* (1), 3.

275. Tsugawa, H.; Kind, T.; Nakabayashi, R.; Yukihira, D.; Tanaka, W.; Cajka, T.; Saito, K.; Fiehn, O.; Arita, M., Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. *Analytical Chemistry* **2016,** *88* (16), 7946-7958.

276. Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S., Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of the National Academy of Sciences of the United States of America* **2015,** *112* (41), 12580-12585.

277. Graves, P. R.; Haystead, T. A. J., Molecular Biologist's Guide to Proteomics. *Microbiology and Molecular Biology Reviews* **2002,** *66* (1), 39-63.

278. Aittokallio, T.; Schwikowski, B., Graph-based methods for analysing networks in cell biology. *Briefings in bioinformatics* **2006,** *7* (3), 243-55.

279. Guo, Z.; Li, Y.; Gong, X.; Yao, C.; Ma, W.; Wang, D.; Li, Y.; Zhu, J.; Zhang, M.; Yang, D.; Wang, J., Edge-based scoring and searching method for identifying condition-responsive protein–protein interaction sub-network. *Bioinformatics* **2007,** *23* (16), 2121-2128.

280. Ideker, T.; Ozier, O.; Schwikowski, B.; Siegel, A. F., Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **2002,** *18* (suppl_1), S233-S240.

281. Cho, H.; Berger, B.; Peng, J., Compact Integration of Multi-Network Topology for Functional Analysis of Genes. *Cell Systems 3* (6), 540-548.e5.

282. Gibbs, D. L.; Baratt, A.; Baric, R. S.; Kawaoka, Y.; Smith, R. D.; Orwoll, E. S.; Katze, M. G.; McWeeney, S. K., Protein co-expression network analysis (ProCoNA). *Journal of Clinical Bioinformatics* **2013,** *3* (1), 11.

283. Reckow, S.; Gormanns, P.; Holsboer, F.; Turck, C., Psychiatric disorders biomarker identification: from proteomics to systems biology. *Pharmacopsychiatry* **2008,** *41* (S 01), S70-S77.

284. Subramanian, A.; Tamayo, P.; Mootha, V. K.; Mukherjee, S.; Ebert, B. L.; Gillette, M. A.; Paulovich, A.; Pomeroy, S. L.; Golub, T. R.; Lander, E. S.; Mesirov, J. P., Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **2005,** *102* (43), 15545-15550.

285. Grindrod, P.; Kibble, M., Review of uses of network and graph theory concepts within proteomics. *Expert Review of Proteomics* **2004,** *1* (2), 229-238.

286. Munoz, J.; Low, T. Y.; Kok, Y. J.; Chin, A.; Frese, C. K.; Ding, V.; Choo, A.; Heck, A. J. R., The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. *Molecular Systems Biology* **2011,** *7* (1).

287. Cristobal, A.; Hennrich, M. L.; Giansanti, P.; Goerdayal, S. S.; Heck, A. J.; Mohammed, S., In-house construction of a UHPLC system enabling the identification of over 4000 protein groups in a single analysis. *The Analyst* **2012,** *137* (15), 3541-8.

288. Cox, J.; Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology* **2008,** *26*, 1367.

289. Reimand, J.; Arak, T.; Vilo, J., g:Profiler—a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Research* **2011,** *39* (suppl_2), W307-W315.

290. Rolland, T.; Taşan, M.; Charloteaux, B.; Pevzner, Samuel J.; Zhong, Q.; Sahni, N.; Yi, S.; Lemmens, I.; Fontanillo, C.; Mosca, R.; Kamburov, A.; Ghiassian, Susan D.; Yang, X.; Ghamsari, L.; Balcha, D.; Begg, Bridget E.; Braun, P.; Brehme, M.; Broly, Martin P.; Carvunis, A.-R.; Convery-Zupan, D.; Corominas, R.; Coulombe-Huntington, J.; Dann, E.; Dreze, M.; Dricot, A.; Fan, C.; Franzosa, E.; Gebreab, F.; Gutierrez, Bryan J.; Hardy, Madeleine F.; Jin, M.;

Kang, S.; Kiros, R.; Lin, Guan N.; Luck, K.; MacWilliams, A.; Menche, J.; Murray, Ryan R.; Palagi, A.; Poulin, Matthew M.; Rambout, X.; Rasla, J.; Reichert, P.; Romero, V.; Ruyssinck, E.; Sahalie, Julie M.; Scholz, A.; Shah, Akash A.; Sharma, A.; Shen, Y.; Spirohn, K.; Tam, S.; Tejeda, Alexander O.; Trigg, Shelly A.; Twizere, J.-C.; Vega, K.; Walsh, J.; Cusick, Michael E.; Xia, Y.; Barabási, A.-L.; Iakoucheva, Lilia M.; Aloy, P.; De Las Rivas, J.; Tavernier, J.; Calderwood, Michael A.; Hill, David E.; Hao, T.; Roth, Frederick P.; Vidal, M., A Proteome-Scale Map of the Human Interactome Network. *Cell* **2014,** *159* (5), 1212-1226.

291. Airoldi, E. M.; Blei, D. M.; Fienberg, S. E.; Xing, E. P., Mixed Membership Stochastic Blockmodels. *Journal of machine learning research : JMLR* **2008,** *9*, 1981-2014.

292. Vallès-Català, T.; Massucci, F. A.; Guimerà, R.; Sales-Pardo, M., Multilayer Stochastic Block Models Reveal the Multilayer Structure of Complex Networks. *Physical Review X* **2016,** *6* (1), 011036.

293. Altelaar, A. F. M.; Munoz, J.; Heck, A. J. R., Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature Reviews Genetics* **2012,** *14*, 35.

294. Dias, M. H.; Kitano, E. S.; Zelanis, A.; Iwai, L. K., Proteomics and drug discovery in cancer. *Drug discovery today* **2016,** *21* (2), 264-277.

295. Ebhardt, H. A.; Root, A.; Sander, C.; Aebersold, R., Applications of targeted proteomics in systems biology and translational medicine. *Proteomics* **2015,** *15* (18), 3193-3208.

296. Shi, T.; Song, E.; Nie, S.; Rodland, K. D.; Liu, T.; Qian, W.-J.; Smith, R. D., Advances in targeted proteomics and applications to biomedical research. *Proteomics* **2016,** *16* (15-16), 2160-2182.

297. Gygi, S. P.; Aebersold, R., Mass spectrometry and proteomics. *Current Opinion in Chemical Biology* **2000,** *4* (5), 489-494.

298. Swainston, N.; Smallbone, K.; Hefzi, H.; Dobson, P. D.; Brewer, J.; Hanscho, M.; Zielinski, D. C.; Ang, K. S.; Gardiner, N. J.; Gutierrez, J. M.; Kyriakopoulos, S.; Lakshmanan, M.; Li, S.; Liu, J. K.; Martínez, V. S.; Orellana, C. A.; Quek, L.-E.; Thomas, A.; Zanghellini, J.; Borth, N.; Lee, D.-Y.; Nielsen, L. K.; Kell, D. B.; Lewis, N. E.; Mendes, P., Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics* **2016,** *12*, 109.

299. Kanehisa, M.; Goto, S., KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **2000,** *28* (1), 27-30.

300. Haggart, C. R.; Bartell, J. A.; Saucerman, J. J.; Papin, J. A., Whole-genome metabolic network reconstruction and constraint-based modeling. *Methods in Enzymology* **2011,** *500*, 411-433.

301. Ghosh, S.; Baloni, P.; Vishveshwara, S.; Chandra, N., Weighting schemes in metabolic graphs for identifying biochemical routes. *Systems and Synthetic Biology* **2014,** *8* (1), 47-57.

302. Madhukar, N. S.; Warmoes, M. O.; Locasale, J. W., Organization of Enzyme Concentration across the Metabolic Network in Cancer Cells. *PLoS ONE* **2015,** *10* (1), e0117131.

303. Croft, D.; O'Kelly, G.; Wu, G.; Haw, R.; Gillespie, M.; Matthews, L.; Caudy, M.; Garapati, P.; Gopinath, G.; Jassal, B.; Jupe, S.; Kalatskaya, I.; Mahajan, S.; May, B.; Ndegwa, N.; Schmidt, E.; Shamovsky, V.; Yung, C.; Birney, E.; Hermjakob, H.; D'Eustachio, P.; Stein, L., Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research* **2011,** *39* (Database issue), D691-D697.

304. Cheung, N.; Mitchell, P.; Wong, T. Y., Diabetic retinopathy. *Lancet* **2010,** *376* (9735), 124-36.

305. Yau, J. W.; Rogers, S. L.; Kawasaki, R.; Lamoureux, E. L.; Kowalski, J. W.; Bek, T.; Chen, S. J.; Dekker, J. M.; Fletcher, A.; Grauslund, J.; Haffner, S.; Hamman, R. F.; Ikram, M. K.; Kayama, T.; Klein, B. E.; Klein, R.; Krishnaiah, S.; Mayurasakorn, K.; O'Hare, J. P.; Orchard, T. J.; Porta, M.; Rema, M.; Roy, M. S.; Sharma, T.; Shaw, J.; Taylor, H.; Tielsch, J. M.; Varma, R.; Wang, J. J.; Wang, N.; West, S.; Xu, L.; Yasuda, M.; Zhang, X.; Mitchell, P.; Wong, T. Y., Global prevalence and major risk factors of diabetic retinopathy. *Diabetes care* **2012,** *35* (3), 556-64.

306. Simo, R.; Villarroel, M.; Corraliza, L.; Hernandez, C.; Garcia-Ramirez, M., The retinal pigment epithelium: something more than a constituent of the blood-retinal barrier--implications for

the pathogenesis of diabetic retinopathy. *Journal of biomedicine & biotechnology* **2010,** *2010*, 190724.

307. Chen, Y. H.; Chen, J. Y.; Chen, Y. W.; Lin, S. T.; Chan, H. L., High glucose-induced proteome alterations in retinal pigmented epithelium cells and its possible relevance to diabetic retinopathy. *Molecular bioSystems* **2012,** *8* (12), 3107-24.

308. Chen, Y. H.; Chou, H. C.; Lin, S. T.; Chen, Y. W.; Lo, Y. W.; Chan, H. L., Effect of high glucose on secreted proteome in cultured retinal pigmented epithelium cells: its possible relevance to clinical diabetic retinopathy. *Journal of proteomics* **2012,** *77*, 111-28.

309. Young, S. P.; Wallace, G. R., Metabolomic analysis of human disease and its application to the eye. *Journal of Ocular Biology, Diseases, and Informatics* **2009,** *2* (4), 235-242.

310. Beltran, A.; Suarez, M.; Rodriguez, M. A.; Vinaixa, M.; Samino, S.; Arola, L.; Correig, X.; Yanes, O., Assessment of compatibility between extraction methods for NMR- and LC/MS-based metabolomics. *Analytical chemistry* **2012,** *84* (14), 5838-44.