UNIVERSITAT ROVIRA I VIRGILI

FROM SPECTROMETRIC DATA TO METABOLIC NETWORKS: AN INTEGRATED VIEW OF CELL METABOLISM

Jordi Capellades Tomàs

The book cover is based on a design from Freepik.com

Jordi Capellades Tomàs

From spectrometric data to metabolic networks:
an integrated view of cell metabolism

PhD Thesis Dissertation

Supervised by

Dr. Oscar Yanes Torrado

Dr. Maria Vinaixa Crevillent

Department of Biochemistry and Biotechnology

UNIVERSITAT ROVIRA I VIRGILI

Tarragona

2019

UNIVERSITAT
ROVIRA I VIRGILI

I STATE that the present study, entitled “From spectrometric data to metabolic networks: an integrated view of cell metabolism”, presented by Jordi Capellades Tomàs for the award of the degree of Doctor, has been carried out under my supervision at the Department of Biochemistry and Biotechnology of this university.

Reus, 25th November 2019

Doctoral Thesis Supervisor/s

VINAIXA
CREVILLENT MARIA
- 77834682Z

Digitally signed by
VINAIXA CREVILLENT
MARIA - 77834682Z
Date: 2019.11.24 10:42:04
+01'00'

Maria Vinaixa Crevillent

OSCAR
YANES
TORRAD
O

Digitally signed
by OSCAR
YANES
TORRADO
Date:
2019.11.24
16:16:47 +01'00'

Oscar Yanes Torrado

Acknowledgements

Primer de tot agraeixo als meus pares tota l'ajuda i els sacrificis que han fet per a que jo hagi arribat fins aquí. A la Núria, també, per ser la germana que escolta, aconsella, anima i fot canya quan toca. Moltíssimes gràcies a tots tres!

Escriure aquesta part de la tesi és una de les coses més agredolces que he fet mai. La veritat és que m'agradaria prendre'm amb molta més calma el fet d'escriure aquestes paraules per no deixar cap detall ni persona fora d'elles. Si una cosa he fet en aquest temps, és aprendre de tots vosaltres.

Òbviament, haig d'agrair l'Oscar i a la Mariona tots aquests anys. No tant com a directors de tesi, sinó com a companys i mentors. Gràcies a tots dos per l'esforç en ensenyar-me a ser millor comunicador i també per totes les infinites correccions.

Dono les gràcies a l'Oscar per dipositar en mi la seva confiança i donar-me la oportunitat de estar en el seu petit, però potent grup, on he gaudit d'experiències que mai hauria imaginat.

Em faltarien pàgines per a descriure el que haig d'agrair a la persona que m'ha marcat més aquest temps, la Maria "Mariona" Vinaixa Crevillent, no et puc dir més: "Gràcies pel temps que has dedicat en ensenyar-me!". No pares mai de transmetre els teus coneixements (no són pocs) a tots nosaltres, sense esperar res a canvi, i ja sabem el valor que té això.

No em puc oblidar pas dels primers companys del *zulo*: el Miguel, la Sara i la Miriam de *Hospi*; la vostra alegria era el millor per començar el dia. Us he trobat a faltar. Gràcies Miguel per dedicar llargues estones a il·luminar-me en el mòn de la RMN i la química orgànica, ets un pou de saviesa i sobretot una gran persona. A la Sara, per ser la millor consellera ja siguin idees per a programar, sobre la vida, emocionals, de l'esport o la nutrició; ets inspiradora. A la Miriam, per ser la companya de fatigues barallant-nos amb R o el que fes falta, i per ensenyar-me de sobre espectrometria i per no deixar mai de ser una dolça, treballadora i optimista.

Ni dels més recents, la Sandra i el Joan. Sandra, ha sigut molt divertit veure com has deixat de ser una *"analítica"* cap quadrada, i igual de divertit ha sigut entendre'ns a l'hora de posar a punt la *xurrera.* Ets una crack i una xina, això si no ho deixaràs mai de ser. Joan, ets terrible i molt bon paio, i el teu punt de vista és refrescant. Ets un gran col·lega i m'ha encantat ajudar-te en tot el que has fet, i el que vindrà. Gràcies al vostre suport moral durant el final de la tesi, i també pels vostres consells culinaris.

No em deixaré pas el meu company de cafès, Enrique "Kike". Gràcies per totes les agradables estones a la cafeteria, no hi ha res millor per desconnectar uns minuts.

D'altra banda m'agradaria agrair totes les experiències amb els companys i amics del DEEEA de Tarragona i Biosfer Teslab, ja fossin professionals, com les reunions i seminaris dels dijous o els esmorzars al "Delta", com més festives, ja siguin dinars, sopars o calçotades. Anomenar-vos a tots seria impossible, ja que aquesta família no para mai de canviar.

Gràcies també a tots els amics que m'heu distret i acompanyat tot aquest temps. Valoro molt tenir tanta bona gent amb qui poder relaxar-se.

Finalment agraeixo de tot cor a l'Anna per tot el suport que m'ha donat durant la llarga creació d'aquest manuscrit. Gràcies a tu, ha sigut molt més fàcil aconseguir aquest objectiu i m'has ajudat més del que creus.

I must also mention all the great colleagues and friends that I found in Birmingham. Specially; the computational metabolomics team, all the football teammates, plus all the Spanish friends that I made inside and outside the campus. Lads, you made dealing with the Midlands' winter quite bearable!

“Don’t fear failure.
Not failure, but low aim,
is the crime.
In great attempts it is
glorious even to fail.”

– *Bruce Lee*

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLE OF CONTENTS

ABSTRACT

ABSTRACT

Understanding the molecular basis of life has been in the spotlight of biochemistry research for more than a century already. Molecular biology has taken medicine forward thanks to technological breakthroughs like DNA sequencing and CRISPR editing. However, in order to understand metabolism we must rely on the study of metabolite profiles and metabolic reactions. The purpose of this thesis is to contribute to this area, which unites the fields of proteomics and metabolomics. Traditionally, omics data analysis treats variables independently even if it is strongly settled that molecular mechanisms involve the interaction of diverse pathways, therefore data should be analyzed correspondingly. A vast amount of metabolic pathways have been described, together with enzymes that are responsible for metabolite transformations, this information has been assembled in databases that, in turn, can be used to build metabolic networks.

In here, we use metabolic networks to predict metabolite dysregulation based on quantitative proteomics profiles. To validate the predictions, it is possible to measure the abundance of metabolites or their flux, namely the rate at which they are transformed, using stable isotope labelling experiments, both measurements can be performed by metabolomics. In this thesis, two different metabolomics-based stable isotope labelling approaches have been developed, one for the study of central carbon metabolites and one for the unbiased detection of deregulated fluxes in other metabolic pathways. These approaches have been tested on different datasets and have proven valuable to obtain remarkable results, unraveling molecular mechanisms in diabetes complications or novel metabolic hallmarks of cancer.

ABSTRACT

ABSTRACT

La biologia molecular ha avançat considerablement gràcies a importants progressos com la seqüenciació del ADN o la seva modificació per CRISPR. Tot i això, per entendre el metabolisme requerim estudiar els perfils metabòlics i les seves reaccions metabòliques. L'objectiu d'aquesta tesi és contribuir en aquest estudi del metabolism, el qual unifica dels camps de la proteòmica i la metabolòmica. Tradicionalment, l'anàlisi de dades òmiques es basa en el tractament independent de les diferents variables encara que està profundament establert que els mecanismes moleculars són controlats per la interacció de diferents molècules, i per tant seria més correcte tractar les dades de la mateixa manera. Avui dia, s'han descrit una gran quantitat de vies metabòliques, incluint els enzims responsables de les transformacions dels metabòlits que les formen, aquesta informació s'ha recopilat en bases de dades, que a la vegada poden ser utilitzades per a construir xarxes metabòliques.

En aquesta tesi, s'han utilitzat xarxes metabòliques per a desenvolupar un algoritme que prediu metabòlits desregulats basant-se en el perfil d'expressió d'enzims gràcies a proteòmica quantitativa. Per a validar tals prediccions, és possible mesurar l'abundància d'aquests metabòlits, o el seu flux, o sigui la velocitat a la que s'han transformat, utilitzant experiments de marcatge amb isòtops estables, mesures completades mitjançant metabolòmica. Aqui, mostrem els productes del desenvolupament de dos mètodes per a l'anàlisi de dades de metabolòmica per a experiments amb isòtops estables: el primer per a la quantificació dirigida del flux en metabòlits del metabolisme central; i un segon, per la detecció no-dirigida de metabòlits marcats amb isòtops en altres vies metabòliques. Aquests mètodes han sigut provats en diferents estudis on han aportat resultats remarcables, revelant nous mecanismes moleculars en una complicació de la diabetes o en relació al metabolisme del càncer.

LIST OF TABLES AND FIGURES

LIST OF TABLES AND FIGURES

LIST OF TABLES AND FIGURES

LIST OF PUBLICATIONS

Capellades J, Navarro M, Samino S, Garcia-Ramirez M, Hernandez C, Simo R, Vinaixa M, Yanes O (2015) “geoRge: A computational tool to detect the presence of stable isotope labelling in LC/MS-based untargeted metabolomics” Analytical chemistry. 88:621-628.

Vinaixa M, Rodríguez MA, Aivio S, **Capellades J**, Gómez J, Canyellas N, Stracker TH, Yanes O (2017) “Positional Enrichment by Proton Analysis (PEPA): A One-Dimensional 1H-NMR Approach for 13C Stable Isotope Tracer Studies in Metabolomics” Angewandte Chemie International Edition. 56:3531-3535.

Senan O, Aguilar-Mogas A, Navarro M, **Capellades J**, Noon L, Burks D, Yanes O, Guimerà R, Sales-Pardo M (2019) ”CliqueMS: A computational tool for annotating in-source metabolite ions from LC-MS untargeted metabolomics data based on a coelution similarity network” Bioinformatics

Bueno MJ, Jimenez-Renard V, Samino S, **Capellades J**, Junza A, López-Rodríguez ML, Garcia-Carceles J, LopezFabuel I, Bolaños JP, Chandel NS, Yanes O, Colomer R, Quintela-Fandino M. (2019) “Essentiality of fatty acid synthase in the 2D to anchorage-independent growth transition in transforming cells” Nature Communications, 10 (1), 5011.

Llinàs-Arias, P, Rosselló-Tortella, M, López-Serra, P, Pérez-Salvia, M, Setién, F, Marin, S, Muñoz, J P, Junza, A, **Capellades, J**, Calleja-Servantes, M E, Ferreira, H J, De Moura, M C, Srbic, M, Martínez-Cardús, A, De La Torre, C, Villanueva, A, Cascante, M, Yanes, O, Zorzano, A, Moutinho, C Y Esteller, M (2019) “Epigenetic loss of the endoplasmic reticulum–associated degradation inhibitor SVIP induces cancer cell metabolic reprogramming” JCI Insight, 4(8)

Soukupova J, Malfettone A, Hyroššová P, Hernández-Alvarez MI, Peñuelas-Haro I, Bertran E, Junza A, **Capellades J**, Giannelli G, Yanes O, Zorzano A, Perales JC, Fabregat I (2017) “Role of the Transforming Growth Factor-β in regulating hepatocellular carcinoma oxidative metabolism” Scientific Reports. 2;7(1):12486.

LIST OF ABBREVIATIONS

LIST OF ABBREVIATIONS

ARPE-19	Human retinal pigment epithelial cell line
ATP	Adenosine triphosphate
BRB	Blood retinal barrier
BSS	Blind source separation
CE	Capillary electrophoresis
CI	Chemical ionization
CID	Collision-induced dissociation
CV	Coefficient of variation
DDA	Data-dependent acquisition
DESI	Desorption electrospray ionization
DM	Diabetes mellitus
DR	Diabetic retinopathy
EC	Enzyme Code
EIC	Extracted ion chromatograms
ESI	Electrospray ionization
FAMEs	Fatty acid methyl esthers
FC	Fold change
FDR	False discovery rate
FTICR	Fourier transform ion cyclotron resonance
GC	Gas chromatography
GPR	Gene-Protein-Reaction rules
H25	Hyperglycemia and hypoxia
H5	Normoglycemina and hypoxia
HILIC	Hydrophilic interaction chromatography
HMDB	Human Metabolome Database
LC	Liquid chromatography
LIT	Linear ion trap
LTQ	Linear quadrupole ion trap
MA	Methoxyamine
m/z	mass-to-charge ratio
MIM	Multiple ion monitoring
MS	Mass Spectrometry
MS/MS	Tandem mass spectrometry
MSI	Metabolomics Standards Initiative
mzRT	Peak with unique m/z and a specific retention time
N25	Hyperglycemia
N5	Normoglycemina and normoxia
NADPH	Nicotinamide adenine dinucleotide phosphate
NIST	National Institute of Standards and Technology

NMR	Nuclear magnetic resonance
ORA	Overrepresentation analysis
PAGE	Polyacrilamide Gel Electrophoresis
PCA	Principal component analysis
PDR	Proliferative diabetic retinopathy
PPI	Protein-protein interaction
PRM	Parallel reaction monitoring
PTM	Post translational modulation
QC	Quality control
QIT	quadrupole ion trap
Q-TOF	Quadrupole-TOF
QqQ	Triple quadrupole
QTrap	Triple-quadrupole ion trap
RP	Reversed-phase
RT	Retention time
SAH	S-Adenosylhomocysteine
SAM	S-Adenosylmethionine
SIM	Single ion monitoring
SCX	Strong cation exchange
SD	Standard deviation
SDS	Anionic sodium dodecyl sulfate
SILAC	Stable isotope labelling by amino acids in cell culture
TCA	Tricarboxylic acid
TMS	Trimethylsilyl
TMT	Tandem mass tags
TOF	Time of flight
TQ	Triple quadrupole
UPLC	Ultra performance liquid chromatography

INTRODUCTION

1. METABOLISM AND ENERGY

1.1 The role of energy in biochemistry

All living organisms must obtain and use energy to live, grow, reproduce and eventually evolve. At the cellular level, this energetic demand is commonly supplied via the transformation of chemical substrates. Metabolism can be defined as the collection of life-sustaining chemical reactions in living organisms.[1] Metabolism plays a central role in life being the source of energy and building blocks for cellular growth as well as ensuring protection against stress factors. Metabolism has evolved to support function of the cell and can roughly be divided into three functional modules: central carbon metabolism, biosynthetic pathways and secondary metabolism.[2]

Organisms can perform a wide variety of metabolic reactions with great efficiency. Chemical compounds taking part in these reactions are referred to as metabolites. Carbohydrates, nucleotides, amino acids, and lipids are among others examples of metabolite families whose atomic composition and structure vary as much as their physico-chemical properties and roles in the cell.[1] Metabolites can also be combined to form more complex macromolecules, such as DNA, RNA and proteins, that acquire whole new biochemical properties and functionalities.

1.2 Metabolic reactions

In living organisms, homeostasis is maintained through regulation of gene expression and enzyme activity, which, in turn, control metabolic fluxes when cells require a higher production of a given compound or family of compounds. Different molecular mechanisms regulate all these interlocked metabolic processes simultaneously, obtaining every product in the amount needed and at the right time. Such control involves a variety of mechanisms operating at different time scales, in order to properly adjust metabolic fluxes when environmental conditions change. If homeostasis fails, disease or death is often the consequence.

Metabolic reactions are bioenergetic processes, and as any other energy exchange processes, are subject to the laws of thermodynamics. Thermodynamic laws and molecular properties are the foundations of metabolic reactions in a living organism, yet evolution has only brought nature a limited number of reaction options compared to those used in organic chemistry labs. In general, these reactions can be[3]:

- **Catabolic.** Breaking down larger molecules to release energy and to generate precursors for further reactions.
- **Anabolic.** Combining molecules together to generate biologically useful molecules and biopolymers.
- **Interconversions.** In which there is no change of order but instead functional groups or isomer configurations are swapped between molecules.

Metabolic reactions can be artificially organized in schematic maps that interconnect compounds in so called metabolic pathways. In a metabolic pathway, the product of one reaction is connected to the following reaction as a reactant, normally associated with information regarding genes and proteins that control such reaction. Biochemists and biologists have classified metabolic pathways depending on their main metabolic purpose, though it must be taken into account that they are all interconnected, and thus free exchange of metabolites happens between them, except if they are compartmentalized in certain organelles. Therefore, metabolism can be viewed as a collection of metabolic pathways in which the concentrations of products and reactants are constantly fluctuating to balance reactions.

Thus, although the rate of metabolite transformation, or flux, through any step of the pathway may be high and variable, the concentration of substrate remains typically constant. This ability of maintaining the intracellular concentrations within limited boundaries is known as homeostasis.[1] Balancing reaction kinetics is necessary in living organisms, otherwise their metabolic systems would collapse.

1.3 Enzymes and enzymatic reactions

In metabolic reactions, the elements responsible for the transformation of metabolites are a family of protein macromolecules known as enzymes. Enzymes themselves are also under tight regulation, which involves a combination of regulators in the form of other proteins and gene processes.

When a chemist performs a reaction in the laboratory, diverse parameters can be applied to increase the yield or rate of a reaction, including temperature, pressure or the use of catalysts. In contrast, most biological systems carry out reactions at the temperature maintained by the organism and at atmospheric pressure[3]. This is possible because biological systems use enzymes are molecular catalysts. An enzyme drives the reaction by providing a specific environment in which a given reaction can occur more rapidly, in other words, enzymes make the reaction thermodynamically favorable. Many biochemical reactions, such as the formation of the amino peptide bond in a protein, would be otherwise unfavorable in isolation. Thus, in order to drive the reaction forward, biochemical systems often couple reactions with a positive-free energy change to those with a large negative-free energy change.[3] Enzyme structure contains a highly specific substrate-binding site that acts lowering the energy of high energy species along the reaction intermediates from starting material to product. This physical pocket in the enzyme structure is known as the active site.[1]

Generally, enzymes are proteins, macromolecules of medium or high molecular weight made up of amino acids by peptide bonds. Their amino acid composition defines their (1) folded tridimensional structure due to non-covalent interactions and (2) an enzymatic activity, determined by one or more amino acid motifs, evolutionarily conserved sequences of amino acids, or patterns of sequences, that are characteristic of certain type of enzymatic reactions. A second group of macromolecules which also have enzymatic activity are RNA molecules, known as ribozymes. Ribozymes were discovered in the 1970s, when the first RNA molecule was found to have catalytic activity proving that nucleic acids are more than mere passive carriers of information.1 Even though ribozymes do not strictly participate in metabolic reactions themselves, their action is essential for other processes related to transcription and translation.

Since enzyme naming is ambiguous and there is an ever-increasing number of newly discovered enzymes, biochemists have adopted a system for naming and classifying enzymes by international agreement. Enzymatic reactions are classified by the scientific community using the EC (Enzyme Consortium) codes which define the type of reactions that occur, the substrates, the products and even the cofactors necessary for it to happen.[1]

Table 1. Human protein enzymes in Uniprot database[4] per main enzyme class.

Main enzyme class	Class function	Number of subgroupsin class	Number of human proteins *Reviewed
EC 1. Oxidoreductases	Transfer of electrons	25	558
EC 2. Transferases	Transfer of functional groups	10	1825
EC 3. Hydrolases	Molecule hydrolysis	13	1606
EC 4. Lyases	Non-hydrolytic addition or removal of groups from substrates	8	156
EC 5. Isomerases	Isomerization, rearrangement of functional groups	6	122
EC 6. Ligases	Molecular binding	6	125
EC 7. Translocases	Movement of molecules across or through membranes	6	63

1.4 Enzymatic regulation

Continuing with the chemistry lab analogy, the function of a catalyst is to increase the rate of a reaction without affecting the equilibrium of the reaction. Consequently, to control the reaction equilibrium cells have different strategies to regulate the availability of the elements that participate in the reaction. Firstly, the concentration of substrates and products can be controlled by regulating their import and export from and to the environment or between intracellular compartments. Secondly, enzymes can be regulated through:[1]

- **Transcription and translation regulation.** One of the strategies to control the concentration of an enzyme is through their genetic expression, i.e. modulating their transcription and translation processes. Gene expression and protein expression are regulated at different molecular levels, for example using transcription factors, enhancers and RNA modification. Once folded into a functional protein, enzymes can also be regulated by localization, inhibition and post-translational modifications mechanisms.

- **Localization.** Since cells are highly organized living systems, reactions themselves are inevitably too, meaning that reactions are compartmentalized in certain parts of the cell. Therefore, if enzymes are locked in a different part of the cytoplasm different from the reactants, their availability is limited until it is necessary.

- **Inhibition.** Another example of enzymatic control is the use of inhibitors, molecules that are able to interact with the enzyme structure or its active site. Enzyme inhibitors interfere with catalysis, slowing or halting enzymatic reactions. Depending on the type of bond, inhibitors are either reversible or irreversible, yet three types of reversible inhibition exist:

  - Competitive inhibitors bind to the enzyme’s active site that prevents binding of the substrate to the enzyme.
  - Uncompetitive inhibitors bind at a separate site but bind only to the enzyme-substrate complex.
  - Mixed inhibitors bind at a separate site, but may bind to either the enzyme or the enzyme-substrate complex.

- **Post translational modification (PTM).** Finally, enzyme activity can also be controlled by modifying the functional groups of their amino acid sequence. The introduction of a charged chemical group or a different functional group can alter the local properties of the enzyme and induce a change in conformation, which can affect the activation status of the enzyme, its localization or its affinity to/for a substrate.[1] PTM modifying groups include a wide range of small chemicals such as phosphoryl, acetyl, methyl, glucosyl, but also larger modifications including ubiquitin and sumo proteins.

2. BIOMEDICINE AND OMIC SCIENCES

2.1 From biochemistry to molecular biology

Compiling and contrasting all the information about life has taken centuries of hard work done by biologists, physicians, chemists, physicists, mathematicians, biochemists and molecular biologists. Early and mid 20th century represented the "golden age of biochemistry" when the foundations of molecular biology were laid.[3]

First, pathways responsible for nutrient utilization and energy production in humans and other organisms were delineated. This was followed by important landmarks such as protein composition and DNA structure characterization among others. This led to definition of the central dogma of molecular biology stating a simplified gene-transcript-protein relationship. The progressive discovery of the function of biomolecules drawn the attention of biomedical research, which was traditionally focused on studying genetic variants and mutations that are associated with different diseases.[3] Common diseases were understood in terms of inherited or somatic mutations that impact gene expression, signal transduction, cellular differentiation and other processes not traditionally viewed in bioenergetic or metabolic terms.[5] However, current evidence is showing that many metabolic perturbations accompany human diseases causing major worldwide burden such as cancer or type 2 diabetes.

A completely refurbished view of metabolism is gaining ground. This is nowadays regarded as a highly-organized network of metabolic reactions, meaning that a perturbation of almost any part of metabolism results in a global response in which a large number of enzymes have to alter their regulation to maintain homeostasis.[2] In other words, metabolism impacts, and is impacted by, virtually every other cellular process when responding to cellular needs.[5] Consequently it is clear that there is a need to go downstream translation to be able to explain complex diseases where fully functional gene variants, instead of mutations, may predispose the individual to suffer from a certain condition. The combined effect of such gene variants and environmental conditions defines the pathological observable phenotype. Modern biomedical research aims at understanding the causes and mechanisms leading to such pathological phenotypes at the molecular level. To this end, ‘omic’ sciences have been progressively introduced.[6]

2.2 The family of omic sciences

Since the discovery of the structure of DNA, molecular biologists got more and more interested in understanding biology in detail at the molecular level. This led to the development of technology capable of recovering the sequence of DNA, the detection of molecules with tiny masses or the characterization of protein binding and its dynamics, giving birth to omic sciences and thus, modern biomedicine.

Omic sciences are a family of interdisciplinary sciences that focus on the generation of a comprehensive profiling of different kinds of molecules in a cell culture, tissue or organism. The output of omic technologies may include information about abundance, identity and modifications of the analyzed molecules. Omic sciences can be applied not only for greater understanding of normal physiological processes but also in disease where they play a role in screening, diagnosis and prognosis as well as aiding our understanding of the etiology of diseases.

Depending on the family of molecules studied, omic sciences can be briefly divided into:[7]

- **Genomics** is the systematic study of an organism's genetic DNA material. Genomics studies normally involve the characterization of genomes in order to compare gene sequences, their variants and regulators across species or generations.

- **Transcriptomics** studies gene expression and gene variability through gene's RNA transcripts. Additionally, non-coding RNA analysis gives information about their regulatory mechanisms.

- **Epigenomics** is the study of epigenetic modifications on the DNA and histones. These modifications play an important role in gene expression and regulation.

- **Proteomics** aims to understand the functional relevance of proteins, including their modifications and interactions. The proteome is defined as the set of all expressed proteins in a cell, tissue or organism.

- **Metabolomics** can be defined as the study of metabolite profiles. The metabolome is the final downstream product of gene transcription and protein expression, therefore, the metabolome is closest to the phenotype.

3. MASS SPECTROMETRY-BASED OMICS

3.1 Mass spectrometers

Mass spectrometry (MS) is a technology applied to measure the abundance of molecular ions in a sample. MS has become invaluable across a broad range of fields and applications, including metabolomics and proteomics. In the end, both metabolomics and proteomics measure ionized molecular species that are obtained by ionizing either metabolites, peptides or proteins, respectively.

Mass spectrometers are an heterogeneous family of instruments capable of ionizing, filtering and measuring the abundance of ions present in a sample. Mass spectrometers typically consist of the following elements: inlet, source, mass analyzer and an ion detection module.[8,9]

- **Sample inlet.** This is an injection system that is used to introduce the content of a sample vial into the ionization source. The pump may be either manual in the case of direct injection or connected to a chromatography system working at high pressures (see Section 3.3).

- **Ionization source**. In which sample analytes are subjected to a physico-chemical process that produces ions suitable for resolution in the mass analyzer, this process is known as “ionization” and takes place in gas phase. This process involves the generation of charged molecules from sample compounds through an energy transfer, generating charged molecular ions from the original molecule. There are two main types of ionization methods: soft ionization methods leave the structure of the original molecule relatively unharmed, while hard ionization methods cause fragmentation of the compound molecular structure into smaller fragment ions.

The most common ionization method for small molecule and peptide analysis is electrospray ionization (ESI), a soft ionization method that uses desolvation through a current at high voltage to produce charged molecular ions. Other soft ionization approaches such as atmospheric pressure chemical ionization (APCI) and atmospheric pressure photon ionization (APPI) are also used, mainly in metabolomics. These ionization methods use either a molecular reaction by interaction with a reagent gas in the former, or ultraviolet light in the latter.

In the case of GC/MS-based metabolomics experiments, it is possible to utilize electron ionization (EI) as well as chemical ionization (CI) with different reagent gases. EI is a highly reproducible hard ionization technique with extended use in GC/MS, while CI has more of a niche use.

Finally, in the case of MS imaging (MSI), a technique that acquires MS spectra from bidimensional solid matrices, matrix-assisted laser desorption/ionization (MALDI), desorption electrospray ionization (DESI) and secondary ion mass spectrometry (SIMS) are different technologies, each with its own ionization particularities, but all apply an ionization energy onto a solid sample instead of a liquid or gas.[14]

- **Mass analyzer.** Mass analysers can be separated into two main groups: (i) trapping mass spectrometers such as ion trap (IT) or Orbitrap and (ii) ion-beam mass spectrometers, such as TOF and quadrupoles. The main difference is that trapping systems resolve ions discontinuously while ion-beam instruments do it continuously.[10]

    - There are many types of mass analyzers, using either static or dynamic fields, and magnetic or electric fields, but all operate according to a differential equation of motion of charged particles: the Lorentz law.

Ion trap (IT), Orbitrap, and ion cyclotron resonance (ICR) mass analyzers separate ions based on their m/z resonance frequency, while quadrupoles (Q) use m/z stability and time-of-flight (TOF) analyzers use time. In consequence, each analyzer possesses different acquisition capabilities that can be tuned depending on the objectives of the analysis: [11]

- Resolution. A metric that measures the amount of mass units that the analyzer is capable of distinguishing between two neighboring ion peaks.

- Sensitivity. A qualitative indicator of the minimal amount of analyte that the analyzer detects. It can be quantitatively determined using a dilution series, and it is reported as the limit of detection.

- Scan speed/rate. The frequency at which the instrument is recording ion abundances, depending on the architecture of the analyzer and the detector, this speed can go from 1 to 100 Hz.

- Mass accuracy. A quantitative measure of how similar the recorded m/z value is to the true m/z ion. This difference is expressed in parts per million (ppm).

- Response range. It is the range of absolute abundance units at which the detector operates, in which we can distinguish: (a) dynamic range, the concentration range which gives a measurable response that goes from the limit of detection to saturation; and the (b) linear range, that is the concentration range at which the response of the detector is linear, and thus considerably more reliable.

- Mass range. It is the range of m/z values at which the analyzer performs with maximal mass accuracy and sensitivity. Some instruments sacrifice mass range in order to maximize scan rate and resolution.

- **Mass detector.** The detector records the charge induced or a change in current produced when an ion passes by or hits a surface, these are associated to abundance of the previously measured molecular ions, a task solved by the instrument software. There are different types of detector hardware yet they are generally chosen to match the performance of the mass analyzer.

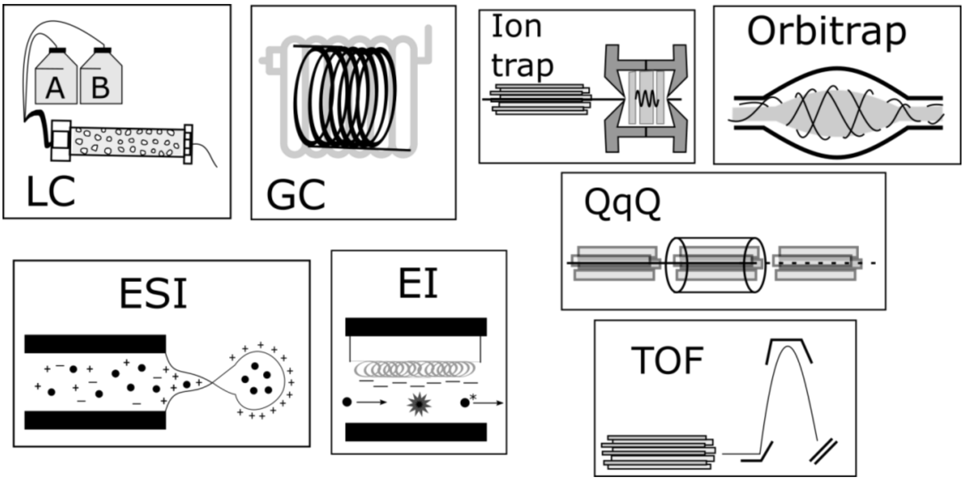


Figure 1. Diagram of parts of mass spectrometers.

In order to improve the capabilities of mass spectrometry systems, hybrid instruments emerged, which balance the advantages and limitations of different instrument types. For instance, triple-quadrupole instruments are still very limited in resolution and mass accuracy, but they offer a high dynamic range and excellent sensitivity, which make them ideal instruments for peptide and small molecule quantitation in targeted analyses.

Plus, a hybrid of the highly efficient quadrupole and a highly accurate and resolving Orbitrap is ideal for both peptide and small molecule detection/identification.[10]

Additionally, ion mass filter technology (as in quadrupoles) or the ability to perform ion fragmentation via tandem mass spectra (see Section 3.2) are also valuable features of hybrid mass spectrometers. Other useful utilities are the ability to shift between positive and negative polarity modes, meaning that they can detect positively or negatively charged ions, some instruments may be even capable of shifting at high rates.

Table 2. Mass analyzers and their capabilities (Adapted from Junot et al. [12]). i.c, internal calibration; e.c., external calibration

Mass analyzer type	Max resolving power (FWHM)	Mass accuracy (ppm)	Max scan rate (Hz)	Max mass range	Dynamic range (Log10 AU range)
QqQ	7.500	5	5	3.000	5-6
IT/LIT	10.000	50	30	4.000	4
qLIT	9.000	50	20	2.000	5-6
TOF	20.000	1 (i.c.)	40	20.000	4-5
IT-TOF	10.000	2 (i.c.)	10	40.000	3-4
qTOF	60.000	2 (i.c.)	100	40.000	4-5

Orbitrap (Exactive)	200.000	1 (i.c.) 3 (e.c.)	8-15 *Depends on resolution	6.000	4-5
LTQ-Orbitrap	240.000	1 (i.c.) 3 (e.c.)	5-10 *Depends on resolution	4.000	3-4
qOrbitrap	150.000	1 (i.c.) 5 (e.c.)	8-12 *Depends on resolution	4.000	3-4
LTQ-FT-ICR (7T)	750.000	<1 (i.c.) 1 (e.c.)	1-5 *Depends on resolution	4.000	3-4
qFT-ICR (7T)	1.000.000	<1 (i.c.) <1.5 (e.c.)	1-5	10.000	3-4
LIT-qOrbitrap (IDX)	500.000	1 (i.c.) 3 (e.c.)	40 *Depends on resolution	2.000	3-4
qTOF (Agilent 6546)	60.000	1 (i.c.)	30	40.000	5-6

3.2 Tandem mass spectrometry

Omics technologies play an important role in biomarker discovery as well as in other stages of the drug discovery and development (e.g. target discovery, mechanism of action or predicting toxicity). In particular, recent progress in tandem mass spectrometry techniques have helped facilitate the realization of the inherent power of proteomics and metabolomics in biomarker discovery, validation and qualification.[13]

In mass spectrometry, the mass-to-charge ratio is a non-informative measure of molecular ion structure, thus this magnitude is hardly traceable into a compound name in metabolomics or amino acid sequence of a peptide in proteomics. In order to attain such objectives, a further level of information can be obtained through fragmentation of molecular ion structure using tandem mass spectrometry.

Tandem mass spectrometry, also known as MS/MS or MS2, involves three distinct steps of (1) selection of a single or multiple ion/s, (2) fragmentation and (3) mass separation of fragments. The fragmentation and separation of fragments may take place in a different space in the case of instruments that contain collision cells such as QqQ, qTOF, or hybrid ion trap/FTMS instruments; or scan time through ion accumulation in ion traps, Orbitrap and FT-ICR MS.[14]

In the case of small molecules, as in metabolites, the structure of each parent/precursor molecular ion-generates a pattern of fragments, or product ions, that are highly specific, but not necessarily unique. On the other hand, MS/MS is used in proteomics to fragment peptide or protein ions in order to elucidate their amino acid sequence, useful indeed for protein identification.

3.3 Tandem mass spectrometry acquisition modes

The acquisition of MS/MS is strictly tied to the hardware design of the mass spectrometer, in which we can distinguish three different methods depending on how the precursor ions are selectively fragmented: targeted, data dependent and data independent acquisition methods. In most cases, tandem mass spectrometry is applied in mass spectrometers coupled to separation systems in order to be efficiently perform and improve the quality of results.

3.3.1 TARGETED ACQUISITION MODES

In mass spectrometry, targeted analysis implies the selection of a few precursor ions, reporting their relative abundance or absolute concentration by extrapolating their abundance on a calibration curve. Targeted methods are characterized by having high sensitivity, wide dynamic range, reliable quantification accuracy and stability. However, their setup is so complex that they are generally restricted to a low number of preselected analytes.

A few MS strategies exist for targeted analysis in which the quantification is based on full scan mode, meaning that no fragmentation is applied, and instead the abundance of a preselected ion is used for quantification. In these, tandem mass spectrometry is not used for identification and instead a previously characterized retention time serves as a guide to quantify previously selected ions in methods known as selected ion monitoring (SIM) in single quadrupoles or multiple ion monitoring (MIM) in triple quadruples or quadrupole-ion trap hybrids.[15]

MS/MS fragmentation brings further advantages since this technology can be applied to monitor and fragment multiple parent ions and study the fragmentation pattern for greater specificity. These include Multiple reaction monitoring (MRM), or Selected Reaction Monitoring (SRM). SRM is typically performed on a triple quadrupole mass spectrometer, which for long has been used for targeted quantification. SRM is the most used strategy for targeted quantification in both metabolomics and proteomics.[15] Also, a similar strategy can also be applied on a Q-Orbitrap instrument, known as

Parallel Reaction Monitoring (PRM), which allows simultaneous scanning of an entire group of fragment ions at higher resolution than a triple quadrupole.[15]

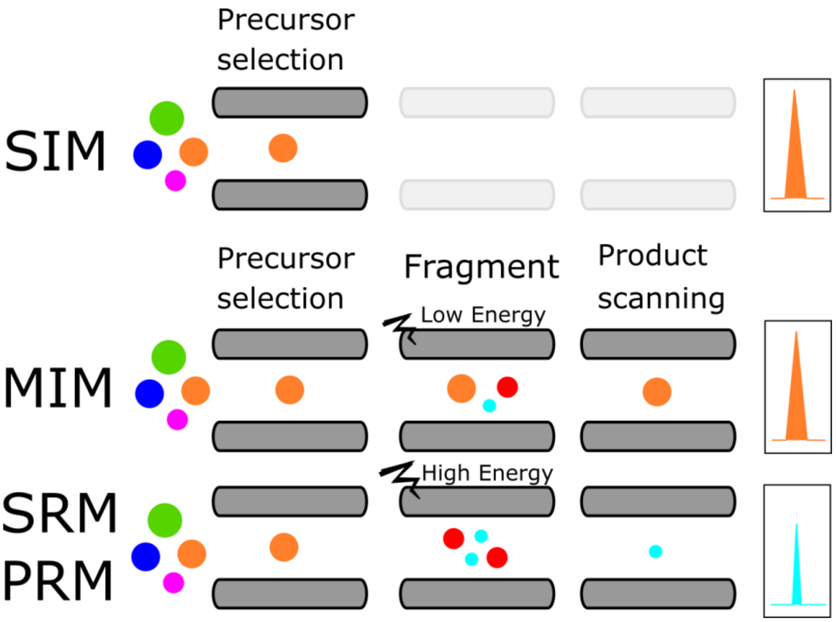


Figure 2. Targeted acquisition modes. Based on Figure 5 from Junot et al.[12]

3.3.2 DATA DEPENDENT ACQUISITION MODES

Modern hybrid instruments have been upgraded with automatable utilities. One of them is the ability to record and filter some ions and then near instantly perform a fragmentation routine for those filtered. This augments the amount of data that can be extracted from a single chromatographic run, including precursor ions and their corresponding fragmented spectrum, yet these approaches have certain limitations.

Data Dependent Acquisition (DDA) implies that fragmentation is only affected if an ion complies with a certain condition in the full scan mode, meaning that in DDA the instruments are sequentially switching between full scan (MS1) and MS2 modes. There are different criteria that can be used to trigger DDA acquisition routines:[8,16]

- **Ion intensity-dependent acquisition.** It uses an intensity threshold to trigger sequential MS/MS acquisition, if a precursor ion exceeds a predefined threshold of intensity fragmentation occurs.

- **Accurate-mass inclusion list-dependent acquisition.** In here a list of accurate masses of expected or predicted metabolites is used to trigger MS/MS acquisition.

- **Isotope pattern-dependent acquisition.** This method takes advantage from the fact that certain atoms generate unique isotopic patterns that can be easily recognized in their mass spectra to trigger MS/MS acquisition.

- **Pseudo neutral loss-dependent acquisition.** This DDA experiment consists in the generation of pairs of full scans, one at low collision energy (i.e., 5 eV) followed by another scan with a higher collision energy ramping (i.e., 20–40 eV), and monitors characteristic m/z differences of ion-pairs (neutral loss) between consecutive low and high collision energy full-scan MS. This can be used to obtain MSMS spectra for a subset of a characterized family of metabolites.

- **Mass defect-dependent acquisition.** Mass defect filter (MDF) was developed to facilitate the detection of both common and uncommon metabolites.

3.3.3 DATA INDEPENDENT ACQUISITION MODES

This approach provides fragment ion information in a non-selective manner by fragmenting all precursor ions within a defined mass range.[17] It requires novel mass spectrometers with fast duty cycles and acquisition times with up to 100 MS/MS scans per second at medium-high mass resolving power. In DIA, all precursor ions in a predefined isolation window (from several Da to a full mass range) are sequentially isolated to acquire multiplexed MS2 spectra and ensure to acquire MS2 spectra for all ions in MS1 scan.[18]

One advantage is that they do not require to be triggered by any event rule (as in data-dependent MS/MS), but the main drawback is that co-eluting molecules with higher intensities (that are usually triggered first in data-dependent MS/MS) and lower abundant ions are still fragmented altogether, incrementing data complexity. Therefore, an obvious disadvantage for this type of analyses is that the direct link between a specific precursor ion and its corresponding product ions is broken, thus these methods produce highly complex MS/MS spectra that require specific data processing software tools.[16,17] DIA methods include:

- **Elevated energy MS (MSE) approach using Q-TOF.** This approach involves alternating between two full-scan functions, one at low collision energy and the other at a higher collision energy range. The mass spectra from low collision energy provide intact molecular ion information, while high collision energy spectra contain fragmentation data useful for structural elucidation.

- **All-ion fragmentation using Orbitraps.** In which a wide range of precursor ions are first recorded and then sent into the HCD cell where they are fragmented.

- **Sequential Windowed Acquisition of All Theoretical Fragment Ion Mass Spectra' (SWATH™) acquisition.** This technology takes advantage of the fast scan speed and narrow selectivity ranges of TOF Sciex (TripleTOF™) mass analyzers, in which the instrument fragments all ions across a given mass range in sequential narrow ranges (20-50 Da) as they elute.[16] SWATH allows a reduction of simultaneously fragmented precursor ions, therefore, decreasing the complexity of MS2 spectra. As a result, the effort to reconstruct the connections between the precursor and product ions is also alleviated. [18]

3.3.4 MULTI-STAGE TANDEM MASS SPECTROMETRY

Last but not least, ion-trap and FT-ICR type instruments are capable of generating multi-stage mass spectrometry (MS^n) spectra, in which MS2 product ions are trapped allowing another isolation and fragmentation to be performed, in MS^n this cycle can be performed multiple (n) times. MS^n generates a list of interrelated MS/MS spectra, also known as ion trees or mass spectral trees, which contain the precursor-product relationships of the different MSn acquisition rounds.

3.3.5 TECHNICAL CONSIDERATIONS

There are two different fragmentation modes for small biomolecule MS/MS-based experiments: collision-induced dissociation (CID) in quadrupoles and higher energy collisional dissociation (HCD), specific of orbital ion trap mass spectrometers. Both CID and HCD fragmentation methods result in complementary fragmentation information because of the different hardware technologies that are implemented in the respective collision cells.[8]

There are two main parameters in the case of tandem mass spectrometry that have a great impact on the resulting fragmentation spectrum. First, the “collision energy” plays an important role in MS/MS spectra generation, especially in CID. Low collision energies preserve most of the precursor ion structure and only few product ions are observed, whereas increasing the collision energy naturally increase product ion abundances toward low m/z ranges, i.e. smaller fragments. Therefore, using a set of different complementary energies is a better source of structural information.

Another important aspect is the “precursor ion isolation width” of the mass selector, since it has an impact on both sensitivity and selectivity during MS/MS data acquisition. In general, selecting narrow precursor ion isolation windows (high resolution precursor isolation) lowers the sensitivity of the precursor ion and thus the intensity of fragment ions. In contrast, widening the isolation window leads to the fragmentation of a larger number of different compounds and results in impure product ion spectra with interfering ions.

3.4 Separation techniques coupled to mass spectrometry

In a single drop of sample, thousands of molecular ions can potentially be detected using a mass spectrometer, but mass-to-charge values would overlap and impede the extraction of ion abundance. Therefore, separation techniques such as chromatography and capillary electrophoresis have been developed to be coupled with mass spectrometry since their main purpose is to, over time, separate compounds based on their physical and chemical properties.

3.4.1 CHROMATOGRAPHY

Chromatography is a separation technique that consists in the use of a mobile phase and stationary phase in order to separate or even purify certain analytes found in an heterogeneous mix. The applications of chromatography are endless, both in research and industry. In the family of omic sciences, both proteomics and metabolomics use chromatography for analyte separation prior to mass spectrometry, an instrument setup named “hyphenated mass spectrometry”. The stationary phase chemically and physically interacts with the analytes in a sample until the mobile phase progressively washes them away in a process known as elution. The amount of time that a single analyte requires to be completely washed away is called elution time, or retention time, such time depends on the strength of the interaction of the analyte with the solid phase and how the mobile phase interferes with such interaction.

Chromatographic separation of analytes prior to MS analyses has several advantages:[9]

- It reduces matrix effects and ionization suppression, side-effects of the ionization procedure from complex samples.
- It is capable of separating isomers.
- It allows for more accurate quantification of separated analytes.

3.4.1.1 Liquid Chromatography

Liquid chromatography (LC) is a separation technique in which the mobile phase is a liquid and the stationary phase is a solid surface, that is variable depending on the analytes to be separated. The technology evolved in what is known as high-performance liquid chromatography (HPLC), which utilizes very small packing particles and a relatively high pressure to maintain a stable and relatively fast flux through the column.

In the last decades chromatographic columns evolved to pack sub-2µm particles, requiring much higher pressures, leading to the definition of ultra-high performance liquid chromatography (UHPLC)[19], also UPLC, capable of better resolved peaks compared to HPLC and in shorter run times. Additionally, in the case of proteomics, a need for high sensitivity using low amounts of sample yielded extremely low flow rate systems known as nanoLC.[19]

Historically LC has been divided into two different subclasses based on the polarity of the mobile and stationary phases. Methods in which the stationary phase is more polar than the mobile phase are termed normal phase liquid chromatography (NPLC) and the opposite method is termed reversed phase liquid chromatography (RPLC).[19]

In proteomics and metabolomics, RPLC is the most common approach used. There is a wide range of applications using RPLC because of the various mobile and stationary phases[20], being C18 silica column using an acidic water/organic mobile phase the standard choice.[21] In C18-bonded silica columns, the selectivity of non- and moderately polar compounds can vary greatly depending on the density of C18 alkyl chains, the accessibility of silanol, or the presence of bonding groups between the silica beads and C18 alkyl groups.

However, chromatography technology has increased the specificity of methods using novel column material designs. A particular example of novel chromatography techniques is Hydrophilic interaction liquid chromatography (HILIC), which utilizes highly polar stationary phases capable of forming a water-rich layer in which hydrophilic solutes are retained based on their polar, ionic and hydrogen bonding interactions.[19] It is a unique type of NPLC chromatography method that can be used to increase coverage of the polar metabolome. The order of elution is inverted compared to RPLC, with hydrophilic compounds being retained longer than hydrophobic compounds.[21] A plethora of HILIC column stationary phases have been developed and can be separated in four categories: anionic, cationic, uncharged and zwitterionic.[21] HILIC separation is used in both metabolomics and proteomics for different reasons. In metabolomics, it has proven suitable for the separation of polar and hydrophilic metabolites[21], whereas in proteomics HILIC is used as an alternative to ion exchange chromatography and for targeted analysis of PTMs.[23]

3.4.1.2 Gas chromatography

Gas chromatography (GC) is a chromatography separation technique in which the mobile phase and the sample are found in a gas phase. GC is a separation method specific for low weight compounds as it only is capable of analyzing volatile compounds, or those that can be chemically modified to be volatile and thermally stable by derivatization.[24]

In gas chromatography, the sample is vaporized and injected into the head of the separation column that then it traverses by the flow of an inert gas employed as the mobile phase. This inert gas is usually helium or an unreactive gas such as nitrogen, although hydrogen is preferred for improved separations. The separation depends on how compounds are adsorbed on the surface of the stationary phase. The stationary phase may be a solid adsorbent, in the case of gas–solid chromatography, or a liquid coating for gas–liquid chromatography. The latter is the most widely used GC technique in metabolomics, due to superior performance and suitability for metabolites. In gas–liquid chromatography, the liquid stationary phase is adsorbed onto a solid inert packing or immobilized on the capillary tubing walls:[24]

- The column is considered packed if the glass or metal column tubing is filled with small spherical inert supports. The packing is an inert support impregnated with 5–20% stationary liquid phase in a thin layer. These liquid phases are primarily silicone-based oils with high temperature stability.

- In a capillary column, tubing walls contain highly absorbent materials that form a thin film of liquid stationary phase that coats their inner wall. Because the tube is open, its resistance to flow is very low, and it is thus referred to as an open tubular column.

Capillary columns although more expensive are a better option for complex mixtures, they offer various advantages relative to packed columns. Higher density and surface of separative elements leads to far superior resolution plus faster and more efficient separation.

Optimizing GC separation requires fine-tuning of a number of variables and their interactions. Both physical (internal diameter, length, and stationary phase) and parametric (temperature and flow velocity) column variables affect the separation process in metabolomics. A range of stationary phases can be applied to metabolome analysis in GC/MS, although methyl-phenyl columns are typically applied (e.g., 95:5 methyl/phenyl and 50:50 methyl/phenyl), plus their length (from 5 to 100 m) and diameter (<0.1 to >0.5 mm) is also variable.[25] In general, most methods use narrow bore fused-silica capillaries with wall-coated 5%(diphenyl)- polydimethylsiloxane (PDMS) as a generic non-polar stationary phase and He as an inert gas with elution gradient programs performed with a maximum temperature under 320 °C within 50 min to reduce column bleed; however, some reports also utilize more polar stationary phases for improved selectivity in GC/MS and/or higher peak capacity in 2D GC × GC/MS, such as 50%(diphenyl)-PDMS and 14%(cyanopropylphenyl)- PDMS bonded phase columns.[19]

3.4.2 CAPILLARY ELECTROPHORESIS

Capillary electrophoresis itself is not strictly a chromatography separation as it lacks a stationary phase, though CE-MS represents a niche separation technique used in both metabolomics and proteomics.[19] In CE, analytes migrate through electrolyte solutions under the influence of an electric field, separated according to their intrinsic electrophoretic mobility, which is dependent on the charge and size of the analyte. CE–MS is often considered a technically challenging approach and limited by poor concentration sensitivity and migration time variability.[26]

CE is applied for resolution of highly polar, charged and labile ions that have poor retention in RP-LC or require complicated sample handling prior to GC/MS, for example multivalent charged ions. Therefore it is reserved for the analysis of special types of multiply charged metabolites or proteins, it has been used for the global and reproducible profiling of native peptides and some metabolites in a clinical setting.[26]

4. PROTEOMICS

Proteins are an incredibly diverse family of molecules even if their sequence can merely be formed by only 20 different amino acids, their order and location in the peptide sequence lead to significantly different protein types, conformations and isoforms, plus the modification of amino acid side chains affects their functionality and stability. Proteomics generally focuses on the global identification and quantification of protein molecules in biological samples. In addition, proteomics also is interested in the study of protein-protein interactions and protein post-translational modifications (PTM).

In a similar manner to metabolomics, proteomics has been growing in terms of throughput due to improved protein extraction protocols and MS instrument development.[29] Though, totally different methodologies may be used to study other protein-related molecular dynamics, these include cell imaging by light and electron microscopy for cellular location, or array and chip experiments for interaction. [30]

4.1 Mass spectrometry-based proteomics

4.1.1 INSTRUMENTATION

Electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) are the two techniques most commonly used to ionize proteins or peptides for mass spectrometric analysis.[29]

Due to the large collection of protein types in complex biological samples, samples need to be simplified prior to protein analysis in order to facilitate the subsequent steps of the workflow. Additionally, enzymatic digestion of a full proteome, usually done by trypsinization, generates hundreds of thousands of peptides. In consequence this level complexity is not compatible with direct MS analysis.[29] Therefore, the first step in most proteomics workflows aims at reducing the sample complexity either by separation, prefractionation or enrichment. [50]

It is important to consider that single-dimension peptide chromatography does not provide sufficient peak capacity to separate peptide mixtures as complex as those generated by proteolysis of protein mixture.[29] With the evolution of liquid chromatography technology, LC/MS has been of great importance to proteomics, since sensitivity and selectivity are further improved due to enhanced sampling of peptides into the mass spectrometer.[11] In proteomics, two-dimensional or even three-dimensional chromatographic separations of peptide mixtures are used in LC-MS/MS, normally a combination of RPLC, HILIC, ion exchange or affinity chromatography separations.[29]

4.1.2 PROTEOMICS STRATEGIES

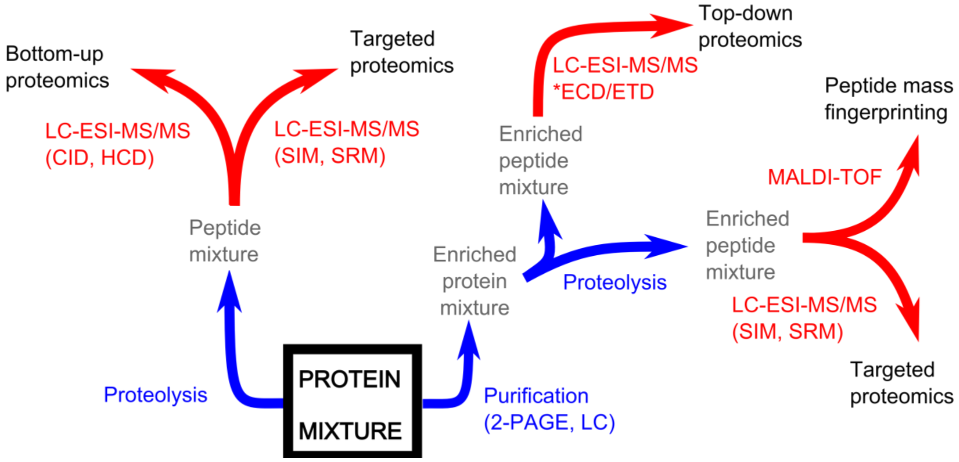


Figure 3. Summary of proteomics workflows to analyze a sample containing a protein mixture. In blue preparation steps prior to detection technologies (red)

There are two main approaches in the field of MS proteomics:[11,29]

- Top-down proteomics consists in the implementation of MS analysis sampling intact proteins, its main advantage is that it reduces the ambiguities tied to the identification of bottom-up or peptide mass mapping approaches, plus it is capable of distinguishing between isoforms. Normally, top down proteomics is applied to highly purified protein samples, e.g. single proteins or simple protein mixtures. Therefore, the source of proteins for top-down analysis may be either a 2DE or 1DE combined with specific LC fractionation.[31]

- Bottom-up proteomics, also shotgun proteomics, is used in the case of high-complexity samples for large-scale analyses, normally applying to quantitative strategies. In bottom-up proteomics, proteins are enzymatically digested into peptides prior to mass analysis, then peptide masses and sequences are used to identify corresponding proteins. Drawbacks of the bottom-up approach include limited protein sequence coverage by identified peptides, loss of labile PTMs, and its main bottleneck which is the elucidation of protein identification from peptide sequences.[31]

4.1.3 QUANTITATIVE PROTEOMICS

In order to compare across different sample groups, MS peptide abundance measurements must be somehow normalized. In quantitative proteomics, samples containing differently labelled proteins or peptides are combined and analyzed by LC-MS/MS for the purpose of identifying what proteins are contained in the sample and determining their relative abundance between groups. Although relative quantification reveals changes in protein levels between two or more states, it is dimensionless and is normally expressed in the form of ratios. Recent advances in MS and bioinformatics now allow the estimation of the absolute amount of proteins: that is, the copy number of proteins per cell. [30]

There is a great collection of methods for protein labelling in quantitative proteomics, that can be divided depending on the mechanisms that are used to label sample proteins:[30]

- In 'stable-isotope labelling with amino acids in cell culture' (SILAC) proteins are labelled metabolically by culturing cells in media that are isotopically enriched (for example, containing 15N salts, or 13C-labelled amino acids). Potentially all peptides can be labelled and the absence of any chemical steps make the method easy to apply as well as compatible with multistage purification procedures.[29]

- In the case of isotope coded affinity tag (ICAT), proteins in each condition are labelled with chemical probes that consist of three elements: a reactive group for amino acid labelling, an isotopically coded linker, and a tag (e.g., biotin) for the affinity isolation of labelled proteins/peptides.[11]

- In isobaric labelling methods, such as iTRAQ, dimethyl labelling or TMT-labelling, proteins are labelled at specific sites with isotopically encoded reagents. The reagents may also contain affinity tags, allowing for the selective isolation of the labelled peptides after protein digestion. Thus enabling their selective isolation and analysis of to specific functional groups or protein classes.

- In enzyme-catalyzed stable isotope labelling proteins are tagged by means of an enzymatic incorporation of 18O from 18O water during proteolysis in the presence of deuterated water. In this method, peptides are labelled at the carboxy terminal when digested.

On the other hand, label-free protein quantification is also an option, yet label-free quantification through spectral counting and/or signal intensity of the detected peptides seems to be the less ideal way to obtain quantitative information, as it suffers from all the drawbacks of mass spectrometry in terms of variability and linearity of the detector's response.[30]

Additionally, absolute protein quantification can be achieved by **targeted proteomics methods** using SRM, MRM or PRM methods to enable reproducible, sensitive and selective protein assays, that are relatively faster to develop and deploy than immunoassays. Even though they are limited to a lower coverage, as it is restricted by manual curation of proteolytic peptides for each targeted protein, it is in principle capable of distinguishing highly similar proteoforms such as isoforms, post-translationally modified proteins and genetic variants. [11]

4.1.4 PROTEIN IDENTIFICATION

The main issue in protein identification in proteomics results from the challenge of elucidating peptide sequences from LC-MS/MS spectra and their subsequent inferred protein entities. The identification of a protein by elucidating the series of fragmented peptides is a complex task tied to stochastic statistics and probability. In high throughput experiments, this process is performed "in silico" in two steps:[29,30]

- The extraction of peptide sequences from CID MS/MS spectra can be accomplished by either searching into existing peptide spectral databases, or alternatively it can be achieved by directly determining the sequence of peptides by analyzing the spectra. Based on stochastic rules of peptide fragmentation and the fact that amino acids have a known fixed mass, a strategy known as "de novo" sequencing.
- Protein identification is solved by algorithms that infer protein identifications based on the peptides sequences detected. This task entails distinguishing correct peptide assignments from false identifications. Incorrect peptide assignments can be prevented by applying filtering criteria based upon database search scores.

However, the rates of false identifications that result from such filters are hard to determine. Consequently, it is therefore important that computational approaches use robust and transparent statistical principles to estimate accurate probabilities indicating the likelihood of identification. The most common approach is the combination of precursor m/z and its fragment ions that are matched to known peptide sequences from large protein databases using search algorithms such as Mascot or SEQUEST, and then use decoy search strategies, in which the MS/MS spectra are competitively matched against random databases to estimate the rate of false positive identifications.

4.2 Result interpretation

Proteomics experiments often generate a vast amount of data. However, the simple identification and quantification of proteins is not sufficient for the full understanding of complex mechanisms occurring in the biological systems. Protein functional annotation through computational tools now occupies a place as important as the protein identification itself. There are different strategies for protein contextualization:[32]

- **Gene ontology-based annotation.** Genome and proteome ontologies are designed with hierarchical classes, communicating definitions with clarity and objectivity, however, keeping extendibility. Ontologies for genes and proteins usually describe the classification of the molecules according to their role in the biological systems, using controlled vocabulary. As a result, collections of protein entities are associated to relevant biological information such as functionality or cellular compartmentalization.

- **Functional enrichment analysis for identification of overrepresented biological mechanisms.** This approach increases the probability of identifying the most pertinent biological processes related to a biological mechanism under study. The goal of this approach is to summarize the biological processes and pathways that are most likely altered given the protein profile. Enrichment scores can be calculated by statistical methods, including Chi-square, Fisher's exact test, Binomial probability and Hypergeometric distribution.

- **Integration of functional annotations through biological network analysis.** The main purpose of this approach is to facilitate the visualization of the results. Proteomics data is displayed in networks built from knowledge-based databases that contain different biological layers of information, such as gene expression and co-expression patterns, protein–protein interactions and coregulation.

5. METABOLOMICS

The metabolome is defined as the collection of small molecules produced by cells, metabolites, it is therefore a direct source of information to study the biochemistry of organisms. Unlike genes and proteins, whose function is respectively subject to epigenetic regulation and post-translational modifications, metabolites serve as direct signatures of biochemical activity and they are therefore easier to correlate with a phenotype.[33]

In this context, metabolomics has become a powerful approach that has been widely adopted for clinical diagnostics, revealing molecular mechanisms underlying drug action, dietary changes, exercise intervention and psychosocial impacts of environment. Moreover, it plays a vital role in systems biology for elucidating the function of unknown genes and enzymes.[19]

5.1 Metabolomics strategies

The main concern in the experimental design of a metabolomics study is actually deciding the number of metabolites that should be measured. This somewhat depends on whether it is possible to localize the source of metabolic variation, aka metabolic pathway(s) that are affected by a given perturbation or challenge. The type of approach used will determine the subsequent analytical protocols.[33]

5.1.1 UNTARGETED METABOLOMICS

Untargeted metabolomics methods aim to simultaneously and unbiasedly measure as many relevant metabolites as possible from biological samples. This is usually reserved to biologically relevant molecular features, in other words, metabolite ions that show differential abundance profiles detected by rigorous data preprocessing and statistical analysis.[33]

The uniqueness of untargeted metabolomics is that it suits "hypothesis-free" testing for discovery of unknown or poorly characterized metabolites of some type of biological or clinical significance. In summary, untargeted metabolomics offers novel insight into complex metabolic mechanisms of drug/toxin action or characterization of complex diseases.[19] Hyphenated-MS platforms have become the setup of choice for most untargeted metabolomics experiments given their higher sensitivity. In relation to instrument setups for MS untargeted studies, metabolite quantification requires a compromise between sample throughput, metabolome coverage, peak quality and peak capacity.[19] While the equipment used in MS untargeted metabolomics varies, chromatography instruments are normally coupled to high-resolution MS equipment that operates at a high scan speed since it is ideal to record the finest peak profiles possible during the chromatographic run. Given the limitations of coverage of each chromatography method, in most cases, orthogonal chromatography separation modes are used in parallel to analyze the same sample as a way to expand metabolome coverage. In the case of polar and nonpolar metabolites, these include HILIC-LC/MS and RP-LC/MS, in aggregate with GC/MS.[34]

5.1.2 TARGETED METABOLOMICS

Targeted metabolomics refers to any analytical method in which a specified list of metabolites is measured. Targeted approaches are commonly driven by a specific biochemical question or a hypothesis that motivates the investigation of a particular metabolic pathway. Methodologically, targeted metabolomics involves an important analytical effort in order to optimize chromatographic separation and ionization responses, commonly done using a commercial standard or a pure extract of the targeted metabolites.[42] Targeted approaches provide a highly sensitive and robust method to proficiently measure dozens of biologically important metabolites with relatively high throughput, therefore analytical platforms used in targeted metabolomics require good sensitivity and specificity in order to measure of low-concentration metabolites that are difficult to detect with alternative methods.[19] Additionally, targeted methods are quantitatively reliable and hence they are sometimes used to achieve absolute quantitation.[33] Absolute quantification can determine the exact molarity for given metabolites in a sample, yielding analytical results independently of the experimental context. To achieve absolute quantification, a calibration curve for each targeted metabolite needs to be constructed using a range of several dilutions.

5.1.3 SEMI-TARGETED METABOLOMICS

It is defined as an "expanded" targeted metabolomics approach that has been proposed for quantitative analysis of hundreds of known standard metabolites, often across multiple analytical platforms, in order to cope with the drawbacks of untargeted approaches while retaining comprehensive metabolome coverage.

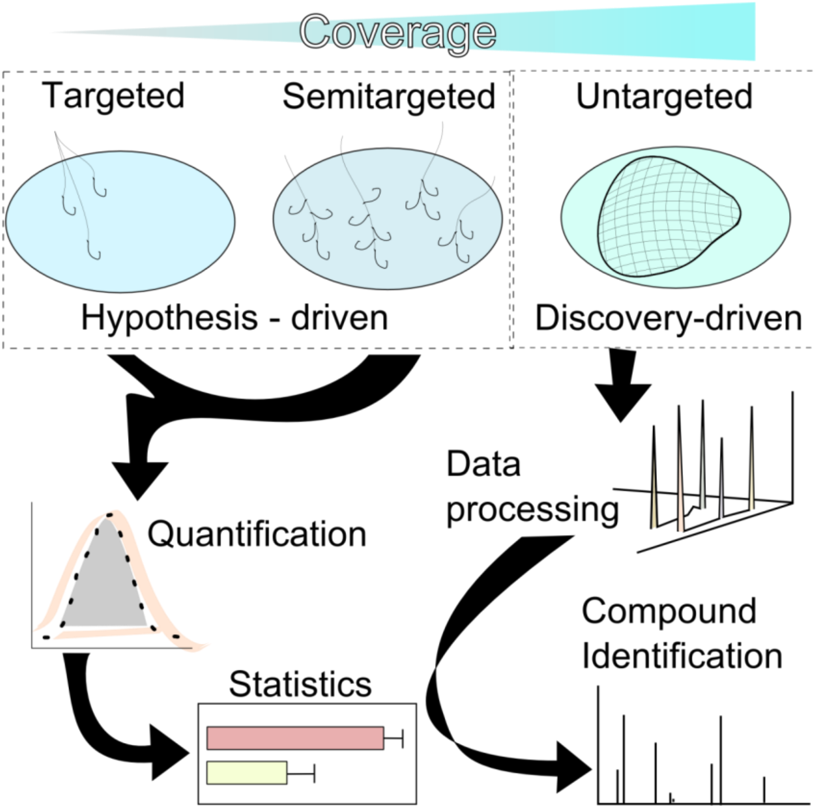


Figure 4. Metabolomics workflows, fishing for biomarkers. Targeted quantification methods are represented as fishing hooks, while untargeted approaches are similar to a fishing net.

5.2 Sample preparation

An indispensable element of any metabolomics protocol involves the extraction of analytes, metabolites, from the studied samples, such as body fluids, cells, and fresh or fixed tissue. The goal of extraction is to obtain quantitative yields of metabolites in the sample, and remove impurities, i.e. proteins.

Extraction protocols may vary depending on the experimental design, the nature of the sample, targeted molecules to be detected and the analytical platform used. An optimal metabolite extraction method ultimately leads to a higher extraction efficiency and analytical sensitivity. However, an increased number of preparatory procedures and fractionation may reduce the analytic throughput.[34] There are two main types of extraction methods: liquid-liquid and solid phase extraction. The former is widespread used in both targeted and untargeted approaches because of its superior throughput and reproducibility.[25]

Extraction methods are designed to exploit a property which differs in the analytes of interest and the rest of extraneous material. In summary, sample preparation includes:[15]

- Enrichment for metabolites of interest: The most common extraction solvents include a combination of water, methanol, acetonitrile, methyl tert-butyl ether and/or chloroform.
    - It is important to consider that sample extraction targets a subset of the metabolome to be extracted from raw specimens. Extraction protocols are normally classified depending on whether the metabolites to be extracted are (a) polar or (b) non-polar. In other words, no single extraction method satisfactorily covers the entire metabolome.[15]
- Removal of impurities that interfere with analysis such as proteins and salts: Protein precipitation is generally achieved by mixing with cold solvents such as acetonitrile, ethanol or methanol in acidic or basic pH, in order to cause denaturation and thus reduce their solubility.

5.3 Mass spectrometry-based metabolomics

In MS metabolomics, the nature of the data detected depends on multiple factors. First metabolomics MS techniques can be divided depending on the how the sample is injected in the mass spectrometer.

Samples can be directly injected to mass spectrometers, a technique known as direct infusion (DI-MS).[35] Alternatively, a separation technique can be used previous to MS analysis, this family of techniques are known as hyphenated MS. Analytes that enter the mass spectrometer are previously separated by a gas-, liquid chromatography (GC and LC, respectively) or capillary electrophoresis (CE).[33] It must be considered that, similar to the diversity of conditions for metabolite extraction, no single chromatographic method is suitable for all classes of metabolites. Since the metabolome is composed of a diverse array of compounds with different physicochemical properties, no single retention mechanism is adequate to resolve complex sample mixtures that vary widely in their polarity, charge, and stability.[15]

Finally, metabolites found in biological samples, such as tissue or organ slices, can also be ionized from a solid surface, or solid-like matrix, providing information on the physical localization of analytes in sample as the spectra is associated to a certain area of the surface, SIMS, DESI and MALDI are MS techniques used to that end, for example in the field of "tissue imaging".[36,37]

5.3.1 LC/MS METABOLOMICS

LC/MS metabolomics is the method that brings the widest metabolome coverage, from small metabolites to large lipids. Such immense coverage comes from the variety of chromatography columns together with recent developments in sensitivity and mass accuracy, LC-MS/MS have been applied to high accuracy instruments typically used in untargeted metabolomics.[38] Normally, LC/MS instruments used in metabolomics are UHPLC systems coupled to hybrid high resolution instruments, such as the qTOF and Orbitrap, in which the standard method of ionization is ESI.

Recent developments in LC column technology have greatly improved separation efficiency while expanding selectivity, yet chromatographic separations are not universal for the metabolome.

Therefore, different combinations of chromatography column and mobile phases setups must be used to increase the coverage of any LC/MS metabolomics experiment. The most common chromatographic separation is RP, but lately metabolomics interest in HILIC has grown given its complementarity to RP analysis. Additionally, the separation of metabolites can be tuned by modifying the chromatographic separation time or the use of mobile phase modifiers (ammonium formate or ammonium acetate), so as to improve the separation stability and efficiency of LC/MS, and improve the MS detection sensitivity.[15]

5.3.2 GC/MS Metabolomics

In the case of low molecular weight metabolites, GC/MS has revealed as a method of choice to detect not only volatile but also non-volatile polar and non-polar metabolites, though its coverage is still significantly lesser than in LC/MS.

GC/MS metabolomics relies on the application of chemical derivatization to analyze most polar and lower weight metabolites. Chemical derivatization can be applied to obtain a broader metabolome coverage to increase metabolite stability and volatility. There are a few derivatization methods, yet the most common approach involves a two-step procedure, first the methoximation of the ketone groups and the silylation of all protonated radicals (alcohols, carboxylic acids, amines, thiols, and phosphates). Interestingly, trimethylsilylation remains as the most widely used derivatizing procedure in GC/MS-based metabolomics given its general applicability to various classes of polar metabolites.[19] Though, there exists no derivatization method that leads to one derivative per metabolite for all compound classes.

Even though there is a great variety of column lengths, diameter and capillary types, 30 m long 0.25 mm internal diameter 95% dimethyl/5% diphenyl polysiloxane, and similar, are the standard choice for GC/MS metabolomics.[39] In the case of MS detectors, TOF and single quadruple are more common, yet higher resolution instruments, such as Orbitrap, may also be used.[40,41] Unlike ESI ionization used in LC/MS, Electron Impact (EI) is the standard ionization method used in GC/MS, that does not suffer from ion suppression or differential adduct formation.[15] Besides, the fragmentation spectrum generated by EI is highly reproducible and an analogue to tandem mass spectra in the case of metabolite identification.

5.3.3 Characteristics of metabolomics data

The extraction of metabolite abundance out of spectral data is an important task in metabolomics. While data analysis in targeted metabolomics is nowadays rather a manual and demanding task aided by commercial software regularly distributed along with the instruments. It is, in contrast, truly a challenge in untargeted metabolomics since the datasets are exceedingly complex in terms of the knowledge of data analysis needed to extract biological information.[25] Nonetheless, it is accepted in the community that quality of data processing can be difficult to assess since there is not any standardized benchmark, thus it relies on data analysis traceability and reproducibility.[43]

In MS metabolomics, acquisition instruments generate data files which contain a collection of successively recorded ion histograms. Each histogram represents counts of ionized molecules impacting the detector during a short time frame (scan), each ion is recorded as a value of mass-to-charge ratio (m/z) eluting at a certain time point.[43]

Due to the intrinsic dynamics of ionization events, ion quantification and limitations of chromatographic separation, untargeted metabolomics data has a few inherent properties that make it noisy and redundant, these features vary depending on the experimental design and platform used, yet they must always be considered in untargeted metabolomics data analysis pipelines :

- **Adducts**. A common issue in MS spectra is the presence of adducts during the event of ionization. Especially in LC/MS where electrospray ionization (ESI) during the desolvation process in the source, yet it also occurs in the rest of soft ionization techniques. This occurrence, together with the detection of natural isotope variants of molecules, increase the number of recorded ion m/z peaks, meaning that spectra contain many more ion features than actual compounds.[44]

  Adduct formation cannot be controlled, yet it can be favored by controlling the composition of the mobile phase. Adduct type is molecule dependent and adducts themselves have different sources depending on: salt concentrations, mobile phases, solvents, impurities from glassware or the sample matrix itself.[44] Dimers or even trimers may appear predominantly in the source, which are concentration dependent, plus they are highly expected if hydrogen bonds are possible within the molecule.[44] Finally, some ions may appear multicharged, increasing the challenge of compound annotation.
- **Isotopes**. It is common to detect naturally occurring isotopes in MS metabolomics, especially in the case of 13C, yet in very high resolution instruments 15N, 18O or 32S must also be considered among other atom isotopes. Their presence correlates with molecular ion abundance and they can potentially mask other ion signals.[25]
- **Retention time shifts**. The chromatographic separation equipment is inherently sensitive to changes in temperature and atmospheric pressure, meaning that in datasets with a large number of samples, metabolite retention time is inevitably subject to variation. Additionally, retention time shifts may result from anomalies in the flow rate, mobile phase composition and column age.[44]

- **Mass accuracy**. Due to detector limitations, mass spectrometers have a window of accuracy when measuring the mass-to-charge ratio of ions, meaning that the true value is within an error tolerance that depends on the instrument type and its calibration.
- **Noise**. MS spectra may also contain a variable amount of chemical noise that has different sources: impurities from buffers and solvents, column/tube degradation species, especially in the solvent front and at the end of the run during column washing. In addition, spectra contain random noise, typical of electronic devices, which can be attributed to the detector response variation caused by its maintenance status, atmospheric pressure or temperature.[43] The main drawback of noise is that it may shift the signal baseline, altering the threshold of intensity that differentiates noise from true ion signals.[44]
- **Performance variation over time**. Closely related to RT shifts and noise, it defines the event of a gradual loss of response in the MS detector. It is a common side-effect of instrument usage and contamination of the detection hardware, the MS counts recorded may vary over time, leading to false intensity records that are lower than expected.[25]

5.4 Data processing

5.4.1 RAW DATA TRANSFORMATION

Data processing consists in low-level processing of raw data using signal processing methods and combining data between measurements. These tasks transform raw series of scans into a format that is easy to use in the subsequent data analysis steps.[43]

In general, MS instruments store data in vendor specific formats that are designed to be used in the corresponding vendor software. However, the community made an effort to design a unique open data format to standardize open source software for data analysis, this format is known as mzML, previously mzXML[45] The transformation of vendor formats into mzML is a common first step in metabolomics data processing.

Using mzML file format, raw data can be imported into data processing software or coding environments. In these, raw data points are transformed into simplified data structures, e.g. matrix and additional metadata that facilitate easy access to characteristics of each observed ion. These characteristics include intensity measurement, m/z and elution time, commonly named retention time, of the ions recorded in each raw data file. The matrix obtained contains three dimensions: m/z, retention time and intensity; different in each sample. The difficulty in this process originates from the fact that algorithms must account for considerable variations in their chromatographic peak shape, peak time span and m/z accuracy.[46]

A large collection of software exists for untargeted metabolomics data analysis, each with its own strengths and weaknesses, yet most of them involve some, or all, of the following data processing steps:[43] base line removal, peak detection/filtering and retention time alignment.

5.4.2 PEAK DETECTION

The purpose of peak picking, or mass trace detection, is to extract as many recorded signals, caused by true molecular ions, i.e. metabolites, as possible. This is done by sequentially checking for ion recordings in which their intensity fits to a model of an ideal shape of an eluting compound, this model normally includes a gaussian-like function. This step also aims to provide accurate quantitative information about ion abundance.[43]

Noise filtering/smoothing and baseline removal are essential steps of peak detection, they involve mathematical noise reduction methods designed to remove random noise from the measurement signal while finding the signal baseline shape and finally subtracting the shape from the raw signal.[43] These methods are typically implemented using traditional signal processing algorithms such as: moving average window, median filter, Savitzky-Golay local polynomial fitting and wavelet transformation.[47]

On the other hand lies deconvolution, an algorithm-based process designed to reverse the effects of convolution, a mathematical operation that combines two differently shaped functions. Deconvolution methods have been typically used in metabolomics to extract different compounds that coelute forming a complex elution peak. The main challenge in the use of deconvolution methods is that complex biological matrices lead to a large number of overlapping peaks, with similar retention times and overlapping isotope patterns. Deconvolution algorithms utilize the assumptions that different fragments from the same molecule have the same retention time as well as that their profiles across multiple samples are highly correlated since they are subject to the same biological variation and systematic error.[47,48] Such methods have been widely used in GC-EI-MS data processing, and also have been applied to LC-DIA-MSMS. In a way, deconvolution is in itself a peak picking and retention time adjustment approach altogether. This signal processing technique estimates the relative area corresponding to each individual peak when multiple peaks overlap within the same spectral region.[47] This strategy involves finding raw data components that are subsequently matched over the different observations, but its main requirement is the presence of several mass traces that converge at the same retention time for deconvolution, thus ideal for DIA-MS and EI-MS.[49]

5.4.3 SAMPLE ALIGNMENT

Spectral alignment is one of the main processing steps in untargeted studies involving datasets with tenths or hundreds of samples. When analyzing multiple samples, the retention time of compounds corresponding to the molecular ion peak may be affected by non-linear shifts.[47] Spectral alignment methods must be therefore applied to correct this undesired variation in the samples that can profoundly affect the quality of the study, they can be divided in two main groups depending on whether the algorithm uses a reference list of peaks to perform the alignment. However, the choice of alignment method usually dictates the type of required downstream data analysis. Some alignment methods allow the analysis of aligned raw signals directly for finding differences between samples, while other alternatives lead to multivariate data analysis.[43]

Referenced alignment algorithms involve the use of an existing peak list to direct the RT correction of the shifts. This reference list may either be one of the other samples or a meta sample created by mixing all previously detected peaks. Pair-wise methods use raw data as input material and generate a set of mappings that transform retention time axis of each run to a common retention time axis by aligning either pairs of samples or multiple samples against a selected reference sample or a template.[20,43]

- Alignment of total ion chromatogram curves (TIC)
- Peak-based alignment.
- Internal Standards as alignment reference

Normally, referenced algorithms involve variations of Warping methods (COW, PTW, DTW, STW,…), which are based on the application of a non-linear transformation to the retention time axis in order to maximize the correlation between the spectra[20]. Non-referenced or reference-free methods avoid the biases introduced by using reference spectra, but at a cost of being more computationally intensive. In these, the alignment is performed over all the spectra or by splitting the spectra into smaller segments and independently align each resulting segment.[43,50]

- Binning and clustering methods.

- Segmenting by applying a constant shift to all spectral points.

5.4.4 FEATURE CORRESPONDENCE

In untargeted metabolomics, the term “feature” is used to describe a data variable that is potentially originated by a metabolite. This term is commonly used in MS untargeted metabolomics to refer to the m/z, or m/z-rt pairs, that are extracted after data processing. In other words, in metabolomics, features are aligned ions that are consistently detected in most samples meaning that they are an actual metabolite with high certainty.[20]

In metabolomics, the process of grouping sample ions with similar m/z values that have a reproducible elution across multiple experimental runs is known as “correspondence determination”, an important task in order to supply the subsequent statistical analysis. The main difficulty within correspondence determination is that, even if previously aligned, m/z values still contain accuracy error. This is similar to the case of retention time alignment, the m/z variation is non-linear itself, meaning that the objective of correspondence is to “align” peaks across samples using both m/z and RT feature values, which is also generally performed with bidimensional warping algorithms or similar methods.[51]

5.4.5 DATA ANALYSIS

Data analysis in untargeted metabolomics involves the transformation and filtering of beforehand processed data, and it is followed by hypothesis testing using statistics. Given the complexity of hyphenated-MS samples, thousands of mz-RT features can be extracted at the end of data processing. However, most of them are not originated by an actual metabolite due to the aforementioned reasons, in other words, during the preprocessing steps, there is no distinction between adduct peaks, natural isotopes, in-source fragments or other impurities. Therefore, careful data analysis is necessary to avoid misleading results.[52]

Metabolomics data analysis mainly consists in the selection of a subset of detected features that comply within certain thresholds of variation across samples and minimum abundance/intensity. Additionally, before any statistical testing is applied, sample quantifications may be necessarily adjusted, which involves data normalization, scale transformation or value imputation if necessary.

FEATURE FILTERING

The following filtering strategies are applied to extract a subset of features that are above an acceptable intensity threshold, since MS/MS experiments are required in untargeted metabolomics for the ultimate step of identification that depends on signal intensity for reliable results. Additionally, low variation (CV) features are also ignored due to that they potentially lack biological importance.[52] And finally, if isotope annotation is used, peaks corresponding to non-monoisotopic signals can be ruled out.

- **Intensity filter.** In order to reduce the number of features that may be passed to statistical testing, one of the first filters involves a simple abundance threshold. The reason for this filter is that, since features of interest (i.e. statistically significant) need to be identified with tandem mass spectrometry, fragmented ions need to have an acceptable intensity otherwise the quality of fragmentation spectra may be compromised. Additionally, this filter serves to eliminate features that are close to noise level.
- **Variation filter.** As mentioned above, the use of chromatography systems and mass-spectrometer instruments implies deviations in metabolite retention time and their quantification response.17,19 To

assess the variability of analytes, metabolomics researchers came up with the idea of continuously injecting identical samples in-between other samples within the chromatographic run 25,52. These samples, known as quality control (QC) samples are considered identical to biological samples, i.e. with equivalent matrix composition by pooling small aliquots of all or representative samples of the study.25

FEATURE ANNOTATION

We must consider that in an untargeted metabolomics study, feature annotation is key because of the high redundancy of hyphenated-MS data, where different features may represent the same metabolite.[53] In particular, it is important to consider that this event, called feature degeneration, is a natural consequence of soft ionization sources such as ESI, which increase the complexity of untargeted metabolomics data due to the contribution of isotopes, adducts and in-source fragments.[54] Feature annotation is therefore essential to reduce false positives and to remove redundant information. Feature annotation leads to grouping features that originate from the same compound which at the same time gives valuable chemical information for metabolite identification.[55] Isotopes can be detected in any type of MS method, though it mainly depends on the signal-to-noise ratio. Sensitivity and resolution have a great impact on the isotopic envelope of metabolites and adducts can increase the feature redundancy.

The ultimate objective of most feature annotation algorithms is to detect features that correspond to monoisotopic peaks of frequent adducts (for example H or NH4 adducts), which are particularly predominant in spectral databases. A large collection of software to perform adduct annotation, and isotope annotation, in untargeted studies have been yearly developed: CAMERA, CliqueMS,… Their annotation algorithms normally include three main steps:[55–59]

- Peak grouping/clustering by:
    - Peak shape correlation.
    - Peak-abundance correlation.
- Cluster peak annotation.
- Neutral mass annotation.

5.5 Compound identification using tandem mass spectrometry

Feature identification is the ultimate goal in untargeted metabolomics, though at the same time it is the main bottleneck for definitive results. Feature identification consists in the identification of experimental features using MS/MS fragmentation in order to translate them into a molecule entity. Performing "a posteriori" targeted or DDA MS/MS experiments are, in practice, not more challenging than a full scan experiment. However, the interpretation of the resulting spectra is indeed challenging and, to a certain point, hard to reproduce/standardize.

5.5.1 IDENTIFICATION CONFIDENCE LEVELS

The Metabolomics Standards Initiative (MSI) was conceived in 2005. The early efforts of MSI were focused on community-agreed reporting standards, which provided a clear description of the biological system studied and all components of metabolomics studies.[60]

The MSI Chemical Analysis working group defined four different levels of metabolite identification to classify the level of identification confidence in the literature, these four levels were later update to five by the Compound Identification work group of the Metabolomics Society at the 2017 annual meeting of the Metabolomics Society. These include different levels of confidence in identification:[61]

- **Level 1.** Valid identification. Requires the identification to be performed with a minimum of two independent and orthogonal properties, the identified compound must be compared against a pure reference standard under identical analytical conditions. This level assures unambiguous 3D structure, including full stereochemistry.

- **Level 2.** Incomplete identification. This differs from Level 1 in the fact that the reference standard is not analyzed under identical conditions, this is the case of comparing acquired MS/MS spectra against an equivalent found in spectral databases. This level at least confirms 2D structure.

- **Level 3.** Putative identification. Identifications at this level are not confident enough to supply a metabolite identity but may supply information about the formula and compound subclass or family, for example in the case of lipids, phosphatidyl-choline family. This level has some certainty about compound formula and part of its structure.

- **Level 4.** Molecular formula annotation. Possibly accompanied by adduct and charge information. These arise when accurate mass and isotopic distribution patterns produce tentative structures from database searches. Note, a single molecular formula typically renders multiple candidate structures.

- **Level 5.** Unknown compound. Completely unidentifiable features, whose MSMS spectra or formula are impossible to annotate.

5.5.2 GC-EI-MS IDENTIFICATION

In most GC/MS metabolomics experiments, EI-MS setting 70 eV as ionization energy and in positive ionization detection mode is used as a universal ion source. These fragmentation patterns are highly reproducible not only in m/z values but also their relative abundance, which increases the accuracy and reliability of the peak assignment by comparison with spectral libraries.[39,40]

The community-standard workflow for GC/MS metabolite profiling typically implies working with deconvoluted mass spectra and identified compounds instead of unidentified peaks, although an unbiased approach based on peak picking followed by manual spectra revision is also acceptable, yet identification then becomes a harder task and normally relies on manual revision of the raw spectra.

A further level of effort for powerful identification results involves the use of retention indices (RIs) based on aliphatic carbon numbers by Kovats, that consists in a standardized collection of alkanes that are injected and thus their retention time used as a reference to normalize the adjacently eluting alkanes. Kovats RIs have been progressively included in mass spectral databases commonly used for GC/MS metabolomics. In combination with EI spectral matching, RI significantly improves metabolite annotation.[62]

Independently of the approach used for data processing, the spectra databases commonly used in GC/MS-based metabolomics are NIST (http://www.nist.gov/srd/nist1a.cfm), the Golm Metabolome Database (GMD, http://gmd.mpimp-golm.mpg.de), and the Fiehn BinBase library (http:// fiehnlab.ucdavis.edu/db), which contain mainly TMS-derivatized metabolites. In addition, HMDB, MassBank and the Madison Metabolomics Consortium Database (MMCD) also contain EI spectra and RI data.[62]

5.5.3 IDENTIFICATION IN LC/MS METABOLOMICS

Ideally, the identification procedure involves the comparison against an MS/MS spectrum of a pure compound corresponding to the expected metabolite hit, so-called compound reference spectrum. However, in practice, levels 2 and 3 are the most commonly reported since level 1 is out of financial reach for the majority of research groups. In general, experimental MS/MS data is searched against a reference MS/MS database, if a definitive match, this is the most conclusive evidence for validating the identification of a metabolite feature using MS/MS[62]. However, prior to MS/MS comparison, a few considerations must be taken when checking whether the experimental MS/MS spectra match that of a reference spectrum found in compound spectral databases. Different criteria of the acquisition methods must be taken into account to consider whether a pair of MS/MS spectra are comparable:[8]

- **Acquisition instrument type.** QqQ, QTOF, ion trap and Orbitrap produce different fragmentation patterns.
- **Mass accuracy.** This affects the precursor ion as well as its fragments, and it depends on the instrument setup used to obtain both the experimental and reference spectrum.
- **Ionization mode and energy.** Polarity, fragmentation method (CID, HCD) and energy used have an impact on MS/MS fragmentation patterns.
- **Adduct of the precursor ion.** Proton adducts overpopulate MS/MS spectra databases but depending on chromatographic conditions other adducts may be detected (+NH4 or +Na)[40]. Plus, some metabolite families are prone to prefer non-proton adducts.

Table 3. Metabolomics databases for identification. Extracted and adapted from Kind et al.[8]

Name	Number of MS/MS spectra	Number of compounds	Available online	Freely available	Instru-ment diversity
NIST14 MS/ MS	193,120	9,344			**+++**
METLIN 2019	> 440,000	>22,000	**+**		**+++**
LipidBlast	212,516	119,200			**+**
MoNA	194,000	68,700	**+**	**+**	**+++**
mzCloud (Thermo Scientific)	182,000	2,800	**+**		**+**
MetaboBASE (Bruker)	26,000	13,000			**+**
GNPS	212,230	12,694	**+**	**+**	**++**
ReSpect	9,000	4,000	**+**	**+**	**+**

MSforID	20,000	1,200			**+**
HMDB	22,198	114,008	**+**	**+**	**++**
NIST17 MS/MS	574,826	13,808			**+++**

5.5.4 AUTOMATIZING METABOLITE IDENTIFICATION

AUTOMATIC IDENTIFICATION

The validation of fragmentation spectra, which relates to both EI and MS/MS spectra, has been traditionally performed in a manual manner by experienced structural elucidation analysts. In summary, a positive identification depends on the similarity of the fragmentation patterns when comparing the reference and the experimental spectra, thus confirming the source of the detected fragments and explaining their relative abundance.[61]

However, given the new capabilities of DDA systems, it is now possible to obtain hundreds of MSMS spectra in a single experimental run, meaning that the process of metabolite identification must be necessarily automatized. Automatic identification tools have been lately developed to speed metabolite identification by applying computational power. As mentioned above, they may start by filtering some candidates from a reference, or library, spectra that could match the conditions of the experimental challenge, basically based on mass accuracy and ionization parameters. Algorithms must take into account that experimental spectra may contain noise and impurities, therefore it is necessary that they unbiasedly calculate the magnitude of similarity between the experimental and reference spectra.

Fragmentation spectra comparison is achieved by using different score-based metrics that are typically calculated considering m/z-relative intensity pairs of the experimental spectrum and library spectra as well as additional parameters such as weighting functions:[8] probability match algorithm, Dotproduct, Jaccard index, Pearson similarity, Jeffries-Matusita distance, or random projection, to name a few.

Additionally, spectral databases are limited, and thus sometimes no reference spectrum is found for an experimental challenge. Then, even an experimented chemist would struggle to extract a compound identification by elucidating the spectra based on experience and chemical knowledge itself.

IN SILICO MS/MS

In order to overcome the challenge of underpopulated spectra[62], a few alternative computational tools have been designed to predict the MS/MS spectrum given a compound structure, known as “in silico” MS/MS spectra generation.

First, molecular structures need to be transformed into a computer readable format that codes its atoms and bonds into a string, nowadays InChI (and InChIKey) is the most used, followed by the older SMILES. These coded chemical structures are available on most spectral databases.[61] In silico MS/MS spectra generation is a daunting task and prone to miscalculations. Four general methods can be distinguished when generating spectra “in silico”, which can be based on:[63]

- **Quantum chemistry.** Uses chemistry principles, but takes longer computing.

- **Machine learning.** Requires diverse training sets but results are quick.

- **Heuristic algorithms.** Limited to certain compound classes, generally rule based.

- **Reaction chemistry.** Based on metabolic reaction pathways, but does not calculate relative fragment intensities.

Table 4. Bioinformatic tools for MS/MS "in silico" identification. Adapted and extended from Blaženović et al.[61]

Name	Fragmentation method
MS-FINDER	Rule-based + Quantum chemistry
CFM-ID	Rule-based + Machine Learning
MetFrag	Rule-based
ChemDistiller	Machine Learning + Quantum chemistry
MAGMa	Rule - based
CSI-FingerID	Machine Learning + Quantum chemistry
iMet	Reaction chemistry + Machine learning

6. METABOLIC NETWORKS

6.1 Networks and graph science

The history of graph theory started in 1736 with the work of the mathematician Leonhard Euler. Graphs are mathematical structures, matrices, used to model pairwise relations between objects. Formally a graph is a symmetric matrix that defines the connections, known as edges or links, between a set of nodes, vertices or points. There are two types of links: directed (unidirectional) or undirected (bidirectional). In many physical applications it is desired that the edges of the graphs support some weights, i.e., real numbers indicating a specific property of the edge. Much later, the term “network” has become a synonym for “graph” in real-world applications of graph science. Graphs are used in different fields of science, for example in quantum physics, statistical physics, electronic systems, oscillation mathematics,…

Complex networks are graph models used to model real world systems, they can be considered as the schematic representation of organized systems in a variety of scenarios, ranging from social and technological to biological and ecological systems, interaction networks or social networks. Examples of complex networks are: metabolic networks, trade networks, protein-protein interaction networks. Their study has become a major field of interdisciplinary research in which mathematicians and physicists have significantly contributed by creating new algorithms to the study of topological and dynamical properties of complex networks.

6.2 Networks in biology

The combination of omics high-throughput technologies has enabled the coordinated study of cellular components at the level of genes, proteins, metabolites, and molecular interactions that occur between them. These studies have resulted in the generation of large-scale datasets that have served as the foundation for the modelling/construction/scaffold of metabolic, regulatory, signaling, and protein-protein interaction networks.[64]

Network science has suggested that biological networks have two distinct structural properties. First, it has been shown that several of these networks are scale-free (most nodes have more connections than average) and they possess a “small world” property (most nodes can be easily interconnected - through other edges- even if they are not directly connected). Second, scale-free networks are suggested to have high error tolerance (tolerance against random failure, robustness to random node removal) and low attack tolerance (vulnerability to the failure/removal of the highly connected nodes).[64] In short, this indicates that the connections in biological networks are organized in a way that there is a high connectivity between most neighbors and that even if they can be structured into subnetworks, the removal of a single node can be circumvented through another path.[65]

In general, these types of networks can be classified on the basis of the nature of the interaction into two broad categories:[64]

- **Influence networks**, where the nature of links represent whether an interaction is present or not, or the type of interaction, such as in protein-protein interaction or signaling networks.

- **Flow networks**, where a specific variable such as mass or energy flow may be conserved at each node, such as metabolic networks.

In the context of biological networks, a metabolic network is a flow network that contains the information of how an organism converts carbon and energy sources using electron donors and acceptors in order to generate biomass, energy, and byproducts.[66]

6.3 Genome-scale metabolic networks

Metabolic networks can be built from different sources of information, the most comprehensive examples are genome-scale reconstructions, knowledge-based models built by a team of experts using information from gene sequencing experiments, that record a collection of metabolic reactions within a biological system.

Genome-scale metabolic models (GMM) are highly detailed data structures that can be used to build metabolic networks that contain known or predicted metabolic reactions in an organism and can therefore serve as functional databases of cell-specific metabolism. Most genome-scale models are annotated with curated gene-protein-reaction associations linking genes with enzymes. These gene–protein-reaction associations are formulated as boolean rules considering isozymes and subunits of protein complexes.[67] In short, GMM can be used to construct metabolic networks containing fine annotation about gene expression, enzyme kinetics and metabolic pathway connection.

The process of building a GMM is a laborious task, which involves iterative series of manual curation and database searching, the latter may be programmatically optimized for time saving.[68] First, a draft reconstruction is generated starting from the genome and including all the genome-encoded metabolic reactions of the targeted organism. The draft reconstruction also includes annotated enzyme, reaction, and pathway data from databases like KEGG[69], BioCyc[70], and BRENDA[71] that serve as sources of the GPR rules.[67]

The earlier GMM were built for microorganisms such as E.Coli. But later, reconstructions of human metabolism were attempted after the publication of the complete human genome sequences in 2004, which required a much larger number of pathways. HumanCyc (in 2004) and Reactome knowledgebase (in 2005) were the first successful attempts to build a curated collection of biochemical reactions in human cells. Later, the first models of human metabolism were published in 2007, these were Recon 1 and EHMN (Edinburgh human metabolic network) reconstructions. Importantly, the sequence of manual curation steps is needed to improve the draft reconstruction, by periodically gathering evidence to prove or disprove the presence of a reaction in the network of the organism. Maintaining models up-to-date improves predictability by expanding and correcting the network content based on emerging biochemical knowledge.

Later, in 2013, a large improvement of the number of metabolic processes was achieved with an updated version of Recon: Recon 2, a consensus model that includes all reactions from an updated version of EHMN, Recon 1, HepatoNet1, and a module for acylcarnitine and fatty-acid oxidation. The following years, Recon2 evolved into the version Recon 2.2 with improved reaction balances and updated gene-reaction associations.[72] Simultaneously, a larger reconstruction named HMR (Human Metabolic Reaction database) was built independently, based on HepatoNet1, Recon1, EHMN, Reactome, HumanCyc, KEGG and the Human Metabolic Atlas. HMR was then extended in 2014 to include lipid metabolism, therefore generating HMR2.

Later, in 2017, a new human metabolic model named iHsa was obtained as an expanded and better curated version of HMR2.[67] Recently, Recon 3D appeared as a result of incorporating HMR2 into Recon 2 together with a number of additional reaction sets, including reactions modelling host-microbe interaction, reactions for simulating drug effects on human metabolism, reactions for absorption of dietary compounds, reactions of lipid metabolism and reactions from metabolomics datasets. 3D protein structures, pharmacogenomics data and atom-atom mappings were also included in the model.[67]

Table 5. Genome-scale metabolic models. Extracted from Angione et al.[67]

Model	Genes	Metabolites	Reactions	GPR rules
HumanCyc	3209	1761	1716	-
Reactome	1180	1131	1216	-
EHMN	2492	3695	6216	-
Recon 1	1905	2766	3744	2307
HMR	3668	6000	8100	6015
Recon 2	2140	5063	7440	4821
HMR2	3765	6006	8181	6071
Recon 2.2	1675	5324	7785	4742
iHSA	2315	5620	8264	5961
Recon 3D	2248	5835	10600	5938

The fields of science in which metabolic networks can be applied are diverse:

- **Contextualization of high-throughput data.** By serving as a framework on which other omics data can be overlaid/mapped for contextualizing high-throughput data and aiding profiling approaches. However, a major challenge still lies in determining an optimal strategy for interpreting the annotated data.

- **Guidance of metabolic engineering.** Traditionally, metabolic engineering has been performed on a small scale through manipulation of a few genes to affect the yield of a target metabolite, in which enzyme modification is directed based on local metabolic knowledge. The inherent drawbacks of using local analysis tools to guide cell-scale metabolic engineering efforts have motivated the use of metabolic networks, bringing up the field of 'systems metabolic engineering'.

- **Directing hypothesis-driven discovery**. Metabolic networks represent summarized collection of previously confirmed hypotheses; therefore, they are ideal for the construction of new hypotheses. Metabolic networks have been used to frame investigations into specific biological questions, using a mix of traditional biological approaches and computational systems-level thinking, they enable systematic hypothesis testing and prediction.

- **Interrogation of multi-species relationships.** Metagenomics studies particularly have shown most ecosystems to be extremely diverse, while higher-eukaryotic biology requires the study of multi-cellular systems. Lately, an increasing effort has been put into modelling such interactions in metabolic network models.

- **Network property discovery.** Biological networks are the ultimate representation of scientific holistic thinking, to be able to detect molecular events or phenomena that would be undetectable by reductionist approaches. GMM have enabled analysis of whole networks rather than individual pathways or genes, and many computational techniques have been developed to probe network properties.

The field of computational systems biology has produced a rich array of methods for network-based analysis, but many of these methods produce results that can be difficult to link to observable phenotypes.

7. TRACING METABOLISM USING STABLE ISOTOPES

7.1 Stable isotopes

Stable isotopes such as 13C or 15N, have the same number of electrons and protons and consequently share the same physicochemical properties, yet they differ in mass due to a different number of neutrons. Among biochemically relevant elements, carbon, hydrogen, nitrogen, oxygen and sulfur all have one or more stable isotopes with measurable abundance in nature.[53] Isotopologues, that is metabolites containing stable isotopes and their unlabelled counterparts, have the same chemical formula and structure and hence generally behave identically during chromatographic separation.[53]

The ability to discover new metabolic pathways by following the distribution of heavy atoms in isotopologues through the metabolic network was exploited during the classical period of biochemical pathway identification, which are the basis of current isotope labelling or isotope tracing methods. During its roots these studies were mostly performed using radioactive isotopes even though they are hazardous, while stable isotopes are not.[53]

With the advent of high sensitivity MS instruments and the chemical synthesis of stable isotope enriched compounds, radioactive isotopes were retired by these novel and lesser dangerous homologues.

7.2 Stable isotope labelling

The major interest in stable isotope labelling for metabolomics stems from its ability to aid in the measurement of dynamic activity of metabolic pathways, in order to provide mechanistic explanations for the perturbations in metabolite levels observed in classical metabolomics studies. Stable isotope labelled precursors with uniform or positional labelling of constituent atoms have been used to generate biological samples that allow the study of their metabolic fate throughout the metabolism in an untargeted way. Similar tracer analyses have also been used for many years alongside metabolomics studies to study specific metabolic pathways of interest in a targeted manner.[73] In the context of systems biology, SIL provides an opportunity for reconstructing and validating both stoichiometric and dynamic computational models.

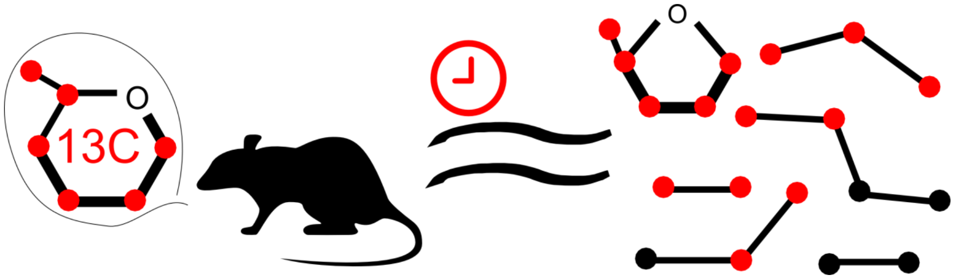


Figure 5. Stable isotope labelling.

Software solutions to routinely extract isotopologue abundances on a large scale are available. Biochemical interpretation of these data still requires a significant level of manual curation, but the growing availability of atom-resolved metabolic networks, where the source of specific atoms in each molecule can be traced from precursor substrates, offers the potential for computational approaches to biological inference and hypothesis generation based on network-wide isotope-labelled metabolomics data.[53]

Apart from the choice of detection platform and separation methods, stable isotope labelling metabolomics requires a few extra considerations given its particularities:

- The selection of labelled nutrient, also called tracer metabolite, depends on the study hypothesis and the known metabolic pathways in the organism of interest.

For example, hypothesis-free metabolomics studies that require extensive metabolite labelling utilize fully labelled carbon sources, such as U-13C-glucose. Other common stable isotope tracers include: U-13C-glutamine to measure TCA cycle anaplerosis, U-13C, 15N-glutamine to detect pathways for carbon and nitrogen assimilation, or 13C-bicarbonate to monitor CO2 incorporation from the atmosphere in the case of autotrophs, such as plants.[74]

- The kinetics of nutrient assimilation differ from system to system, an utterly important consideration in steady-state labelling studies. Systems must be characterized in order to assure that isotopologues reach a concentration high enough to be detectable in the quantification platform used. In particular, in the case of secondary metabolites and many macromolecules which need a significant higher culturing time with labelled nutrients, while labelling of central carbon metabolites may occur within seconds.[74,75]

7.3 Qualitative assessment of metabolic fluxes

Stable isotope tracing experiments can be used to follow the fate of atoms from a substrate downstream into biochemical reactions. Analysis of steady-state labelling patterns in downstream metabolites provides information as to their origin and rates of production. By computing fractional enrichments using stable isotope tracing experiments, a series of qualitative or semiquantitative flux predictions around each analyzed metabolites can be obtained. Of mention, direct interpretation of data from stable isotope tracing experiments cannot be used to quantify network-wide flux analysis. This approach stops short of modeling network-wide flux maps. However, it is relatively straightforward and has become increasingly accessible due to the widespread adoption of high-resolution mass spectrometers.[76,77]

7.4 Quantitative assessment of metabolic fluxes

Fluxomics is an interdisciplinary science that aims to quantitatively determine the turnover of metabolites through a network of enzymatic reactions (metabolic flux). Fluxomics integrates experimental measurements of metabolic fluxes with mathematical models to determine the absolute flux through the metabolic network.[78] Metabolic fluxes are governed by the interplay of gene expression, protein concentration, protein kinetics, regulation, metabolite concentration and thermodynamic driving forces. These metabolic fluxes are the central trait of cellular metabolism representing its integrated functional output response and consequently becoming the best signature of the cellular phenotype.[79] However, metabolic fluxes cannot be directly determined from other -omics data. Zamboni et al.[80] nicely illustrates this concept by comparing flux to traffic. Traffic, regarded as the mobility response within a city, results from the relationship among parking lots (mRNAs), roads (proteins), and cars (metabolites).

However, knowing the number of parking lots, roads and cars is not enough to inform on traffic (i.e., knowing whether cars are stalled or allowed to freely move). Similarly, quantifying molecular levels of mRNA, proteins and metabolites using -omics technologies does not provide a measurement of the metabolic fluxes. In fact, metabolic fluxes cannot be directly measured but need to be inferred from other observables.[74] They are usually estimated through mathematical modeling of metabolic reactions. In general terms, this modeling uses genome-scale metabolic reconstructions and linear programming to quantitatively estimate distributions of metabolic fluxes. Metabolic reactions enclosed in the reconstruction are formalized to a set of mass balance equations resulting from a matrix representation of the stoichiometric coefficients of the metabolites participating in each reaction. Assuming steady-state conditions, linear programming can be applied to compute a vector of quantitative fluxes that maximize a predefined objective function and solves this set of mass balance equations.[76] This model-based quantitative estimation of metabolic fluxes is challenging, particularly for large reconstruction models holding a far large number of reactions than metabolites. In such cases there is more than one solution to mass balance equations and quantitative fluxes remain undetermined. Here, experimental constraints other than steady-state mass balance assumption are of great help to find the range of feasible fluxes across a metabolic network (i.e., metabolites uptake/secretion rates determined from medium measurements).

Depending on the optimized objective function we can distinguish two main model-based approaches to quantify metabolic fluxes: Flux Balance Analysis (FBA) and 13C Metabolic Flux Analysis (13C MFA).[76,78,79] Both use stoichiometric, thermodynamic and experimental constraints to find the range of feasible fluxes across a metabolic network and then find the flux distributions within that space that optimize a given objective function. In the case of 13C-MFA this objective function is set to minimize the difference between simulated and experimentally measured 13C enrichment in metabolites whereas the objective function definition for FBA is not that clear, particularly in FBA-biomedical applications.[81] Nowadays, 13C metabolic flux analysis (13C-MFA) is the predominant technique used for quantitatively estimate intracellular fluxes.[82]

References

(1) Lehninger, A. L., Nelson, D. L., & Cox, M. M. *Lehninger principles of biochemistry*; New York: Worth Publishers, 2000.

(2) Nielsen, J. *Annu. Rev. Biochem.* **2017**, *86* (1), 245–275.

(3) Jonsson, A. L.; Roberts, M. A. J.; Kiappes, J. L.; Scott, K. A. *Essays Biochem.* **2017**, *61* (4), 401–427.

(4) The UniProt Consortium. *Nucleic Acids Res.* **2017**, *45* (D1), D158–D169.

(5) DeBerardinis, R. J.; Thompson, C. B. *Cell* **2012**, *148* (6), 1132–1144.

(6) Manzoni, C.; Kia, D. A.; Vandrovcova, J.; Hardy, J.; Wood, N. W.; Lewis, P. A.; Ferrari, R. *Brief. Bioinform.* **2018**, *19* (2), 286–302.

(7) Horgan, R. P.; Kenny, L. C. *Obstet. Gynaecol.* **2011**, *13* (3), 189–195.

(8) Kind, T.; Tsugawa, H.; Cajka, T.; Ma, Y.; Lai, Z.; Mehta, S. S.; Wohlgemuth, G.; Barupal, D. K.; Showalter, M. R.; Arita, M.; Fiehn, O. *Mass Spectrom. Rev.* **2018**, *37* (4), 513–532.

(9) Lei, Z.; Huhman, D. V; Sumner, L. W. *J. Biol. Chem.* **2011**, *286* (29), 25435–25442.

(10) Fischer, R.; Bowness, P.; Kessler, B. M. *Proteomics* **2013**, *13* (23-24), 3371–3386.

(11) Vidova, V.; Spacil, Z. *Anal. Chim. Acta* **2017**, *964*, 7–23.

(12) Junot, C.; Fenaille, F.; Colsch, B.; Bécher, F. *Mass Spectrom. Rev.* **2014**, *33* (6), 471–500.

(13) Mikami, T.; Aoki, M.; Kimura, T. *Curr. Mol. Pharmacol.* **2012**, *5* (2), 301–316.

(14) Glish, G. L.; Burinsky, D. J. *J. Am. Soc. Mass Spectrom.* **2008**, *19* (2), 161–172.

(15) Zhou, J.; Yin, Y. *Analyst* **2016**, *141* (23), 6362–6373.

(16) Fenaille, F.; Barbier Saint-Hilaire, P.; Rousseau, K.; Junot, C. *J. Chromatogr. A* **2017**, *1526*, 1–12.

(17) Ma, S.; Chowdhury, S. K. *Bioanalysis* **2013**, *5* (10), 1285–1297.

(18) Wang, R.; Yin, Y.; Zhu, Z.-J. *Anal. Bioanal. Chem.* **2019**.

(19) Kuehnbaum, N. L.; Britz-McKibbin, P. *Chem. Rev.* **2013**, *113* (4), 2437–2468.

(20) Nordström, A.; O'Maille, G.; Qin, C.; Siuzdak, G. *Anal. Chem.* **2006**, *78* (10), 3289–3295.

(21) Contrepois, K.; Jiang, L.; Snyder, M. *Mol. Cell. Proteomics* **2015**, *14* (6), 1684–1695.

(22) Ali, I.; Aboul-Enein, H. Y.; Singh, P.; Singh, R.; Sharma, B. *Saudi Pharm. J. SPJ Off. Publ. Saudi Pharm. Soc.* **2010**, *18* (2), 59–73.

(23) Boersema, P. J.; Mohammed, S.; Heck, A. J. R. *Anal. Bioanal. Chem.* **2008**, *391* (1), 151–159.

(24) Rahman, M. M.; Abd El-Aty, A. M.; Choi, J.-H.; Shin, H.-C.; Shin, S. C.; Shim, J.-H. In *Analytical Separation Science*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2015; pp 823–834.

(25) Dunn, W. B.; Broadhurst, D.; Begley, P.; Zelena, E.; Francis-McIntyre, S.; Anderson, N.; Brown, M.; Knowles, J. D.; Halsall, A.; Haselden, J. N.; Nicholls, A. W.; Wilson, I. D.; Kell, D. B.; Goodacre, R. *Nat. Protoc.* **2011**, *6* (7), 1060–1083.

(26) Ramautar, R.; Somsen, G. W.; de Jong, G. J. *Electrophoresis* **2017**, *38* (1), 190–202.

(27) Rabilloud, T.; Chevallet, M.; Luche, S.; Lelong, C. *J. Proteomics* **2010**, *73* (11), 2064–2077.

(28) Oliveira, B. M.; Coorssen, J. R.; Martins-de-Souza, D. *J. Proteomics* **2014**, *104*, 140–150.

(29) Aebersold, R.; Mann, M. *Nature* **2003**, *422* (6928), 198–207.

(30) Altelaar, A. F. M.; Munoz, J.; Heck, A. J. R. *Nat. Rev. Genet.* **2013**, *14* (1), 35–48.

(31) Yates, J. R.; Ruse, C. I.; Nakorchevsky, A. *Annu. Rev. Biomed. Eng.* **2009**, *11* (1), 49–79.

(32) Carnielli, C. M.; Winck, F. V.; Paes Leme, A. F. *Biochim. Biophys. Acta - Proteins Proteomics* **2015**, *1854* (1), 46–54.

(33) Patti, G. J.; Yanes, O.; Siuzdak, G. *Nat. Rev. Mol. Cell Biol.* **2012**, *13* (4), 263–269.

(34) Lu, W.; Su, X.; Klein, M. S.; Lewis, I. A.; Fiehn, O.; Rabinowitz, J. D. *Annu. Rev. Biochem.* **2017**, *86*, 277.

(35) González-Domínguez, R.; Sayago, A.; Fernández-Recamales, Á. *Bioanalysis* **2017**, *9* (1), 131–148.

(36) Lee, D. Y.; Bowen, B. P.; Northen, T. R. *Biotechniques* **2010**, *49* (2), 557–565.

(37) Boughton, B. A.; Hamilton, B. In *Advances in experimental medicine and biology*; 2017; Vol. 965, pp 291–321.

(38) Cui, L.; Lu, H.; Lee, Y. H. *Mass Spectrom. Rev.* **2018**, *37* (6), 772–792.

(39) Lai, Z.; Fiehn, O. *Mass Spectrom. Rev.* **2018**, *37* (3), 245–257.

(40) Mastrangelo, A.; Ferrarini, A.; Rey-Stolle, F.; García, A.; Barbas, C. *Anal. Chim. Acta* **2015**, *900*, 21–35.

(41) Papadimitropoulos, M.-E. P.; Vasilopoulou, C. G.; Maga-Nteve, C.; Klapa, M. I. In *Methods in molecular biology (Clifton, N.J.)*; 2018; Vol. 1738, pp 133–147.

(42) Lu, W.; Bennett, B. D.; Rabinowitz, J. D. *J. Chromatogr. B. Analyt. Technol. Biomed. Life Sci.* **2008**, *871* (2), 236–242.

(43) Katajamaa, M.; Orešič, M. *J. Chromatogr. A* **2007**, *1158* (1-2), 318–328.

(44) Want, E. *Bioanalysis* **2009**, *1* (4), 805–819.

(45) Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Römpp, A.; Neumann, S.; Pizarro, A. D.; Montecchi-Palazzi, L.; Tasman, N.; Coleman, M.; Reisinger, F.;

Souda, P.; Hermjakob, H.; Binz, P.-A.; Deutsch, E. W. *Mol. Cell. Proteomics* **2011**, *10* (1), R110.000133.

(46) Tautenhahn, R.; Böttcher, C.; Neumann, S. *BMC Bioinformatics* **2008**, *9* (1), 504.

(47) Alonso, A.; Marsal, S.; Julià, A. *Front. Bioeng. Biotechnol.* **2015**, *3*, 23.

(48) 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. *Anal. Chem.* **2016**, *88* (19), 9821–9829.

(49) Want, E. J.; Coen, M.; Masson, P.; Keun, H. C.; Pearce, J. T. M.; Reily, M. D.; Robertson, D. G.; Rohde, C. M.; Holmes, E.; Lindon, J. C.; Plumb, R. S.; Nicholson, J. K. *Anal. Chem.* **2010**, *82* (12), 5282–5289.

(50) Sugimoto, M.; Kawakami, M.; Robert, M.; Soga, T.; Tomita, M. *Curr. Bioinform.* **2012**, *7* (1), 96–108.

(51) Smith, R.; Ventura, D.; Prince, J. T. *Brief. Bioinform.* **2015**, *16* (1), 104–117.

(52) Vinaixa, M.; Samino, S.; Saez, I.; Duran, J.; Guinovart, J. J.; Yanes, O.; Vinaixa, M.; Samino, S.; Saez, I.; Duran, J.; Guinovart, J. J.; Yanes, O. *Metabolites* **2012**, *2* (4), 775–795.

(53) Chokkathukalam, A.; Kim, D.-H.; Barrett, M. P.; Breitling, R.; Creek, D. J. *Bioanalysis* **2014**, *6* (4), 511–524.

(54) Mahieu, N. G.; Patti, G. J. *Anal. Chem.* **2017**, *89* (19), 10397–10406.

(55) Domingo-Almenara, X.; Montenegro-Burke, J. R.; Benton, H. P.; Siuzdak, G. *Anal. Chem.* **2018**, *90* (1), 480–489.

(56) Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T. R.; Neumann, S. *Anal. Chem.* **2012**, *84* (1), 283–289.

(57) Senan, O.; Aguilar-Mogas, A.; Navarro, M.; Capellades, J.; Noon, L.; Burks, D.; Yanes, O.; Guimerà, R.; Sales-Pardo, M. *Bioinformatics* **2019**, *35* (20), 4089–4097.

(58) Broeckling, C. D.; Afsar, F. A.; Neumann, S.; Ben-Hur, A.; Prenni, J. E. *Anal. Chem.* **2014**, *86* (14), 6812–6817.

(59) Uppal, K.; Walker, D. I.; Jones, D. P. *Anal. Chem.* **2017**, *89* (2), 1063–1067.

(60) Salek, R. M.; Steinbeck, C.; Viant, M. R.; Goodacre, R.; Dunn, W. B. *Gigascience* **2013**, *2*, 13.

(61) Blaženović, I.; Kind, T.; Ji, J.; Fiehn, O. *Metabolites* **2018**, *8* (2).

(62) Vinaixa, M.; Schymanski, E. L.; Neumann, S.; Navarro, M.; Salek, R. M.; Yanes, O. *TrAC Trends Anal. Chem.* **2016**, *78*, 23–35.

(63) Blaženović, I.; Kind, T.; Torbašinović, H.; Obrenović, S.; Mehta, S. S.; Tsugawa, H.; Wermuth, T.; Schauer, N.; Jahn, M.; Biedendieck, R.; Jahn, D.; Fiehn, O. *J. Cheminform.* **2017**, *9* (1), 32.

(64) Mahadevan, R.; Palsson, B. O. *Biophys. J.* **2005**, *88* (1), L07–L09.

(65) Pavlopoulos, G. A.; Secrier, M.; Moschopoulos, C. N.; Soldatos, T. G.; Kossida, S.; Aerts, J.; Schneider, R.; Bagos, P. G. *BioData Min.* **2011**, *4* (1), 10.

(66) Yilmaz, L. S.; Walhout, A. J. *Curr. Opin. Chem. Biol.* **2017**, *36*, 32–39.

(67) Angione, C. *Biomed Res. Int.* **2019**, *2019*, 8304260.

(68) Aurich, M. K.; Thiele, I. Humana Press, New York, NY, 2016; pp 253–281.

(69) Kanehisa, M.; Goto, S. *Nucleic Acids Res.* **2000**, *28* (1), 27–30.

(70) Karp, P. D.; Billington, R.; Caspi, R.; Fulcher, C. A.; Latendresse, M.; Kothari, A.; Keseler, I. M.; Krummenacker, M.; Midford, P. E.; Ong, Q.; Ong, W. K.; Paley, S. M.; Subhraveti, P. *Brief. Bioinform.* **2019**, *20* (4), 1085–1093.

(71) Schomburg, I.; Jeske, L.; Ulbrich, M.; Placzek, S.; Chang, A.; Schomburg, D. *J. Biotechnol.* **2017**, *261*, 194–206.

(72) 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. *Metabolomics* **2016**, *12* (7), 109.

(73) Srivastava, A.; Kowalski, G. M.; Callahan, D. L.; Meikle, P. J.; Creek, D. J. *Metabolites* **2016**, *6* (4).

(74) Jang, C.; Chen, L.; Rabinowitz, J. D. *Cell* **2018**, *173* (4), 822–837.

(75) Metallo, C. M.; Vander Heiden, M. G. *Mol. Cell* **2013**, *49* (3), 388–398.

(76) Cascante, M.; Marin, S. *Essays Biochem.* **2008**, *45*, 67–81.

(77) Sauer, U. *Mol. Syst. Biol.* **2006**, *2*, 62.

(78) Sauer, U. *Mol. Syst. Biol.* **2006**, *2* (1), 62.

(79) Zamboni, N.; Fendt, S.-M.; Rühl, M.; Sauer, U. *Nat. Protoc.* **2009**, *4* (6), 878–892.

(80) Sauer, U.; Zamboni, N. *Nat. Biotechnol.* **2008**, *26* (10), 1090–1092.

(81) Foguet, C.; Jayaraman, A.; Marin, S.; Selivanov, V. A.; Moreno, P.; Messeguer, R.; de Atauri, P.; Cascante, M. *PLOS Comput. Biol.* **2019**, *15* (9), e1007310.

(82) Long, C. P.; Antoniewicz, M. R. *Nat. Protoc.* **2019**, *14* (10), 2856–2877.

OBJECTIVES

OBJECTIVES

This thesis is mainly aimed at developing experimental and computational methods to predict and detect a profile of metabolites with imbalanced fluxes/abundances in a cellular phenotype. To accomplish this, three main sub-objectives have been established:

- To develop a computational framework to infer metabolic imbalanced fluxes/abundances from quantitative proteomics data.
- To establish a targeted methodology for high-throughput metabolite flux determination of central carbon metabolism pathways by using stable isotope labelling experiments.
- To develop a method aimed to assess metabolic fluxes beyond central carbon metabolism pathways by unbiased detection of fractional isotopologue enrichment in stable isotope labelling experiments.

This thesis is conformed by three chapters each one covering one of the individual sub-objectives described above. In Chapter 1 a novel data analysis approach is developed to predict a set of metabolites with imbalanced fluxes/abundances from quantitative proteomics experiments. This approach is based upon statistical inference on quantitative proteomic data considering the connectivity of the metabolic network resulting from tailoring a genome-scale human reconstruction model using detected enzymatic proteins. To further confirm predicted imbalanced fluxes, we have developed a series of methodologies to experimentally measure them in a high-throughput manner. Thus, in Chapter 2, a singular and unique method directed towards the qualitative assessment of flux profile in central carbon metabolism pathways is developed. This is done by measuring how the label of an enriched substrate propagates through the metabolite intermediates in such pathways. Finally, a method expanding the coverage of the empirical assays of metabolic flux profiles beyond central carbon metabolism is presented in Chapter 3.

RESULTS

CHAPTER 1. Inferring metabolite alterations within metabolic networks using proteomics

1.1 Introduction

Proteins are the key components for the three main categories of biological networks, the most common being protein–protein interaction networks, signaling networks, and metabolic networks. Proteomics endeavors to study protein levels, post-translational modifications and protein-protein interactions, which is subsequently used to understand cellular processes, including metabolic regulation. Proteins have different roles in metabolism; proteins with enzymatic activity can control metabolite concentrations and metabolic fluxes, and non-enzymatic proteins can mediate the uptake and transport of metabolites, or activate/deactivate signaling cascades in metabolic regulation. Defects on protein expression levels have been thoroughly reported as the cause of multiple diseases and have become the spotlight for many researchers in the fields of molecular and systems biology.[1–4]

Typically, transcript abundances have been used as a surrogate for protein measurements to study metabolism, however with the improvement of proteomic technologies and analytical approaches over the last decade, the sensitivity and the range of proteins quantified have increased dramatically which has also expanded the landscape of enzymatic proteins in proteomic datasets.[5] Similarly, metabolomics attempts to accurately measure the abundance of metabolites as a readout of the metabolic activity and physiological state of a cell.[6] Together, these omic sciences can provide profound knowledge on the metabolic regulation and metabolic rewiring of cells under different environmental stimuli.

Genome-scale metabolic models computationally describe gene-protein-reaction associations for entire metabolic genes in an organism, and can be used to predict metabolic fluxes for various systems-level metabolic studies. A well-known example of a metabolic model is Recon 2[7] an expansion of Recon 1, the earliest comprehensive human genome-scale 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 metabolic reactions, annotated with gene associations and organelle localization. It is possible to extend it further with information on gene-protein linkage from other databases such as KEGG[8] or REACTOME[9]. Genome-scale metabolic models can be used to build metabolic networks, or graphs, which are node-link representations using the connectivity between metabolites based on the reactions in which they take part. The topology of biochemical networks, including protein–protein interaction networks, signaling networks, and metabolic networks, with transcriptomic and proteomic data can be used to predict the impact of perturbation patterns in the absence of the kinetic parameters.[10] These studies have revealed the interconnectivity of genes and proteins, and the presence of functional modules where gene and protein expression are coordinated.[11].

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 (see Figure 6).

Herein, a data analysis workflow that employs the connectivity of a metabolic network, based on the Recon 2 structure, is presented to unbiasedly test metabolites within the network by evaluating the level of proteins associated with their transformation as a whole, instead of separately.

To develop and validate this novel analysis we obtained both proteomics and metabolomics data from 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).

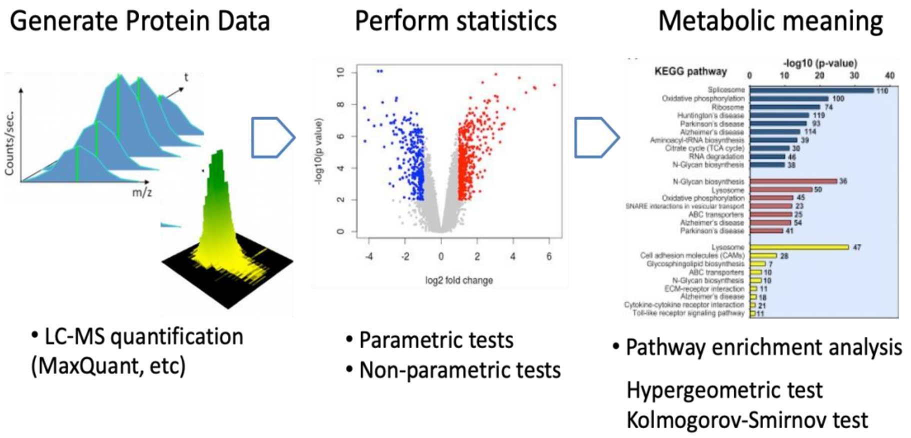


Figure *6.* Conventional proteome analysis. Differentially expressed proteins are identified by using a statistical test (e.g., t test analysis with Bonferroni correction), followed by annotation (e.g., pathway) enrichment analysis.

Diabetic retinopathy is a common long-term diabetes complication occurring to type II diabetes patients that leads to vision impairment or blindness. 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[290]. 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.[12] 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.

1.2 Methods

In order to build the reaction-based metabolic network and annotate each reaction with Uniprot accession identifiers it is important to consider the internal organization of the used database. In here we parsed and built an annotated metabolic network for two different well-known databases: KEGG and Recon 2. The main difference between KEGG and Recon databases is that, in the latter some genes are annotated as gene clusters in the form of gene-protein-reaction rules (GPRs).

Our aim is to generate a comprehensive metabolic network but with low redundancy, so it only does contain indispensable edges, and relate them to a series of genes and protein identifiers. In Figure 7 we show how this is achieved in RECON and KEGG for a well-characterized enzymatic reaction (Lactate dehydrogenase, EC 1.1.1.27), displaying the singularities of each database:

- In the case of Recon SBML files, the XML structure contains two main lists of elements: a list of reactions that in turn contains information about associated modifiers - genes -, reactants and products - metabolites- ; and a list of species, in other words, entities that participate in reactions, metabolite lists that collect metadata such as formula or name, and modifiers, which are combinations of one or more genes that code for proteins that perform such reaction (GPRs), Uniprot identifiers are included in the metadata of each modifier. Using this information, we built a metabolic network characterized by 2642 nodes (metabolites) and 11894 edges (4302 reactions), with a total of 2598 Uniprot identifiers associated to them. However, we decided to ignore some ubiquitous and non-informative metabolites (a total of 35, such as H2O, H+, …) and transport reactions (reactions that do not alter metabolite abundance), reducing the complexity of the network.

- In the case of KEGG, we parsed different files from the FTP downloadable database. The reaction file contains information about reactions, importantly reactant pairs and enzyme code annotation of each reaction. Reactant pairs relate metabolites that are transformed into one another in a reaction, in addition, they can be distinguished with different subtypes (main, cofac, ligase and leave). Using other files found in the database (hsa_enzyme.list and hsa_Uniprot.list) it is possible to relate each enzyme code to a list of individual gene(s) that in turn can be linked to their coded Uniprot identifiers. To build the KEGG metabolic network, we used only main reaction pairs, as other reactions normally involve less relevant metabolic transformations. This resulted in 7181 nodes (metabolites) and 12641 edges (9028 reactions), with a total of 3596 Uniprot identifiers associated to them.

Once the scaffold metabolic networks are built, they must be adapted to proteomics data. Since protein expression depends on multiple factors and proteomics is generally not sensitive enough for a full coverage of our database, metabolic networks must be subset based on detected proteins. In here, the Recon metabolic network is filtered based on its GPRs, meaning that a reaction is present if at least a full modifier (gene-protein set) is annotated, thus a protein is detected per each gene according to the logic of the GPRs, in Figure 7 modifiers translation into gene and protein entities is shown in the dashed square. While, in the case of KEGG, which lacks GPRs, a reaction is used if at least a protein of one of the related genes is found in the proteomics dataset. Therefore, for the exact same proteomics dataset, KEGG is prone to have a higher number of metabolic reactions in the network than that of Recon. In both databases, we tested building of networks enforcing the detection of a protein per associated gene, such approach led to poor results leading to a high proportion of disconnected nodes.

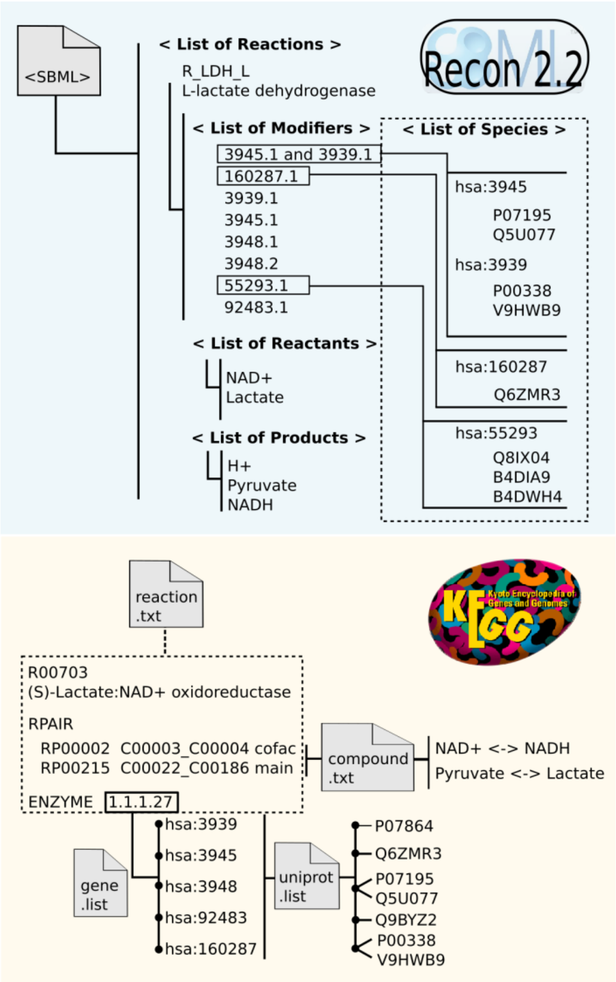


Figure 7. Extraction of reaction, reactants and their associated genes and proteins for Recon 2 SBML and KEGG databases.

Even though, apparently the results seem to indicate that KEGG generates a larger network, we found that a large proportion of nodes are actually disconnected, which also occurs in the case of Recon but is less dramatic.

The overlap in terms of protein identifiers included 1894 Uniprot identifiers, about 55-60%.

In addition, we saw that in the case of Recon, we retained a larger proportion of its reaction collection (Figure 8). Namely, out of the total of 7440 reactions found in Recon 2, 3286 can be linked to at least one protein and gene identifier, while for KEGG the proportion was limited to 2181 out of 7876 (*2181+5695*) reactions. Even if KEGG contains a larger collection of enzymatic reactions (reactions annotated with an EC code): 8520 (*5695+644+2181*) to 4302 (*3286+1016*).

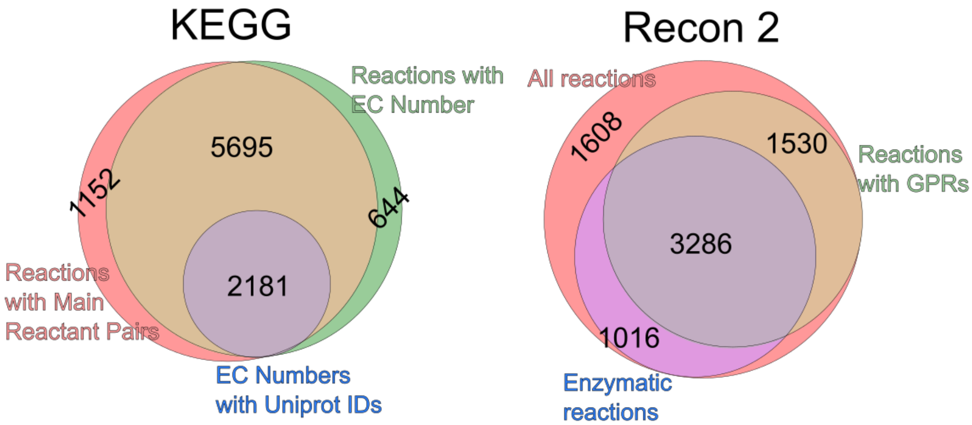


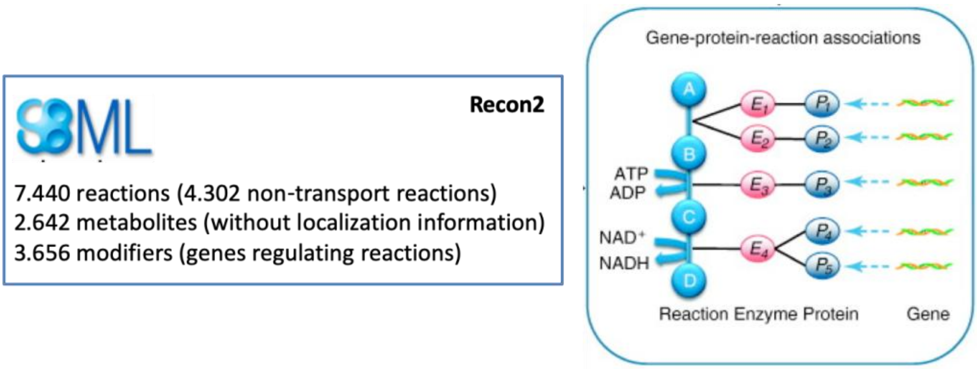
Figure 8. Venn diagrams depicting the result of reaction filtering for metabolic network building in KEGG and Recon 2. Note: In the case of Recon, reaction localization was ignored.

1.3 Results

Regulatory mechanisms in response to mutations, environmental insults or perturbations may cause changes in the cell's metabolic profile, leading to the accumulation of some metabolites and an increased (or decreased) flux through certain metabolic pathways. When this takes place in the context of a pathological state, these changes are referred to 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.

Recon 2 builds into a manually curated metabolic network, which can be enriched with gene-protein annotations, and protein identifiers linked to the corresponding Recon reactions (Figure 9). Extending Recon 2 annotation with associated proteins yields a new layer of useful data for metabolism interrogation and network perturbations. Here we used a database (SBML format) of 7.440 reactions, 2.642 metabolites and 3.656 associated genes.

Using the metabolic network linkage (Figure 10 A), we explored the number of associated proteins per metabolite, namely proteins that participate in the reactions that consume or produce a given metabolite (Figure 10 B), revealing that about 40% of metabolites are potentially regulated by more than 10 enzymes.



SBML
Recon2

7.440 reactions (4.302 non-transport reactions)
2.642 metabolites (without localization information)
3.656 modifiers (genes regulating reactions)

Figure 9. SBML files from Recon 2. Containing human gene-protein-reaction associations: 7.440 reactions including 3.656 genes, 3.090 protein (Uniprot) identifiers and 2.642 associated metabolites. Finally, there are 2.598 proteins associated to metabolic reactions that regulate 1.705 metabolites.

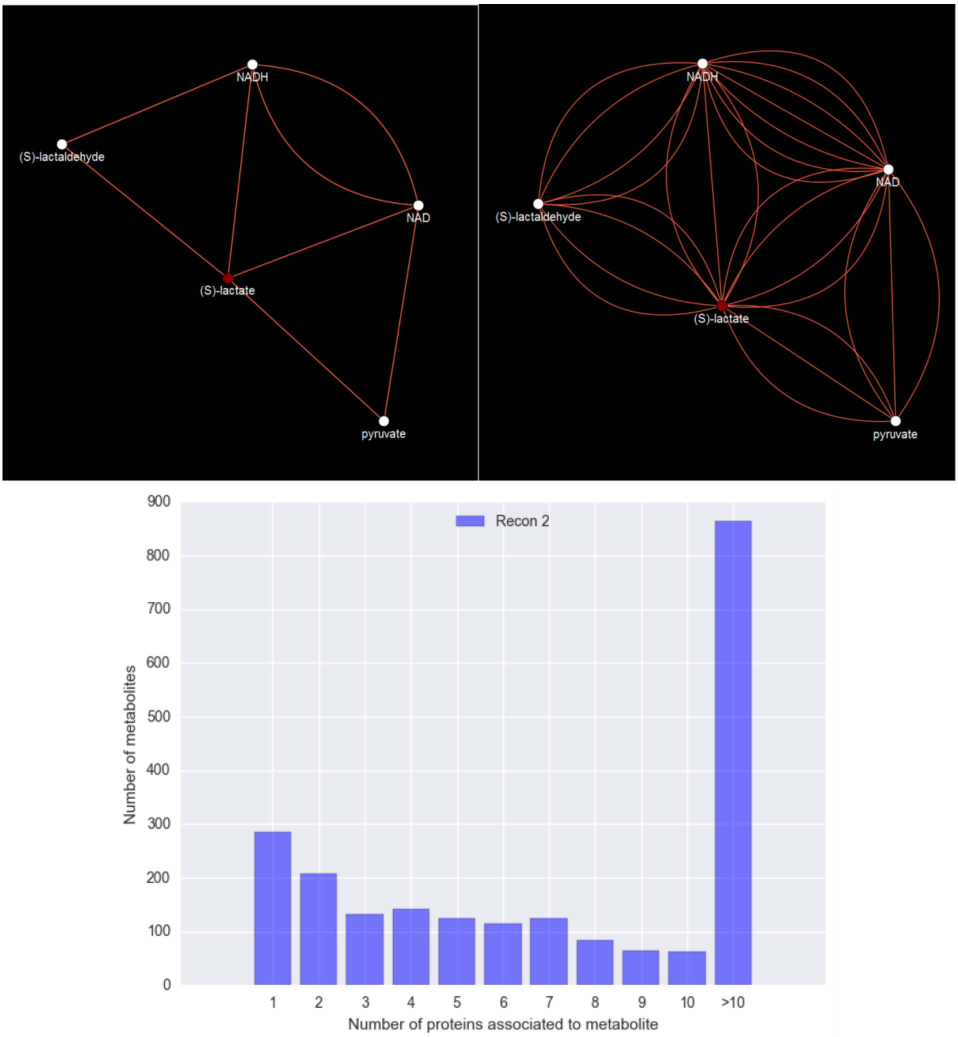
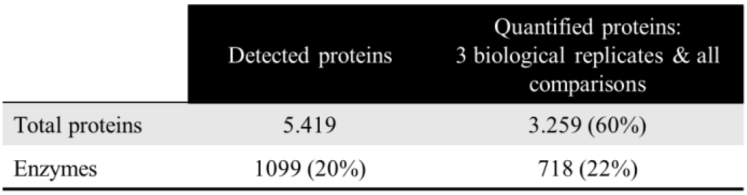


Figure 10. Recon metabolic networks. (A, top) In a reaction-based network, substrates and products are connected by one edge indicating the metabolic reaction between the metabolites (nodes). In an enzyme-based network, edges are all possible proteins with enzymatic activity catalyzing the interconversion of depicted metabolites (nodes). (B, bottom) Bar plot representing the number of proteins (enzymes) involved in metabolic reaction for every metabolite.

Next, we investigated the coverage of our TMT-based quantitative proteomics analysis of ARPE-19 cells exposed to 5mM glucose and normoxia (N5), 25mM glucose and normoxia (N25), 5mM glucose and hypoxia (H5), or 25mM glucose and hypoxia (H25) in the Recon 2 Out of 5,419 identified proteins, 1,099 (20%) were proteins with enzymatic activity. Of these, 718 were quantified in all four experimental conditions and the three biological replicates. Subsequent analyses were performed on this subset (Figure 11).



	Detected proteins	Quantified proteins: 3 biological replicates & all comparisons
Total proteins	5.419	3.259 (60%)
Enzymes	1099 (20%)	718 (22%)

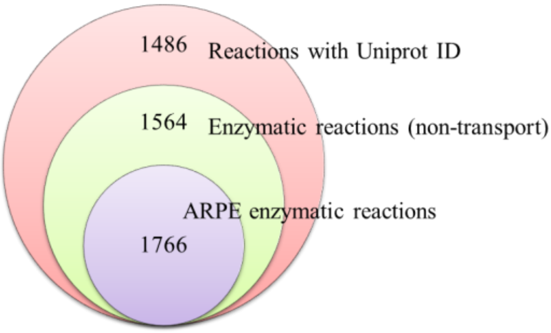


Figure 11. Summary of the quantitative proteomic analysis on RPE cells. The venn diagram shows the overlap between all reactions in Recon2 with Uniprot ID (4.816 in total), the enzymatic reactions (3.330 reactions) and finally, enzymatic reactions found in ARPE cells (1.766 reactions).

The 718 enzymes expressed in ARPE-19 cells were associated with metabolites based on the reactions they are involved in Recon 2. For most metabolites the number of proteins was large enough to generate a distribution of values (Figure 12A), so a binomial test could be applied to calculate if that distribution is equally centered around the non-change, meaning that the null hypothesis considers that the values are evenly distributed (Figure 12B). The binomial test was performed for each metabolite in the metabolic network, and those metabolites with a significant p-value, that is, whose distribution of protein abundance (fold-change) is different from control (N5), were considered altered metabolites.

Based on the 3,259 proteins, a total of 1,766 reactions were filtered out of the 4,320 Recon 2 non-transport reactions, resulting in mapping 1,390 metabolites out of the 2,000 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%) were lipids, and the rest were associated with amino acid and nucleotide metabolism, one carbon metabolism, glutathione metabolism or pentose phosphate pathway among others, as reported in Table 6 (pathway enrichment analysis using IMPALA).

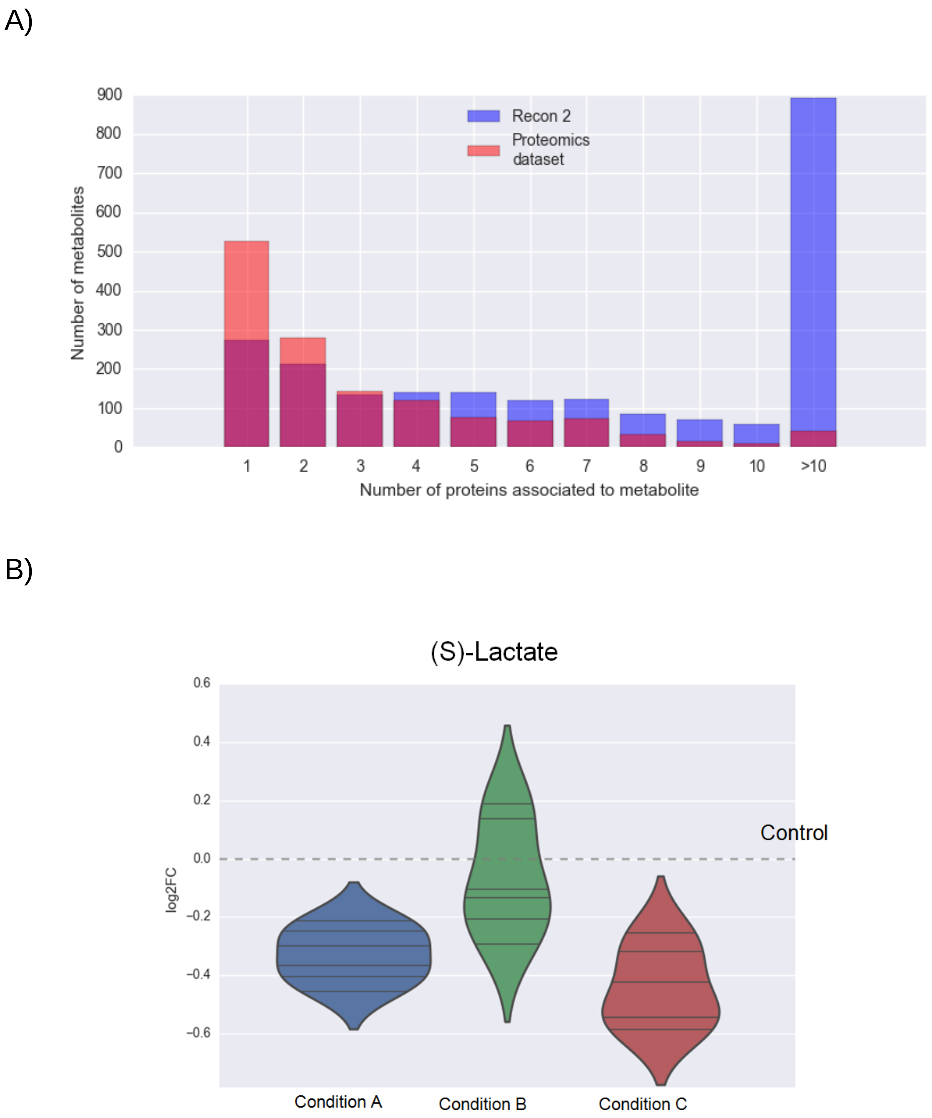


Figure 12. Interrogating the metabolic network with proteomics. (A) Bar plot representing the number of detected and quantified proteins (enzymes) in ARPE-19 involved in metabolic reaction for every metabolite reported in Recon 2. (B) Information from the enzyme-based network is used to statistically test the collective abundance of all the experimentally detected enzymes **associated** with every metabolite under different experimental conditions relative to a control.

Table 6. 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

Most metabolites were significantly altered due to an overall down-regulation of the protein expression profile associated to them. To understand and validate the high number of lipid-related candidates, we performed an untargeted lipidomic profiling of the ARPE-19 cells. We found an important number of significantly down-regulated features (Figure 13A) in hyperglycemic and hypoxic conditions (N25 and H25), which mainly corresponded to triacylglycerol (TG) species (Figure 13B).

With the aim of validating other altered metabolites and enriched pathways, we performed untargeted metabolomics of the polar cell extracts. Among the most important metabolites identified, we highlight glutathione (transsulfuration pathway and one carbon metabolism) that was down-regulated in N25 and H25 (Figure 14), and N-acetyl-neuraminic acid, N-acetyl-glucosamine, alpha-D-Fucose and GDP-Fucose (Asparagine N-linked glycosylation pathway, Figure 13). N-acetyl-neuraminic, N-acetyl-glucosamine and GDP-Fucose were down-regulated only in hypoxic with hyperglycemic condition (H25).

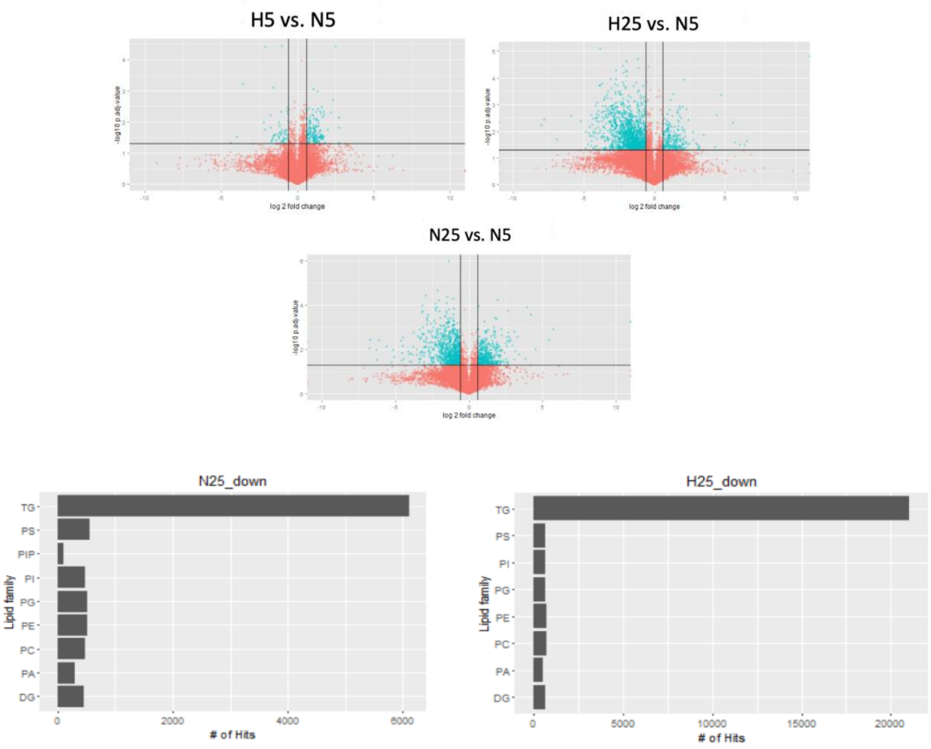


Figure 13. Lipidomics results. (A) Volcano plots of the lipidomic analysis. (B) Distribution of lipid families annotated in LipidMaps from the down-regulated features (blue) in N25 and H25. TG: triacylglycerol; PS: phosphatidylserine; PI: phosphatidylinositol; PG: phosphatidylglycerol; PE: phosphatidyletanolamine; PC: phosphatidylcholine; PA: phosphatidic acid; DG: diacylglycerol.

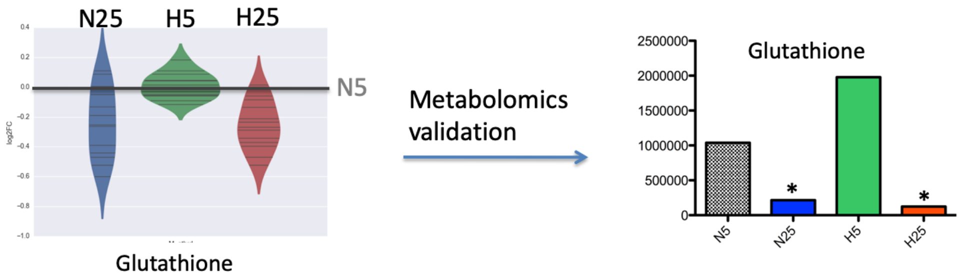


Figure 14. Example of dysregulated metabolite (glutathione). Predicted from the collective enzyme abundance distribution (left) and further experimental validation by LC-MS (right).

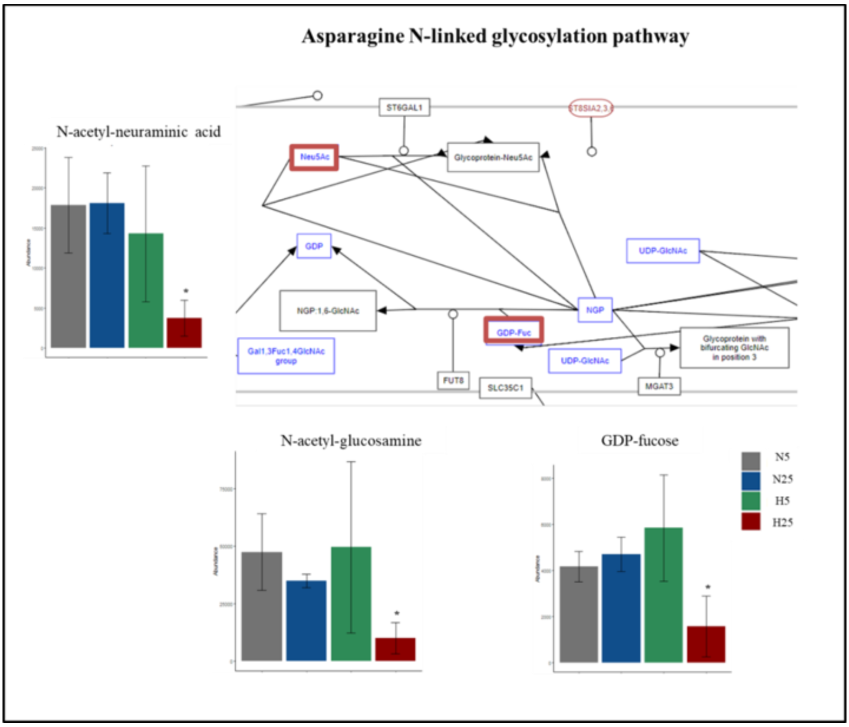


Figure 15. Significantly altered (*) metabolites in Asparagine N-linked glycosylation pathway. N-acetyl-neuraminic acid, N-acetyl-glucosamine and GDP-Fucose.

We further explored the clinical relevance of our results by analysing human vitreous humor samples from patients at different stages of DR, including 14 controls (healthy retinas), 4 non-proliferative DR (NPDR), an early stage of DR, and 12 proliferative DR (PDR), the latest stage of DR. Our NMR and MS-based metabolomic analysis confirmed some of the altered pathways and metabolites reported in RPE cells by the metabolic network analysis. The most relevant metabolites are involved in trans-sulfuration, one carbon metabolism and methylation pathways, including N5-formyl-THF, S-Adenosylhomocysteine (SAH), S-Adenosylmethionine (SAMe) and methionine, which appeared up-regulated in PDR (Figure 14).

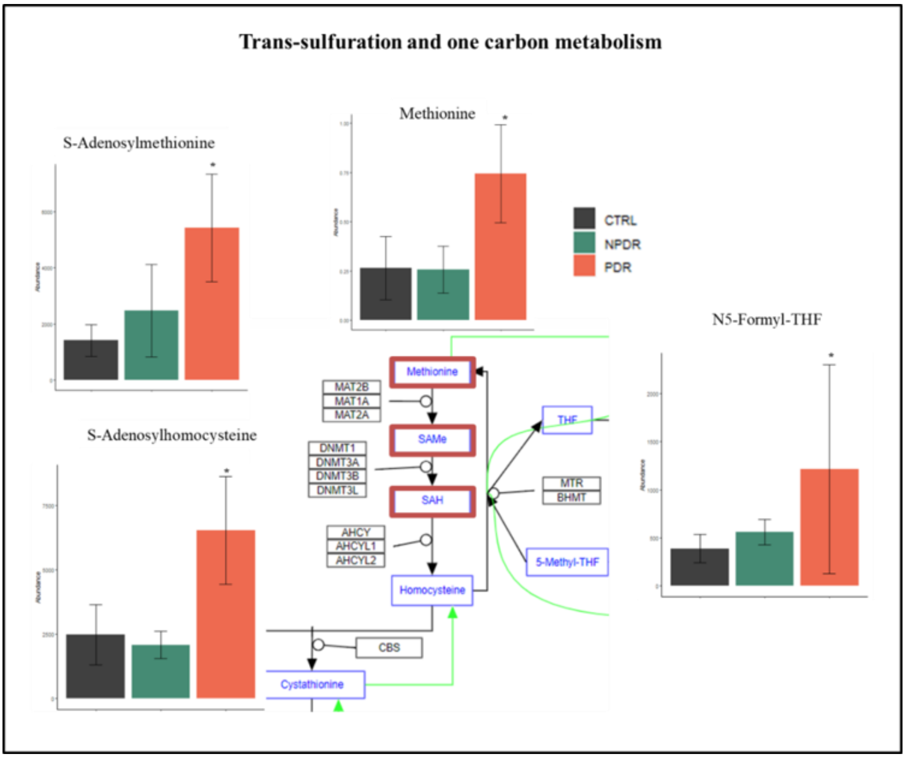


Figure 16. Trans-sulfuration and one carbon metabolism pathway. Showing exemples of experimentally detected altered metabolites found in vitreous humour.

Remarkably, the ketone bodies 3-hydroy-butyrate and acetoacetate were up-regulated in PDR (Figure 17), suggestion an alteration in the metabolism of fatty acids at the late stage of DR.

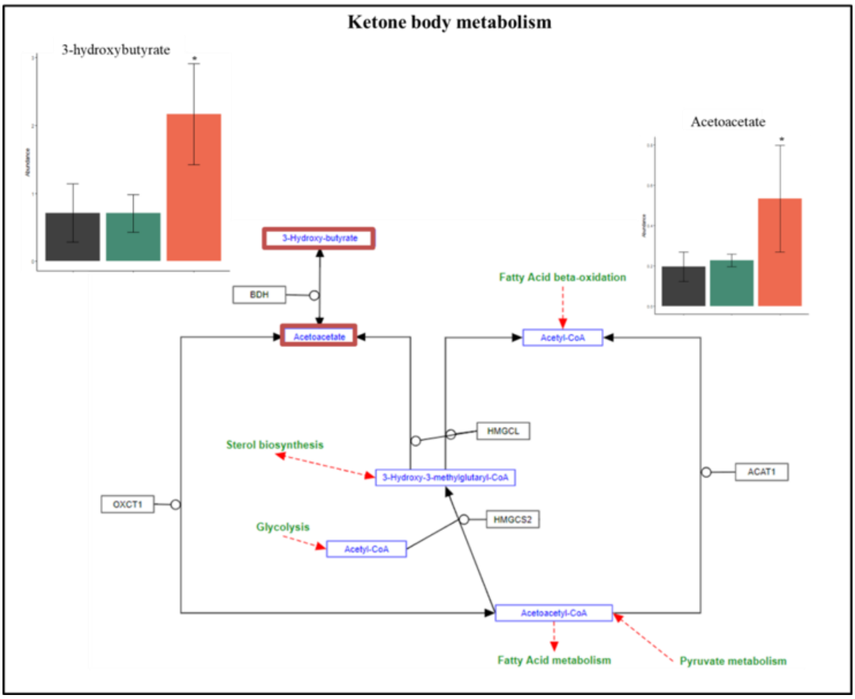


Figure 17. Ketone body metabolism. Showing experimentally detected 3-hydroxybutyrate and acetoacetate significantly up-regulated in the vitreous humor of PDR patients.

1.4 Discussion and conclusions

Cellular response to genetic and environmental perturbations is often reflected and/or mediated through changes in the metabolism. Such metabolic changes are often exerted through transcriptional changes induced by complex regulatory mechanisms coordinating the activity of different metabolic pathways. It is difficult to map such global protein expression responses by using traditional statistical methods, because many genes in the metabolic network have relatively small changes at their transcription/translation level. We therefore have developed an approach that integrates quantitative proteomics data with topological information from a human genome-scale metabolic model, showing that it is possible to reveal metabolites around which significant and coordinated protein expression changes occur in response to environmental perturbations, namely hypoxia and/or hyperglycemia.

The significant metabolites, defined as reporter metabolites elsewhere[13], mark spots in the metabolism where there is substantial regulation either to maintain homeostasis or adjust the concentration of the metabolite to another level required for proper functioning of the metabolic network under different environmental conditions. Conventional proteome analysis, in which differentially expressed proteins are identified by using a statistical test (e.g., t test analysis with Bonferroni correction), did not enable identification of the overall effect of hypoxia and/or hyperglycemia on the metabolism. Similar to protein-protein interaction networks, standard statistical analysis at the level of individual proteins, which assumes that metabolic enzymes are independent variables, show a strong bias toward highly up- and down-regulated proteins. In contrast, our metabolic network-based approach exploits the highly interconnected nature of the metabolic network enabling identification of subtle but coordinated and collective protein expression changes at the system level. This idea originates from the observation that expression difference increases with metabolic network distance, demonstrating that genes closer to each other in metabolic network tend to have, on average, higher level of coexpression.[14,15]

The integration of quantitative proteomics data from RPE cells under hyperglycemic and hypoxic conditions with metabolic networks resulted in the prediction of altered metabolites involved in one carbon metabolism[16] and associated pathways such as trans-sulfuration, methionine and folate cycle and methylation reactions. Remarkably, many of these predicted metabolites were further analyzed by LC-MS, GC/MS and/or NMR and their altered abundance in human RPE cells and vitreous humor of DR patients was also confirmed. Recently, Malaguarnera et al.[17] found significantly higher plasma levels of homocysteine in NPDR patients compared to a control group, and in PDR patients compared to a control group and NPDR patients. The severity of diabetic retinopathy was associated with lower folic acid and red cell folate levels, and a significant difference was observed between PDR and NPDR groups. SAM/SAH ratio in plasma was inversely associated with retinal sclerotic vessel abnormalities and retinopathy in subjects with T2D[18].

In addition, our results highlight the importance of lipid (TG and fatty acids) and ketone bodies metabolism in RPE cells. Typically, ketone bodies are produced from acetyl-CoA, mainly in the mitochondrial matrix of liver cells when carbohydrates are so scarce that energy must be obtained from breaking down of fatty acids. However, extrahepatic ketogenesis has been also demonstrated in tumor cells, astrocytes of the central nervous system, the kidney, pancreatic β cells, retinal pigment epithelium (RPE), and even in skeletal muscle[19]. Recent observations underscore the importance of ketone bodies as vital metabolic and signaling mediators when carbohydrates are abundant[20].

In summary, this methodology shows the significance of studying metabolic problems from a network point of view and will raise awareness on future proteomics studies.

References

(1) Altelaar, A. F. M.; Munoz, J.; Heck, A. J. R. *Nat. Rev. Genet.* **2013**, *14* (1), 35–48.

(2) Dias, M. H.; Kitano, E. S.; Zelanis, A.; Iwai, L. K. *Drug Discov. Today* **2016**, *21* (2), 264–277.

(3) Ebhardt, H. A.; Root, A.; Sander, C.; Aebersold, R. *Proteomics* **2015**, *15* (18), 3193–3208.

(4) Shi, T.; Song, E.; Nie, S.; Rodland, K. D.; Liu, T.; Qian, W.-J.; Smith, R. D. *Proteomics* **2016**, *16* (15-16), 2160–2182.

(5) Gygi, S. P.; Aebersold, R. *Curr. Opin. Chem. Biol.* **2000**, *4* (5), 489–494.

(6) Patti, G. J.; Yanes, O.; Siuzdak, G. *Nat. Rev. Mol. Cell Biol.* **2012**, *13* (4), 263–269.

(7) 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. *Metabolomics* **2016**, *12* (7), 109.

(8) Kanehisa, M.; Goto, S. *Nucleic Acids Res.* **2000**, *28* (1), 27–30.

(9) 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. *Nucleic Acids Res.* **2011**, *39* (Database issue), D691–D697.

(10) Santolini, M.; Barabási, A.-L. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (27), E6375–E6383.

(11) Kustatscher, G.; Grabowski, P.; Rappsilber, J. *Mol. Syst. Biol.* **2017**, *13* (8), 937.

(12) Simó, R.; Villarroel, M.; Corraliza, L.; Hernández, C.; Garcia-Ramírez, M. *J. Biomed. Biotechnol.* **2010**, *2010*, 190724.

(13) Patil, K. R.; Nielsen, J. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (8), 2685–2689.

(14) Kharchenko, P.; Church, G. M.; Vitkup, D. *Mol. Syst. Biol.* **2005**, *1* (1), 2005.0016.

(15) Spirin, V.; Gelfand, M. S.; Mironov, A. A.; Mirny, L. A. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (23), 8774–8779.

(16) Ducker, G. S.; Rabinowitz, J. D. *Cell Metab.* **2017**, *25* (1), 27–42.

(17) Malaguarnera, G.; Gagliano, C.; Salomone, S.; Giordano, M.; Bucolo, C.; Pappalardo, A.; Drago, F.; Caraci, F.; Avitabile, T.; Motta, M. *Clin. Ophthalmol.* **2015**, *9*, 1437–1442.

(18) Van Hecke, M. V; Dekker, J. M.; Nijpels, G.; Teerlink, T.; Jakobs, C.; Stolk, R. P.; Heine, R. J.; Bouter, L. M.; Polak, B. C. P.; Stehouwer, C. D. A. *Clin. Sci. (Lond).* **2008**, *114* (7), 479–487.

(19) Adijanto, J.; Du, J.; Moffat, C.; Seifert, E. L.; Hurle, J. B.; Philp, N. J. *J. Biol. Chem.* **2014**, *289* (30), 20570–20582.

(20) Puchalska, P.; Crawford, P. A. *Cell Metab.* **2017**, *25* (2), 262–284.

CHAPTER 2. Exploring the use of gas chromatography coupled to chemical ionization mass spectrometry (GC-CI-MS) for stable isotope labelling in metabolomics

2.1 Introduction

Metabolomics is nowadays a well-established branch of the omics field. The ability to detect and quantify hundreds of different small organic compounds in complex biological samples makes metabolomics a powerful source of biological information. Due to the chemical diversity of naturally occurring metabolites, metabolomics is rich in analytical instrumentation and configurations, with GC/MS and LC-MS being the most common techniques.

GC/MS has been the analytical platform of choice for measuring volatile compounds for decades[1]. However, with the introduction of chemical derivatization, and the inherent high-resolution chromatographic separations and low MS background noise, GC/MS-based metabolomics has evolved into a powerful analytical platform to produce compound abundance data in multiple types of targeted and untargeted metabolomic experiments[2–7]. GC systems are typically coupled to a single (GC-sQ) or triple quadrupole mass spectrometers (GC-QqQ) for targeted metabolomics, while untargeted analyses are generally performed using time-of-flight (GC-TOF) or Orbitrap (GC-Orbitrap) mass spectrometers. The most widely used ionization method in GC/MS is electron impact ionization (EI), a hard ionization strategy that generates highly reproducible and characteristic fragmentation spectra[1]. However, softer ionization techniques based on chemical ionization (CI) are alternative methods, where molecular ions of analyzed compounds are kept (mostly) intact[8].

The determination of relative or absolute metabolite concentrations by GC/MS can also be used to investigate the rate at which these compounds are being transformed into other intermediates by metabolic (i.e., enzymatic) activity, a field known as stable-isotope labellingand fluxomics[9]. Stable isotopes, such as ^{13}C and ^{15}N, are used to track the fate of a labelled nutrient, e.g. ^{13}C-glucose or $^{13}C/^{15}N$-glutamine, being a much safer alternative to the formerly used radioactive isotopes.[10–12] Metabolites enriched with stable isotopes maintain the original chemical structure and biochemical properties, but the different isotopic composition (i.e., isotopologues) cause a shift in their mass, which is detectable by MS. In this regard, GC-EI MS has become a platform of choice for stable isotope tracing studies and fluxomics approaches[13–19]. The coverage of GC-EI MS is comprehensive enough to cover most central carbon metabolism (i.e., glycolysis, TCA, amino acids, pentose phosphate pathway) thanks to the use of chemical derivatization. Although several tools for automated analysis of GC-EI MS data are available, including open source software[20–23], a few have been developed for stable-isotope labelling[24–30]. This is probably due to the difficulty in deconvoluting the intensity of overlapping isotopologues from EI mass spectra. Interestingly, the use of GC-CI MS for stable-isotope tracing is largely unexplored[31–35]. In here, we have studied the suitability of GC-CI-MS for stable-isotope tracing using multiple analytical configurations based on low-resolution and high-resolution mass spectrometry. We prove that isobutane is an ideal ionization gas by its ability to yield intact protonated molecular ions, and coupled to cutting-edge high-resolution GC/MS instruments, currently provide the best configuration for isotopologue quantification. In addition, we have developed an R-package called isoSCAN capable of processing GC-CI MS data from stable-isotope labelling studies.

2.2 Results

2.2.1 Comparing electron ionization (EI) with chemical ionization (CI) for stable isotope labelling

In order to optimize the detection of isotopologues using GC/MS, we first compared the performance of the most commonly used ionization sources in GC/MS: EI and CI.

In addition, we tested two different reagent gases for CI, namely methane and isobutane. Taking lactate as a reference metabolite in a complex biological sample, we found that CI with isobutane yielded most suitable spectra for stable isotope labelling experiments in comparison with EI (Figure 18). CI is a soft ionization method in which the abundance of a metabolite is concentrated in a few ions, facilitating the detection of isotopic enrichment, such as m+3 for lactate (m/z = 237.1) (Figure 18).

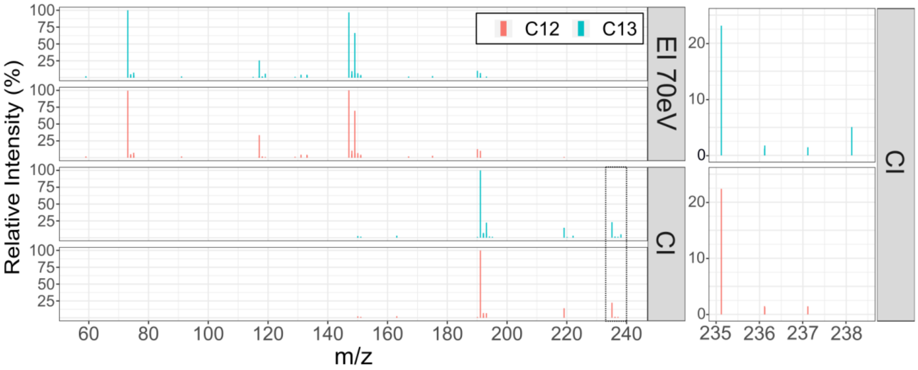


Figure 18. Comparison of Lactate 2TMS spectra. Obtained from non-labelled (red) and $^{13}C_6$-Glucose labelled (light blue) cell cultures. (Top) shows electron impact ionization fragment patterns. (Bottom) shows ions produced by chemical ionization, intact M+H ions are squared and zoomed in.

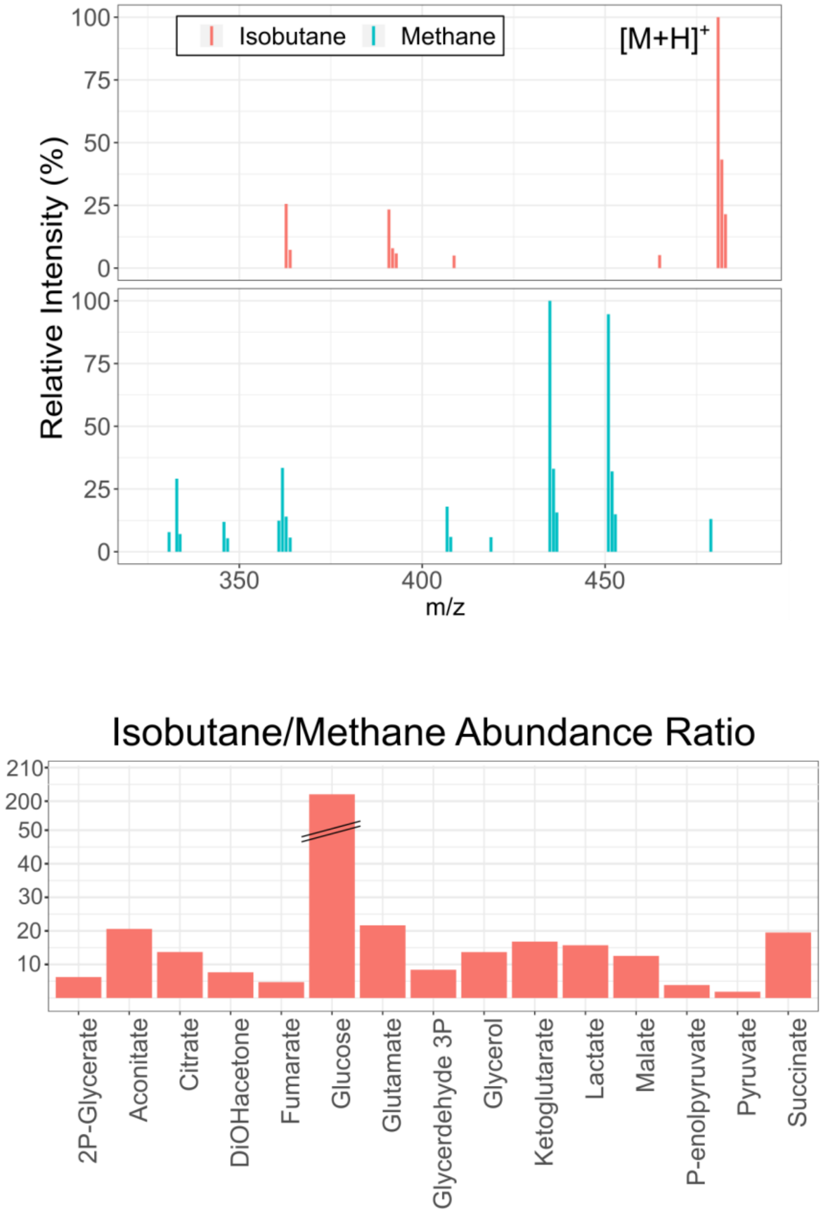


Figure 19. Chemical ionization and isobutane advantages. A,top. Chemical ionization spectra of Citrate 4TMS using Methane and Isobutane as reagent gases. B,bottom. Protonated (M+H+) ion abundance ratios (isobutane/methane) for glycolysis and TCA compounds.

In contrast, EI at 70 eV is a hard ionization method causing extensive fragmentation of compound structures in both non-labelled and labelled samples. This complicates the process of determining the amount of isotope labelling because the isotopic enrichment is scattered in multiple fragments. Notably, the use of isobutane as reagent gas produced mass spectra characterized by protonated molecular ions ($[M+H]^+$) with practically undetectable in-source fragmentation, unlike methane gas that induced unwanted fragmentation of the protonated molecular ion (Figure 19A). This observation applied also to all intermediates of the glycolytic pathway and TCA cycle (Figure 19B), where the $[M+H]^+$ ion was >5-fold more intense using isobutane than methane. Similarly, low-energy EI at 15 eV [29] also produced greater fragmentation of the molecular ion than CI-isobutane (Figure 20).

Finally, we show that CI-isobutane can ionize efficiently most relevant metabolites in central carbon metabolism, including intermediates of glycolysis, TCA cycle, pentose phosphate pathway, amino acids, urea cycle and polyamines (Table 2.2, see end of chapter 2). In summary, CI with isobutane resulted in a broad coverage of metabolites and greater signal intensity of the $[M+H]^+$ molecular ion for all standard metabolites tested.

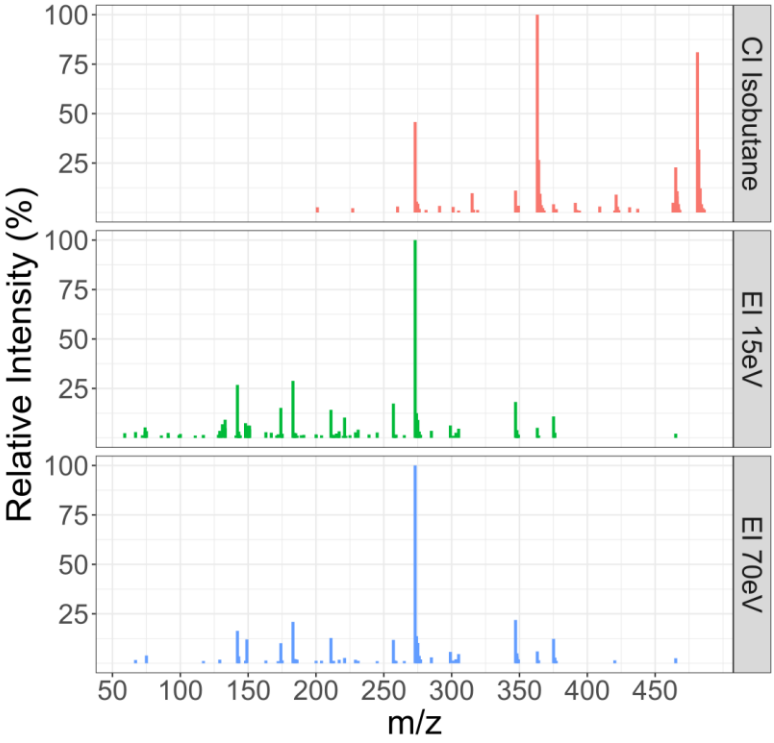


Figure 20. Spectra obtained for Citrate 4TMS in EI 15eV and 70eV ionization. Showing lower intensity for intact ion (*m/z 481.19*) than in the case of CI and isobutane as a reagent gas.

2.2.2 Optimizing data acquisition modes in GC-CI MS for stable isotope labelling

The collection of mass analyzers used in hyphenated mass-spectrometry is very broad nowadays. We compared, therefore, the sensitivity, selectivity and isotope ratio fidelity of different mass analyzers and data acquisition modes for stable isotope labelling using GC CI-isobutane MS. In particular, we used a low-mass resolution configuration based on GC-CI triple quadrupole MS, and two high-mass resolution configurations based on GC-CI qTOF MS and GC-CI qOrbitrap MS to analyze the same labelled and non-labelled samples. In the case of the quadrupole configuration, we explored two acquisition modes to find the most sensitive and selective for our purposes: selected ion monitoring (SIM) and narrow scan (NS) acquisition (Figure 21).

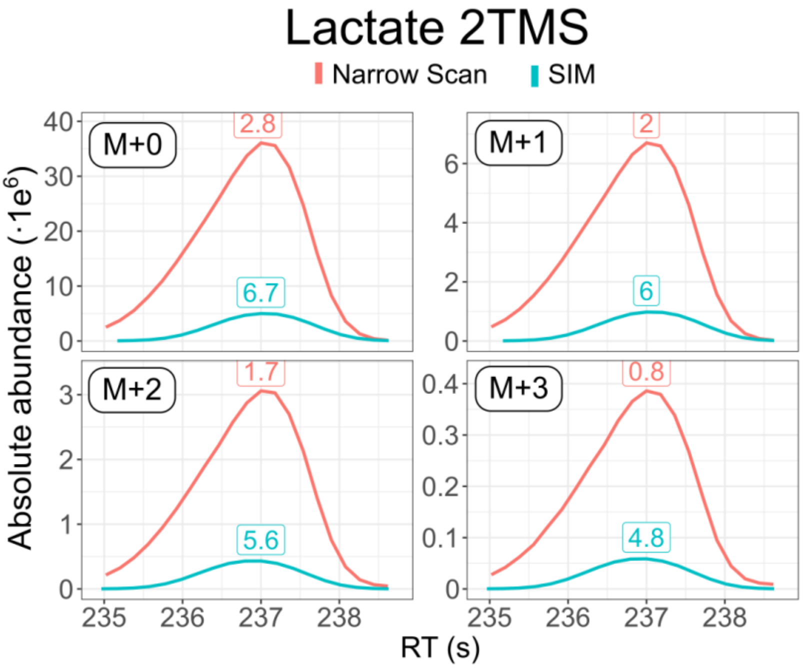


Figure 21. EICs obtained for corresponding Lactate 2TMS isotopologues in a triple quadrupole. Comparing NS (red) and SIM (blue) acquisition SNR and recorded units. Squared values are rounded Log10(SNR) values calculated per each EIC, which are several times larger for SIM compared to NS.

SIM acquisition method was set to monitor single ions, that is, every isotopologue based on the number of carbons in each metabolite. In contrast, NS acquisition method was set to register a cluster of ions within a mass range that covers the full pattern of isotopologues of the derivatized metabolite (typically 8-10 Da), e.g. in the case of lactic acid (2TMS, m/z

235.1107) the quadrupole is set to scan a mass range from m/z 232.0 to 240.0. Despite the greater signal intensity of NS, SIM yielded better signal-to-noise ratio values, likely due to the greater ion selectivity of SIM (Figure 21). The isotope ratio fidelity of NS and SIM, calculated as the relative deviation to the theoretical isotopic ratio, was similar for both acquisition methods which had a median value of ∼10%, indeed fidelity relies on ion quantification, therefore low signal-to-noise ratio values foster fidelity values (Figure 22).

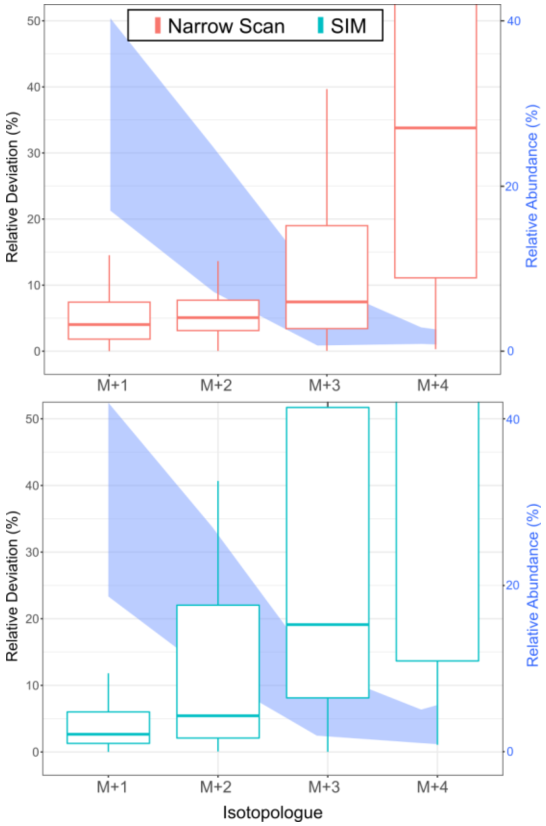


Figure 22. Isotopologue abundances relative deviation in relation to theoretical natural abundances detected in a triple quadrupole. Blue shade indicates the range of relative theoretical abundance per ion.

Next, we compared the sensitivity, selectivity and isotope ratio fidelity of two high-mass resolution instruments: a GC-Orbitrap mass analyzer acquiring in full scan mode at ~60.000 resolution and a GC-qTOF mass analyzer acquiring in full scan mode at ~40.000 resolution. The deviation of isotope ratio of GC-CI qOrbitrap MS and GC-CI qTOF MS were <15% from the theoretical values, yet only the qOrbitrap was capable of clearly resolving natural isotopes from silicon (29Si and 30Si) from carbon, while in the case of the qTOF they were practically indistinguishable and thus were summed (Figure 23). As expected, the qOrbitrap showed a much higher mass accuracy (<1ppm) compared to the QTOF (<6ppm).

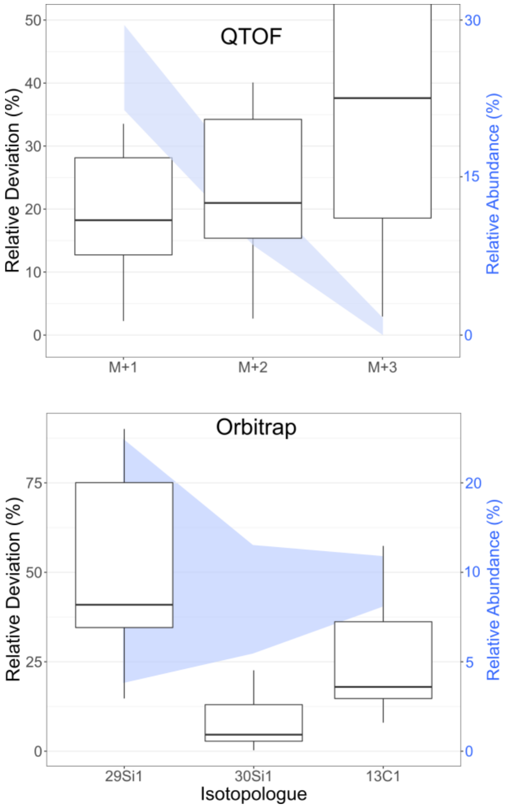


Figure 23. Isotopologue abundances relative deviation in relation to theoretical natural abundances detected in QTOF and qOrbitrap instruments. Blue shade indicates the range of relative theoretical abundance per ion.

Remarkably, the qOrbitrap in full scan mode yielded greater sensitivity than SIM and NS acquisition, particularly when detecting labelled isotopologues of low abundant compounds detected in biological samples (Figure 24). Interestingly, when the quadrupole was set to filter the same mass range as in NS (i.e., 8-10 Da) with the aim of reducing the ion population in the orbitrap mass analyzer, we did not observe an increase in sensitivity in comparison with full scan acquisition (data not shown). As expected, both high-mass resolution instruments have the additional advantage that are more selective than SIM and NS at low-mass resolution, because the former can distinguish artifacts or compounds that could interfere with the isotopologues of interest.

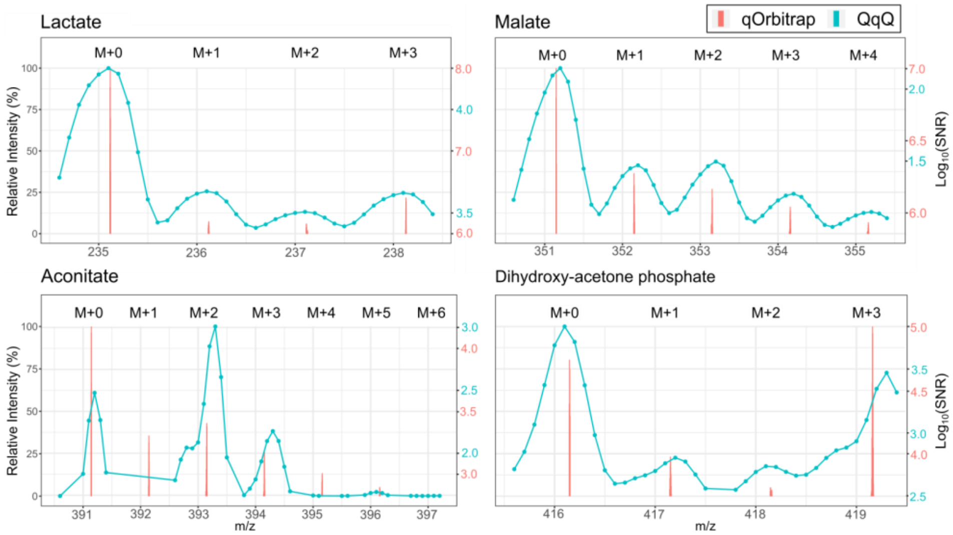


Figure 24. Comparison of relative abundances of isotopologue ions. Different compounds in triple quadrupole (blue) and qOrbitrap instruments (red). The spectra belong to a 13C labelled sample.

In summary, our results demonstrate that the latest generation of high-mass resolution GC/MS instruments in full scan mode have similar or superior sensitivity than quadrupole-based mass analyzers, making GC-CI (with isobutane) Orbitrap MS the most suitable configuration for the detection of isotopologues due to its greater mass resolution and accuracy.

2.2.3 ISOSCAN: AN R PACKAGE TO PROCESS TARGETED GC-CI-MS DATA

To extend high throughput capabilities of GC-CI-MS we have developed isoSCAN, an open-source R-based computational tool (**https://github.com/jcapelladesto/isoSCAN**). isoSCAN encloses a complete workflow to perform high-throughput compound-driven isotopologues quantification in GC/MS data (either high or low resolution). isoSCAN capabilities have been assayed in several GC/MS isotope labelling studies[36–39] isoSCAN computational workflow is summarized in Figure 6. First, it requires as input data (i) GC–MS files in open standard formats (mz(X)ML); (ii) a list of compounds to target with their associated molecular formulas, monoisotopic mass values, and estimated retention times (Table 2.2). Then, isotopologue peaks are picked within a specified retention time range through local maxima detection. This peak-picking step slightly differs from low to high mass resolution data. In the case of low resolution instruments, isotopologue mass is computed by adding corresponding enrichment isotope mass difference (i.e., 1.003355 for 13C) as many times as the maximum number of atoms present in the molecular formula (i.e., up to six times in the case of glucose (C6H12O6)). For high-resolution data, the quantification supports both profile and centroided data, yet for computational reasons we recommend centroiding.

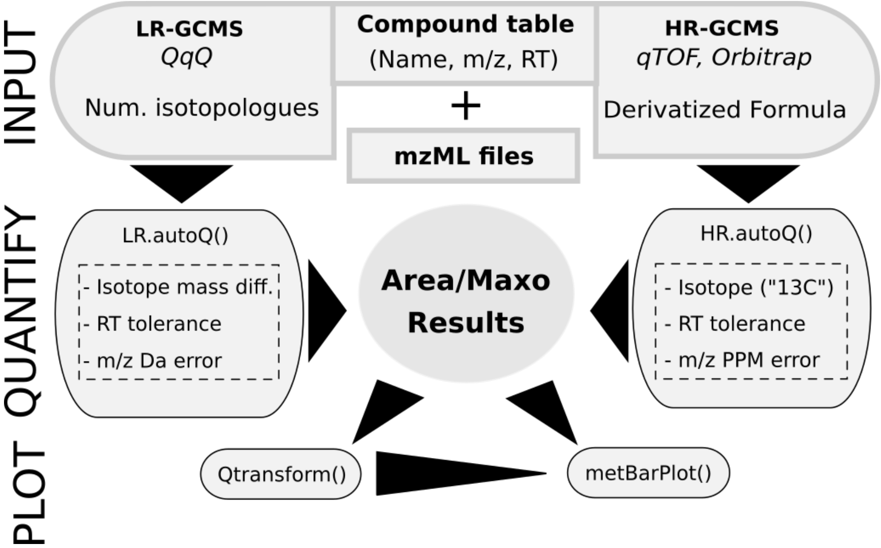


Figure 25. Data input and function hierarchy in isoSCAN.

isoSCAN was developed to automatically quantify all the isotopologues of a given list of compounds. isoSCAN requires a list of compound formulas, their molecular ion m/z and their estimated retention times, in order to extract the abundances of isotopologue (Table 2.2). Additionally, the package includes functions for data transformation and plotting (Figure 25). The main feature of isoSCAN is that it searches, and quantifies, isotopologue peaks by looking for a local maximum using the retention time value as a guide. However, the peak finding algorithm is different depending on instrument resolution:

- **LR.autoQ.** In the case of low resolution instruments, mainly nominal mass data, isotopologue masses are simply calculated by the addition of the mass difference corresponding to the labelling isotope of choice.
- **HR.autoQ.** In the case of high resolution instruments, we observed that isotopologue mass calculation needs to be refined as data may include other isotopic distributions (Si, N or O) that are relatively less

intense but still must be considered for accurate quantification. Therefore, we use enviPat[40] R package to calculate and merge the theoretical isotopic distribution for every possible labelled isotopologue, each is searched independently. Afterwards, isotopologue abundances are added by considering the number of labelled atoms to which each isotopologue corresponds (see below, Table 2, for an example in the case of isotopologues M+0 to M+3 of 2TMS Lactate).

Table 7. Isotopologues to consider in HRMS for Lactate 2TMS (C9H24O3Si2). Each color determines ions that should be summed for correct isotopologue estimation.

m/z	Num ^{12}C	Num ^{13}C	Num ^{28}Si	Num ^{29}Si	Num ^{30}Si
236,1258	9	0	2	0	0
237,1254	9	0	1	1	0
237,1292	8	1	2	0	0
238,1226	9	0	1	0	1
238,1287	8	1	1	1	0
238,1325	7	2	2	0	0
239,1261	8	1	1	0	1
239,1321	7	2	1	1	0
239,1359	6	3	2	0	0
240,1294	7	2	1	0	1
240,1354	6	3	1	1	0
241,1327	6	3	1	0	1

2.3 Discussion and conclusions

MS is currently the preferred analytical technique used for measuring isotopic labelling of intracellular metabolites. In this regard, GC/MS is characterized by low cost and relatively simple maintenance, which makes it more affordable than LC-MS equipment by many individual laboratories. A desirable scenario of stable-isotope labelling is to measure isotopologue distributions in intact metabolites, however, the inherent fragmentation of EI (the most widespread used ionization mode in GC) cause scattered isotope patterns. To minimize fragmentation, different analytical strategies have been implemented, including chemical derivatization with N-methyl N-(*tert*-butylsilyl)trifluoroacetamide (MTBSTFA)[14], EI ionization at 15 eV [29] or the use of chemical ionization[35]. The latter has been traditionally implemented using methane as reagent gas[29], however, isobutane produces superior analyte signal abundance to methane [41], resulting also in a predominant protonated adduct ion and less fragmentation. Here we have used this previous knowledge to demonstrate that CI-isobutane outperforms CI-methane, EI at 70eV and EI at 15eV for isotopologue analysis by GC/MS. We have also shown that the metabolic coverage of CI-isobutane allows analysis of most relevant metabolites in central carbon metabolism with good ionization efficiency, including intermediates of glycolysis, TCA cycle, pentose phosphate pathway, amino acids, urea cycle and polyamines. In addition, we have benefited from this better suited ionization source to explore different acquisition modes from the most common mass analyzers in metabolomics: time-of-flight (TOF), orbitrap, and quadrupole. Quadrupoles act as low-resolution mass filters and can be placed in series (e.g., triple quadrupoles) to measure only a predefined targeted subset of ions, offering in principle the best sensitivity for measuring a single metabolite. Our results showed that monitoring every isotopologue by SIM or opening the quadrupole 8-10 Da to scan the whole cluster of isotopologues of a metabolite, yielded similar results with regard to sensitivity; that is to say, despite the better signal-to-noise ratios of SIM or the higher absolute intensity counts of the narrow scan acquisition, the same isotopologues were detected for low abundant metabolites. The isotope ratio accuracy is slightly better for qOrbitrap due to its superior resolution and sensitivity. Remarkably, our results indicate that cutting-edge high-resolution GC/MS instruments have a comparable, or even better, performance than quadrupole systems, improving isotopologue quantification for low abundant compounds (e.g., dihydroxyacetone-phosphate and aconitate) in complex biological samples. Barely detectable isotopologues in SIM and/or NS showed a considerably higher response in full scan acquisition mode with HRMS instruments such as GC Orbitrap MS. In addition, exact mass instruments bring the advantage of distinguishing m/z interferences that may hinder the quantification of isotopologues.

Finally, we have developed a versatile open-source software that can automatically process low- and high-resolution isotopic labelling data produced with GC CI-isobutane coupled to quadrupole, TOF and orbitrap instruments.

Table 8. Compounds resolved using CI-isobutane detection and quantification. TMS (trimethyl-silyl); MA (methoxamine). *Method 2 also includes part of TCA cycle and some aminoacids

Compound Name	RT (min)	Compound Formula	Detected derivative	Derivative Formula	Detected m/z
Method 1. Glycolysis, TCA and Aminoacids					
Proline	3,76	C5H9NO2	1TMS	C8H18NO2Si	188,1101
Pyruvate	3,9	C3H4O3	TMS+ MA	C7H16O3NSi	190,0891
Lactate	4,01	C3H6O3	2TMS	C9H24O3Si2	236,1263
Valine	5,91	C5H11NO2	3TMS	C11H30NO2Si3	262,1653
Isoleucine	5,95	C6H13NO2	2TMS	C12H29NO2Si2	276,1808
Leucine	6,22	C6H13NO2	2TMS	C12H29NO2Si2	276,1808
Glycine	6,37	C2H5NO2	3TMS	C11H30NO2Si3	292,1571
Succinate	6,84	C4H4O4	2TMS	C10H23O4Si2	263,1134
Alanine	6,97	C3H7NO2	3TMS	C12H32NO2Si3	306,1727
Serine	7,03	C3H7NO3	3TMS	C12H31NO3Si3	322,1683
Fumarate	7,22	C4H2O4	2TMS	C10H21O4Si2	261,0978
Oxaloacetate	8,2	C4H2O5	2TMS+ MA	C11H24O5NSi2	306,1193
Malate	8,95	C4H4O5	3TMS	C13H31O5Si3	351,1479
Aspartate	8,85	C4H7NO4	3TMS	C13H31NO4Si3	350,1632
Urea	9,28	CH4N2O	2TMS	C7H21N2OSi2	205,1182

Ketoglutarate	9,88	C5H4O5	2TMS+MA	C12H26O5NSi2	320,1349
Phosphoenolpyruvate	10,18	C3H3O6P	3TMS	C12H30O6PSi3	385,1087
Glutamate	10,35	C5H7O4N	3TMS	C14H34O4NSi3	364,1795
Glyceraldehyde 3-phosphate	11,44	C3H5O6P	3TMS+MA	C13H35O6NPSi3	416,1509
Aconitate	11,61	C6H3O6	3TMS	C15H31O6Si3	391,1428
Dihydroxyacetone phosphate	11,66	C3H5O6P	3TMS+MA	C13H35O6NPSi3	416,1509
Glutamine	11,83	C5H9O3N2	3TMS	C14H35O3N2Si3	363,1955
Glycerate 2-phosphate	11,98	C3H5O7P	4TMS	C15H40O7PSi4	475,1588
Glycerate 3-phosphate	12,18	C3H5O7P	4TMS	C15H40O7PSi4	475,1588
Citrate	12,25	C6H5O7	4TMS	C18H41O7Si4	481,1929
Glucose	12,87	C6H12O6	5TMS+MA	C22H56O6NSi5	570,2953
Ribose 5-phosphate	13,76	C5H11O8P	5TMS+MA	C21H54NO8PSi5	620,2505
Arginine	14,43	C6H14N4O2	5 TMS	C21H52N3O2Si5	520,0919
Fructose 6-phosphate	15,05	C6H12O9P	6TMS+MA	C25H65O9NPSi6	722,3012
Glucose 6-phosphate	15,12	C6H12O9P	6TMS+MA	C25H65O9NPSi6	722,3012
Lysine	18,09	C6H14N2O2	4TMS	C18H46N2O2Si4	435,2708

Compound Name	RT (min)	Compound Formula	Detected derivative	Derivative Formula	Detected m/z
Method 2*. Nitrogen metabolism					
Creatine	15,23	C4H9N3O2	3TMS	C13H31N3OSi3	330,1841
Agmatine	16,6	C5H14N4	3TMS	C14H35N3Si3	330,2203
Arginino-succinate	16,87	C10H18N4O6	5TMS	C11H31N4O10Si5	607,3362
Putrescine	16,97	C4H12N2	4TMS	C16H45N2Si4	377,2643
Ornithine	17,1	C5H12N2O2	3TMS	C14H37N2O2Si3	349,2148
Cadaverine	17,71	C5H14N2	4TMS	C17H47N2Si4	391,2799
Histamine	17,71	C5H9N3	3TMS	C14H34N3Si3	328,2047
Tyramine	18,36	C8H11NO	3TMS	C17H36NOSi3	354,2091
Citruline	19,31	C6H13N3O3	3TMS	C15H38N3O3Si3	392,2206
Spermidine	20,5	C7H19N3	5TMS	C22H60N3Si5	506,3613

References

(1) Fiehn, O. *Curr. Protoc. Mol. Biol.* **2016**, *114*, 30.4.1–30.4.32.

(2) Dunn, W. B.; Broadhurst, D. I.; Atherton, H. J.; Goodacre, R.; Griffin, J. L. *Chem. Soc. Rev.* **2011**, *40* (1), 387–426.

(3) Maria Vinaixa, †; Sonia Marín, ‡; Jesús Brezmes, *,†; Eduard Llobet, †; Xavier Vilanova, †; Xavier Correig, †; Antonio Ramos, ‡ and; Sanchis‡, V. **2004**.

(4) Krilaviciute, A.; Heiss, J. A.; Leja, M.; Kupcinskas, J.; Haick, H.; Brenner, H. *Oncotarget* **2015**, *6* (36), 38643–38657.

(5) 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. *PLoS One* **2011**, *6* (2), e16957.

(6) Bouatra, S.; Aziat, F.; Mandal, R.; Guo, A. C.; Wilson, M. R.; Knox, C.; Bjorndahl, T. C.; Krishnamurthy, R.; Saleem, F.; Liu, P.; Dame, Z. T.; Poelzer, J.; Huynh, J.; Yallou, F. S.; Psychogios, N.; Dong, E.; Bogumil, R.; Roehring, C.; Wishart, D. S. *PLoS One* **2013**, *8* (9), e73076.

(7) Sreekumar, A.; Poisson, L. M.; Rajendiran, T. M.; Khan, A. P.; Cao, Q.; Yu, J.; Laxman, B.; Mehra, R.; Lonigro, R. J.; Li, Y.; Nyati, M. K.; Ahsan, A.; Kalyana-Sundaram, S.; Han, B.; Cao, X.; Byun, J.; Omenn, G. S.; Ghosh, D.; Pennathur, S.; Alexander, D. C.; Berger, A.; Shuster, J. R.; Wei, J. T.; Varambally, S.; Beecher, C.; Chinnaiyan, A. M. *Nature* **2009**, *457* (7231), 910–914.

(8) Kopka, J.; Schauer, N.; Krueger, S.; Birkemeyer, C.; Usadel, B.; Bergmuller, E.; Dormann, P.; Weckwerth, W.; Gibon, Y.; Stitt, M.; Willmitzer, L.; Fernie, A. R.; Steinhauser, D. *Bioinformatics* **2005**, *21* (8), 1635–1638.

(9) Jang, C.; Chen, L.; Rabinowitz, J. D. *Cell* **2018**, *173* (4), 822–837.

(10) Gibbs, M.; Kandler, O. *Proc. Natl. Acad. Sci. U. S. A.* **1957**, *43* (6), 446–451.

(11) Katz, J. *Med. Sci. Sports Exerc.* **1986**, *18* (3), 353–359.

(12) Batista Silva, W.; Daloso, D. M.; Fernie, A. R.; Nunes-Nesi, A.; Araújo, W. L. *Plant Sci.* **2016**, *249*, 59–69.

(13) Zamboni, N.; Fendt, S.-M.; Rühl, M.; Sauer, U. *Nat. Protoc.* **2009**, *4* (6), 878–892.

(14) Higashi, R. M.; Fan, T. W.-M.; Lorkiewicz, P. K.; Moseley, H. N. B.; Lane, A. N. In *Methods in molecular biology (Clifton, N.J.)*; 2014; Vol. 1198, pp 147–167.

(15) Sauer, U. *Mol. Syst. Biol.* **2006**, *2*, 62.

(16) Schlotterbeck, G.; Ross, A.; Dieterle, F.; Senn, H. *Pharmacogenomics* **2006**, *7* (7), 1055–1075.

(17) Zamboni, N. Springer, Berlin, Heidelberg, 2007; pp 129–157.

(18) Wiechert, W.; Möllney, M.; Petersen, S.; de Graaf, A. A. *Metab. Eng.* **2001**, *3* (3), 265–283.

(19) Jang, C.; Chen, L.; Rabinowitz, J. D. *Cell* **2018**, *173* (4), 822–837.

(20) Bunk, B.; Kucklick, M.; Jonas, R.; Münch, R.; Schobert, M.; Jahn, D.; Hiller, K. *Bioinformatics* **2006**, *22* (23), 2962–2965.

(21) Lei, Z.; Li, H.; Chang, J.; Zhao, P. X.; Sumner, L. W. *Metabolomics* **2012**, *8* (S1), 105–110.

(22) Hiller, K.; Hangebrauk, J.; Jäger, C.; Spura, J.; Schreiber, K.; Schomburg, D. *Anal. Chem.* **2009**, *81* (9), 3429–3439.

(23) 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. *Anal. Chem.* **2016**, *88* (19), 9821–9829.

(24) Ji, H.; Zhang, Z.; Lu, H. *Metabolomics* **2018**, *14* (5), 68.

(25) Wills, J.; Edwards-Hicks, J.; Finch, A. J. *Anal. Chem.* **2017**, *89* (18), 9616–9619.

(26) Wei, X.; Shi, B.; Koo, I.; Yin, X.; Lorkiewicz, P.; Suhail, H.; Rattan, R.; Giri, S.; McClain, C. J.; Zhang, X. *Anal. Chim. Acta* **2017**, *980*, 25–32.

(27) Ferrazza, R.; Griffin, J. L.; Guella, G.; Franceschi, P. *Bioinformatics* **2017**, *33* (2), 300–302.

(28) Dagley, M. J.; McConville, M. J. *Bioinformatics* **2018**, *34* (11), 1957–1958.

(29) Mairinger, T.; Sanderson, J.; Hann, S. *Anal. Bioanal. Chem.* **2019**, *411* (8), 1495–1502.

(30) Selivanov, V. A.; Benito, A.; Miranda, A.; Aguilar, E.; Polat, I. H.; Centelles, J. J.; Jayaraman, A.; Lee, P. W. N.; Marin, S.; Cascante, M. *BMC Bioinformatics* **2017**, *18* (1), 88.

(31) De Mas, I. M.; Selivanov, V. A.; Marin, S.; Roca, J.; Orešič, M.; Agius, L.; Cascante, M. *BMC Syst. Biol.* **2011**, *5*, 175.

(32) Kalderon, B.; Korman, S. H.; Gutman, A.; Lapidot, A. *Am. J. Physiol. Metab.* **1989**, *257* (3), E346–E353.

(33) Benito, A.; Polat, I. H.; Noé, V.; Ciudad, C. J.; Marin, S.; Cascante, M.; Benito, A.; Polat, I. H.; Noé, V.; Ciudad, C. J.; Marin, S.; Cascante, M.; Benito, A.; Polat, I. H.; Noé, V.; Ciudad, C. J.; Marin, S.; Cascante, M. *Oncotarget* **2017**, *8* (63), 106693–106706.

(34) Schricker, T.; Albuszies, G.; Kugler, B.; Wachter, U.; Georgieff, M. *Nutrition 10* (4), 342–345.

(35) Mairinger, T.; Steiger, M.; Nocon, J.; Mattanovich, D.; Koellensperger, G.; Hann, S. *Anal. Chem.* **2015**, *87* (23), 11792–11802.

(36) Bueno, M. J.; Jimenez-Renard, V.; Samino, S.; Capellades, J.; Junza, A.; López-Rodríguez, M. L.; Garcia-Carceles, J.; Lopez-Fabuel, I.; Bolaños, J. P.; Chandel, N. S.; Yanes, O.; Colomer, R.; Quintela-Fandino, M. *Nat. Commun.* **2019**, *10* (1), 5011.

(37) Llinàs-Arias, P.; Rosselló-Tortella, M.; López-Serra, P.; Pérez-Salvia, M.; Setién, F.; Marin, S.; Muñoz, J. P.; Junza, A.; Capellades, J.; Calleja-Cervantes, M. E.; Ferreira, H. J.; Moura, M. C. de; Srbic, M.; Martínez-Cardús, A.; Torre, C. de la; Villanueva, A.; Cascante, M.; Yanes, O.; Zorzano, A.; Moutinho, C.; Esteller, M. *JCI Insight* **2019**, *4* (8).

(38) Cano-Crespo, S.; Chillarón, J.; Junza, A.; Fernández-Miranda, G.; García, J.; Polte, C.; R. de la Ballina, L.; Ignatova, Z.; Yanes, Ó.;

Zorzano, A.; Stephan-Otto Attolini, C.; Palacín, M. *Sci. Rep.* **2019**, *9* (1), 14065.

(39) Soukupova, J.; Malfettone, A.; Hyroššová, P.; Hernández-Alvarez, M.-I.; Peñuelas-Haro, I.; Bertran, E.; Junza, A.; Capellades, J.; Giannelli, G.; Yanes, O.; Zorzano, A.; Perales, J. C.; Fabregat, I. *Sci. Rep.* **2017**, *7* (1), 12486.

(40) Loos, M.; Gerber, C.; Corona, F.; Hollender, J.; Singer, H. *Anal. Chem.* **2015**, *87* (11), 5738–5744.

(41) Newsome, G. A.; Steinkamp, F. L.; Giordano, B. C. *J. Am. Soc. Mass Spectrom.* **2016**, *27* (11), 1789–1795.

CHAPTER 3. geoRge: a computational tool to detect the presence of stable isotope labelling in LC/MS-based untargeted metabolomics

3.1 Introduction

In metabolism, the level of intracellular metabolites remains stable and relatively constant by homeostatic mechanisms that regulate the flow of producing and consuming metabolic reactions in the cell[1]. Many diseases, however, are caused by disturbances in homeostasis that provoke changes in metabolic fluxes and thus metabolite levels[2].

Global metabolite profiling, also known as untargeted metabolomics, has become a powerful approach to provide information about metabolite levels[3], which can be later used to interrogate mechanistic biochemistry in health and disease[4]. Despite that, conventional untargeted metabolomics experiments result in a static snapshot of metabolism that does not provide details about the evolution of metabolic fluxes aimed at maintaining homeostasis and metabolite concentrations. In this regard, modern stable-isotope metabolic flux analysis is becoming an essential tool to unravel mechanisms of metabolic regulation and identify therapeutic targets in diseased cells[5]. This approach makes use of stable isotopically-labelled substrates (e.g., ^{13}C-glucose, ^{13}C-glutamine) to trace back the cellular fate of labelled atoms into the structures of transformed metabolites. With the advent of comprehensive metabolic profiling technologies such as LC/MS and GC/MS, it opens possibilities to broaden the coverage of stable-isotope tracing studies, allowing for an unbiased mapping of fluxes through multiple metabolic pathways[6–9].

However, global tracking of isotopic labels in untargeted LC/MS-based datasets has challenges that hamper the extraction and relative quantification of isotope peaks from labelled data. Lately, a collection of workflows and bioinformatic tools have been developed to automatically detect the isotopologues (i.e., labelled metabolite peaks) formed when employing stable-isotope labelling. Among these stand out X^{13}CMS[6], mzMatch-ISO[7,10] and MetExtract[8] which, each in their own way, share a similar principle for annotating and quantifying monoisotopic and corresponding isotope-labelled peaks of metabolites. These tools are based on a brute force approach by iterating over all MS signal data in each mass spectrum using the theoretical mass difference between the light and heavy stable isotope, such as ^{12}C and ^{13}C. Based on our experience this type of iterative approach may lead to false positive results due to the massive number of peaks in the untargeted LC/MS-based datasets.

Here we describe a novel approach for the annotation and relative quantification of isotope-labelled mass spectrometry data that benefits from a simple and robust trait of the labelled mass spectra when it is compared to the unlabelled equivalent. Spectral peaks originated from labelled metabolites show higher intensity or may even appear as new peaks in the mass spectra, in consequence being distinguishable features by statistical testing. We have exploited this trait to develop geoRge, a new computational tool written in the open language R that runs on features (unique m/z and RT) detected as using the widespreadly-used XCMS package[11,12]. The automated untargeted isotope annotation and relative quantification capabilities of geoRge are demonstrated by the analysis of LC/MS data from a human retinal pigment epithelium cell line (ARPE-19) grown on normal and high glucose concentrations that mimics diabetic retinopathy conditions *in vitro*.

3.2 Results and discussion

3.2.1 COMPUTATIONAL WORKFLOW

Our computational approach works on the basis of an experimental design that includes replicates of unlabelled and labelled (e.g., D-[U-^{13}C]-glucose) biologically equivalent samples. The computational tool uses four functions that are described below (Figure 26).

(i) **Detection of putative incorporations**: the function *PuInc_seeker* generates a list of mzRT features that we consider as isotopologues, namely metabolites that incorporate stable isotope atoms from the labelled source, mentioned in the following as PuInc. To do this, the function performs a Welch's t-test to compare labelled and unlabelled biologically equivalent samples from the XCMS matrix of mzRT features, that is, the xcmsSet object. We set default p-value and fold-change thresholds at 0.05 and 1.2 respectively--- yet this can be changed by the user--- to filter out non-significant features and retain those that are significantly up-regulated in the labelled samples by comparison with the unlabelled ones (Figure 24A). *PuInc_seeker* entirely relies on statistical criteria for finding PuInc features. This allows us to assess the biological variability of the metabolic fate of labelled atoms at the first step of the workflow.

Although the effort of generating more samples (i.e. an identical unlabelled sample for every labelled sample) might be seen as a minor limitation, this constitutes a major advantage of our method. Within the same experiment, our workflow allows studying both metabolites pool size (by comparing unlabelled samples) and stable isotope enrichment (by comparing unlabelled vs. labelled samples), which constitute key complementary metabolic results.

Note: the reliance on XCMS to detect all isotopologues in the initial pre-processing step is critical. Due to different experimental factors, there is not a perfect combination of parameters in XCMS---or any other existing peak-picking tool--- that can detect all relevant features in an untargeted LC/MS-based approach. This may impose some limitations to any computational approach for global tracking of isotopic labels since it is common for metabolites to be present as a combination of multiple isotopologues, many with low peak abundance.

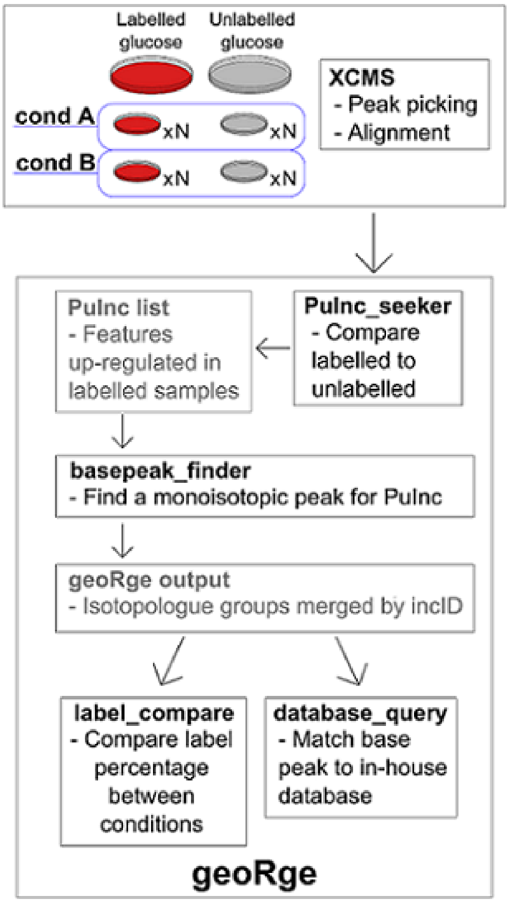

Labelled glucose
Unlabelled glucose
cond A
cond B
xN
XCMS
- Peak picking
- Alignment
PuInc list
- Features up-regulated in labelled samples
PuInc_seeker
- Compare labelled to unlabelled
basepeak_finder
- Find a monoisotopic peak for PuInc
geoRge output
- Isotopologue groups merged by inclD
label_compare
- Compare label percentage between conditions
database_query
- Match base peak to in-house database
geoRge

Figure 26. Workflow for geoRge experiments. The experimental design includes N replicates of unlabelled and labelled biologically equivalent samples from different experimental conditions. geoRge implements four functions (PuInc_seeker, basepeak finder, label_compare and database_query are highlighted in black) on the output matrix of XCMS to find putative isotopologues and the monoisotopic (unlabelled) mass peaks of these isotopologues. Finally, it annotates metabolites using an internal database and the accurate mass of the monoisotopic mass peaks, and compares the relative labelling of each isotopologue between experimental conditions.

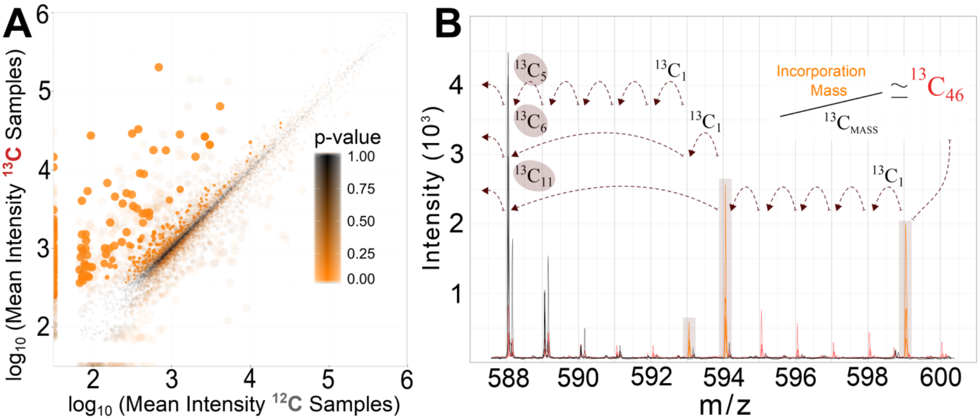


Figure 27. Representation of geoRge's principle. A) Metabolites that incorporate labelled atoms/moieties from the labelled source will appear anew or with greater intensity in the mass spectrum. The mzRT features (dots in the scatter plot) passing a fold-change and p-value filter are associated with metabolites incorporating one or more stable isotope atoms. B) In order to find the unlabelled monoisotopic peak of the labelled metabolites in the LC/MS dataset, the mass of the up-regulated features is divided by the mass of the stable isotope atom (in carbon experiments is 13.003355). The resulting value corresponds to the maximum number of carbon atoms the given mass could have. This is used to obtain a vector of masses that are searched in the unlabelled samples using a certain retention time and mass error window. Herein, the example for ADP-mannose, showing three different labelling moieties, either $^{13}C_5$, $^{13}C_6$ and $^{13}C_{11}$, which correspond to the different sugar substructures that conform the structure of this nucleotide sugar.

(ii) **Finding the monoisotopic (unlabelled) mass peak**: for each PuInc feature the function *basepeak_finder* looks for the monoisotopic (unlabelled) mzRT feature in the unlabelled samples, mentioned in the following as base peak. To do this, the function computes the maximum number of carbon atoms every PuInc can possibly contain by dividing its exact mass (mzmed using XCMS nomenclature) by the atomic mass of the stable isotope (e.g., 13.003355 for ^{13}C). The resulting number is used to calculate a vector of masses that correspond to all candidate base peaks (Figure 27B). In order to find the presence of such base peaks in the unlabelled samples, the function determines a retention time window centered at the median retention time value (rtmed using XCMS nomenclature) of the PuInc feature and a mass error of 6.5 ppm. Any peak that does not fall within these two criteria is discarded. Next, both intensity and signal-to-noise (S/N) thresholds are applied to avoid annotation of unresolved peaks as base peaks. The mean intensity threshold is calculated in the unlabelled samples for each condition independently, which by default is set to 2000 counts (maxo value using XCMS nomenclature) --- yet the user can tune it as necessary. The minimum S/N applied is calculated from the 20th percentile of the total distribution of S/N in the XCMS dataset. Yet, if still more than one feature is found as base peak, the function only retains base peaks with an intensity of at least 70% of the highest candidate base peak. The aim of these thresholds is to minimize false positive base peaks due to the natural isotopic distribution of metabolites, instrumental noise or other unresolved peaks. Finally, every PuInc features linked to the same candidate base peak are grouped in one “incorporation ID”, mentioned in the following as incID. If more than one base peak is found for the same PuInc feature, the different base peaks are given distinct incIDs.

This process is fairly time-efficient; the time required for completing the described workflow depends on the dimensions of the XCMS matrix, defined by the number of samples and mzRT features. Though the rate is about 500 mzRT features per minute on an average-performance computer (4GB RAM and a 2.3GHz dual core processor), meaning that a complete analysis for a 10.000 mzRT feature dataset of 20 samples is performed in less than 25 minutes.

(iii) **Percentage of labelling**: the function *label_compare* calculates the percentage of each PuInc (i.e., isotopologue) out of the sum of all PuInc in one IncID--- plus the base peak if present in the labelled samples. For each PuInc feature within a IncID group, its intensity is divided by the sum of all the other intensities and multiplied by 100. This is repeated for every replicate sample independently. Due to limitations of peak-picking algorithms, the output of this function is not intended to be an absolute quantification of the percentage of stable isotope labelling, but rather a relative quantification allowing a comparison of the level of isotope incorporation between experimental conditions (e.g., KO vs. wild-type). Manual extraction of raw signals from selected metabolites is preferred in order to quantify the percentage of labelling of isotopologues in an absolute manner.

(iv) Database query: the function *database_query* matches automatically the candidate base peaks against the Human Metabolome Database (HMDB 3.6)[13,14] with the objective of knowing the putative identity of the labelled metabolites. The default settings of *database_query* include the most common ion adducts ($M+H^+$, $M-H_2O+H^+$, $M+Na^+$, $M+NH_4^+$, $M-H^-$, $M-H_2O-H^-$) for each ionisation mode, and a 12 ppm mass accuracy---yet these settings can be changed by the user. The user can easily extend the default library by incorporating additional metabolites to build a personal database.

3.2.2 A COMPARATIVE ANALYSIS OF GEORGE AND X^{13}CMS IN THE STUDY OF HUMAN RETINAL PIGMENT EPITHELIAL CELLS EXPOSED TO HYPERGLYCEMIC CONDITIONS

Most investigations into the pathogenesis of diabetic retinopathy have been concentrated on the neural retina since this is where clinical lesions are manifested[15–17]. Recently, however, various abnormalities in the structural and secretory functions of retinal pigment epithelium (RPE) that are essential for neuroretina survival, have been found in diabetic retinopathy[18–20]. To illustrate the complete geoRge workflow described above, we investigated the changes in the labelling pattern occurring in human RPE cells exposed to hyperglycemic conditions, the major component of the diabetic milieu. In addition, we compared the results of geoRge with the outcome of $X^{13}CMS$[6], since both approaches rely entirely on XCMS parameters for mzRT feature selection, unlike MzMatch-ISO[7] and MetExtract[8]. In this way, differences can be fully attributed to the alternative strategies for annotating and quantifying monoisotopic and isotope-labelled peaks (i.e., isotopologues) of geoRge and $X^{13}CMS$. ARPE-19, a human RPE cell line, was cultured in triplicate in both unlabelled D-glucose and labelled D-[U-^{13}C]-glucose under 2 biological conditions: normal glucose concentration (5.5 mM) and high glucose concentration (25 mM). The samples were analyzed using LC-ESI-qTOF MS in negative ionisation mode (see Methods section for further details). Raw LC/MS files have been deposited in MetaboLights with accession number MTBLS213.

The XCMS dataset containing 14,607 mzRT features was used as input for geoRge and $X^{13}CMS$. From 270 and 411 PuInc in normoglycemic and hyperglycemic conditions, respectively, 377 PuInc were associated with monoisotopic (unlabelled) base peaks in geoRge, totaling 271 IncID (Supplementary File 1 found in the Supporting Information of the paper[21]). In contrast, $X^{13}CMS$ detected 1,176 isotopologues (Supplementary File 2 found in the Supporting Information of the paper[21]). The density distribution of m/z ions among the isotopologues found by geoRge and $X^{13}CMS$ is shown in Figure 25A, indicating that geoRge leads to greater detection of putative ^{13}C incorporations in the low mass range than $X^{13}CMS$. By contrast, $X^{13}CMS$ seems to find isotopologues evenly across the entire mass range, closely resembling the distribution of m/z ions of the xcmsSet object. The relationship between isotopologues shown in Figure 25B revealed that 21% of geoRge's ^{13}C incorporations are unique while the rest are also detected by X^{13}CMS. Yet the 879 putative incorporations by X^{13}CMS (75% of the total) that are not detected by geoRge is a remarkable fact. A detailed examination of these isotopologues according to the number of ^{13}C atoms incorporated, demonstrates that there is a predominance of 1-3 ^{13}C and >18 ^{13}C in X^{13}CMS compared with geoRge (Figure 28C).

A manual inspection of the MS signals in the 1-3 ^{13}C range reveals natural abundance isotopic distributions, whereas incorporations of >18 ^{13}C (reaching up to 104 ^{13}C) relates mainly to false positives. This can be also observed through a scatter (volcano) plot (Figure 28D) that relates fold-change to p-value between replicates of unlabelled and labelled hyperglycemic equivalent samples. The vast majority of mzRT features labelled as isotopologues by X^{13}CMS are also present in the unlabelled samples with similar or even higher intensity (Figure 28D-E). Only 4% of the isotopologues correspond to significantly upregulated mzRT features in the labelled samples (Figure 28F). After database searching in the HMDB[13,14] and METLIN[22], 45% and 57% of the candidate base peaks matched with known metabolites as $[M-H]^-$ ions, respectively. When this was done with X^{13}CMS, only 27% and 38% matched with known metabolites.

In summary, while geoRge detects apparently three times less incorporations of ^{13}C than X^{13}CMS, our results suggests that the latter approach generates many false positives that require extensive manual inspection.

We further studied the changes in the labelling pattern occurring in ARPE-19 cells exposed to hyperglycemic and normoglycemic conditions using exclusively geoRge. Our results indicate that high glucose concentrations activate additional metabolic reactions in RPE cells. Many metabolites belong to the central glycolytic route and nucleotide biosynthetic pathways. The identity of some of these metabolites was confirmed by MS/MS both in unlabelled and labelled samples. In order to elucidate the effect of hyperglycemic conditions in ARPE-19 cells, the percentage of ^{13}C labelling between normoglycemic and hyperglycemic conditions was compared using a Welch's t-test for each PuInc feature (this process is performed by *label_compare* function). Out of 171 unique labelled metabolites, 88 appeared differentially labelled.

Among these, the percentage of labelled D-glucose (m+6 ^{13}C peak at 185.0774) was, expectedly, enriched in hyperglycemic conditions relative to the normoglycemic samples. Interestingly, downstream intermediates of the glycolytic pathway such as 1,3-bisphosphoglycerate (m+3 ^{13}C peak at 267.9597), either 2- or 3-phosphoglycerate (m+3 ^{13}C peak at 187.9961), and phosphoenolpyruvate (m+3 ^{13}C peak at 169.9861) appeared to incorporate similar percentages of the stable isotope in hyperglycemic and normoglycemic conditions. However, metabolites involved in and generated from the TCA cycle showed a clear reduction in ^{13}C labelling in hyperglycemic conditions. These included malate (m+2 ^{13}C peak at 135.0225) and citrate (m+2 ^{13}C peak at 193.02879), glutamate (m+2 ^{13}C peak at 148.0537) and carnosine (m+3 ^{13}C peak at 228.1092), a metabolite that is synthesized mainly through glutamate-histidine transformation (Figure 29).

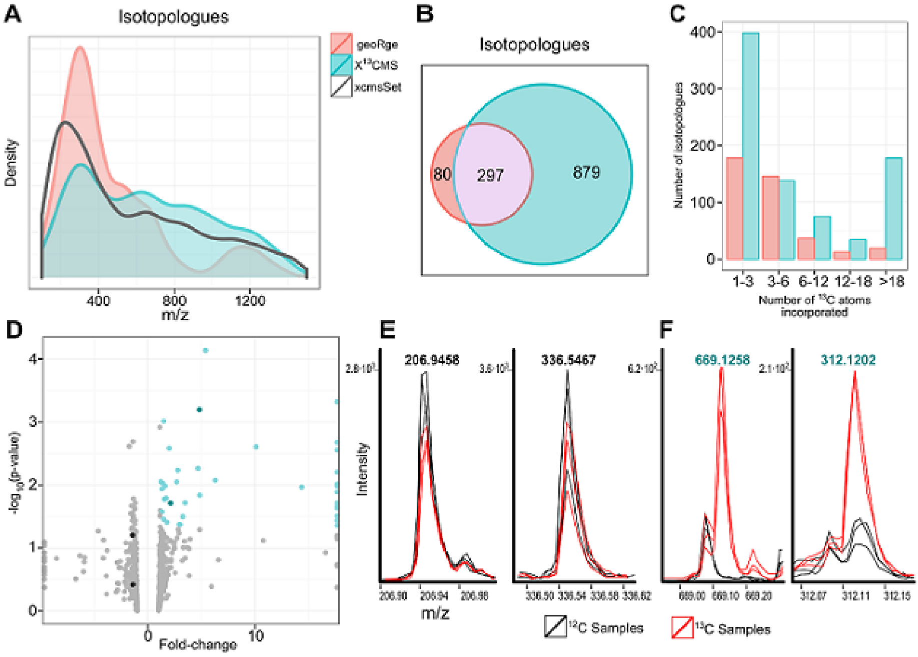


Figure 28. Comparative analysis of geoRge and X^{13}CMS. A) Density distribution of m/z ions among the isotopologues found by geoRge (red area) and X^{13}CMS (green area). The black line shows the distribution of m/z ions of the xcmsSet object. B) Venn diagram shows the relationship between isotopologues detected using geoRge and X^{13}CMS. C) Bar plot shows the distribution of isotopologues in geoRge and X^{13}CMS according to the number of ^{13}C atoms incorporated. D) Scatter (volcano) plot of the 879 unique isotopologues by X^{13}CMS that relates the fold-change to p-value between the three replicates of unlabelled and labelled equivalent hyperglycemic samples. Green dots are significantly upregulated mzRT features in the labelled samples, which correspond to 4% of the total. Two black and two dark green dots are highlighted because their mass spectra is displayed in E) and F).

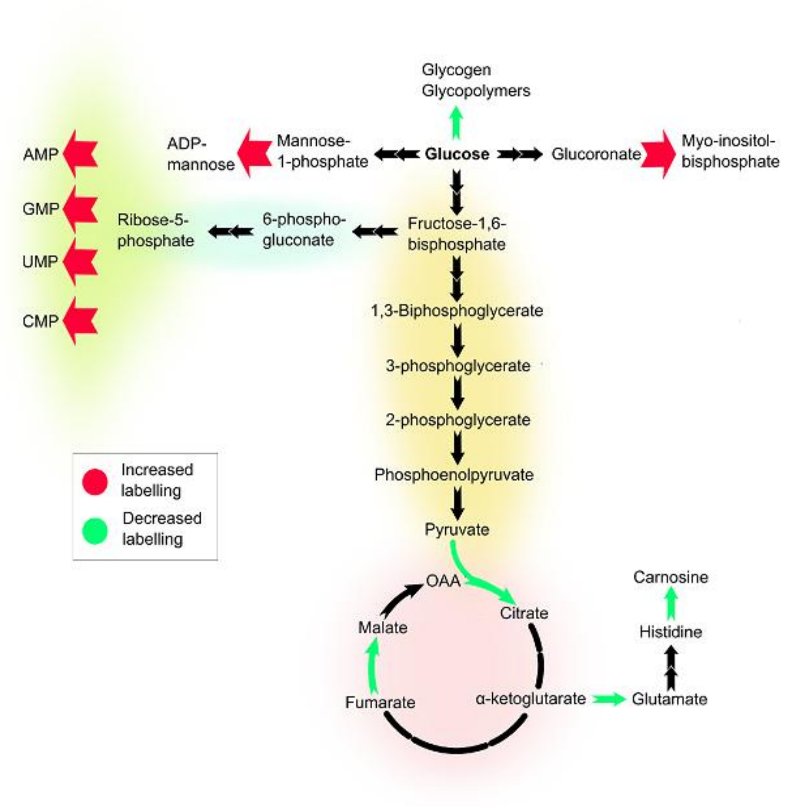


Figure 29. Summary of the effect of hyperglycemic conditions on ARPE-19 cells. TCA metabolites appear to have a diminished labelling proportion in comparison to normoglycemic conditions. Alternative pathways that promote labelling of pentose and hexose sugars in nucleotides or nucleotide sugar structures appear to have a much higher labelling proportion under high glucose concentrations. In contrast, the labelling percentage of polysaccharides (i.e., glycopolymers) is decreased relative to normoglycemic conditions.

Moreover, several metabolites containing pentose and hexose groups, as well as polymers consisting of glucose units were differentially labelled in hyperglycemic samples relative to the normoglycemic ones. Among these, myo-inositol bisphosphate (m+6 ^{13}C peak at 345.0118) and ADP-mannose (m+6 ^{13}C peak at 594.1006 corresponding to labelled mannose) was enriched in hyperglycemic conditions. Many nucleoside and nucleotide structures such as cytidine, uridine, CMP, UMP, AMP and GMP were highly enriched by m+5 ^{13}C in the hyperglycemic samples, which correlates with the incorporation of labelled ribose into their structures (Figure 30A). Finally, different polymers of glucose units tentatively assigned to sucrose, maltotriose and glycogen appeared to incorporate less ^{13}C atoms in the hyperglycemic samples relative to the normoglycemic ones. Although we could not identify these polymers by MS/MS due to the lack of reference spectra in databases, the pattern of isotopologues reinforces the accurate mass matching assignments. Sucrose, a disaccharide combination of the monosaccharides glucose and fructose was enriched in the m+12 ^{13}C peak at 353.1512; maltotriose, a trisaccharide consisting of three glucose molecules was enriched in the m+6, m+12 and m+18 ^{13}C peaks at 509.1867, 515.2067 and 521.2270 respectively; and glycogen, which according to HMDB contains 4 units of glucose, was enriched in the m+12 and m+24 ^{13}C peaks at 677.2598 and 689.3020 respectively (Figure 30B). Note that the m/z 683.2759 was not annotated as m+18 ^{13}C peak because the intensity of this peak in unlabelled samples is higher than in labelled samples. The m/z 683.2759 may correspond to a glycogen form containing 6 glucose units, unfortunately this mass is not linked with any compound in databases.

3.3 Conclusions

geoRge is an open easy-to-use tool that has been designed for rapid global tracking of stable isotope labelling in untargeted metabolomics. Here, we have demonstrated that the detection of new mass spectral peaks that appear in samples that were fed an isotopically labelled precursor is a more robust strategy for tracking the fate of stable isotopes than iterative approaches over all MS signal data that are characteristic of existing tools. We anticipate that this approach will have the same outcome when using other atom variants (^{15}N, ^{33}S and ^{18}O). Importantly, the raw LC/MS experimental data described herein is publicly available through MetaboLights[23] to ensure data traceability and reproducibility, and enabling for comparison with other existing and future approaches. Similarly, geoRge is available as an R script at **https://github.com/jcapelladesto/geoRge**.

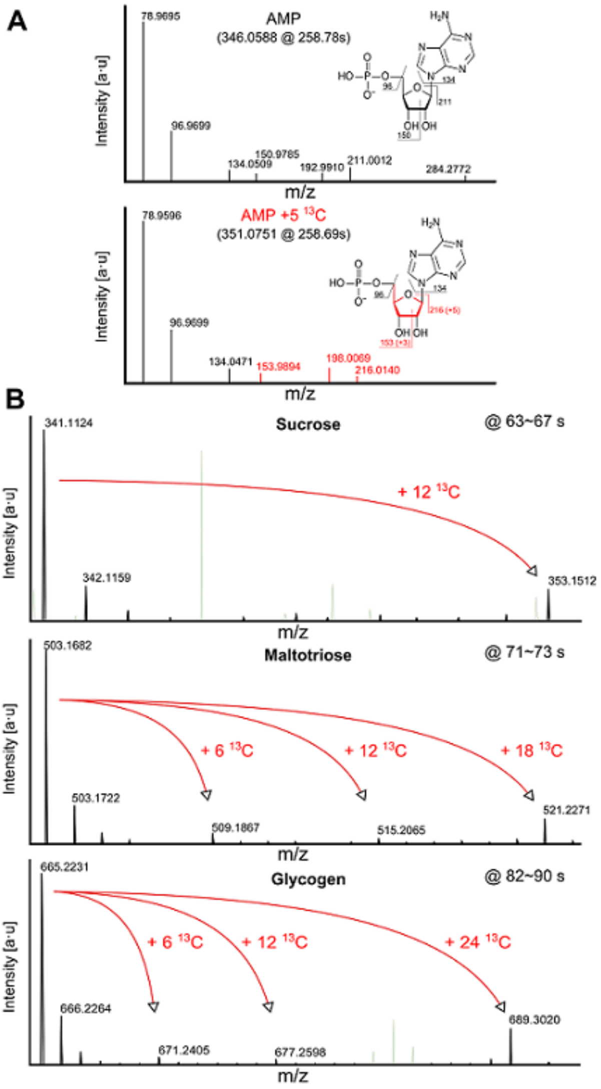


Figure 30. Labelling pattern of sugar-containing compounds. A) MS/MS spectra of unlabelled AMP and its m+5 ^{13}C isotopologue, showing that the fate of D-[U-^{13}C]-glucose atoms is the ribose subunit of the nucleotide. Only the fragments including part or the complete pentose structure gain the equivalent mass difference of ^{13}C. B) Hexose polymers showing the recursive incorporation pattern of m+6 ^{13}C units.

References

(1) Metallo, C. M.; Vander Heiden, M. G. *Mol. Cell* **2013**, *49* (3), 388–398.

(2) Schoors, S.; Bruning, U.; Missiaen, R.; Queiroz, K. C. S.; Borgers, G.; Elia, I.; Zecchin, A.; Cantelmo, A. R.; Christen, S.; Goveia, J.; Heggermont, W.; Goddé, L.; Vinckier, S.; Van Veldhoven, P. P.; Eelen, G.; Schoonjans, L.; Gerhardt, H.; Dewerchin, M.; Baes, M.; De Bock, K.; Ghesquière, B.; Lunt, S. Y.; Fendt, S.-M.; Carmeliet, P. *Nature* **2015**, *520* (7546), 192–197.

(3) Zamboni, N.; Saghatelian, A.; Patti, G. J. *Mol. Cell* **2015**, *58* (4), 699–706.

(4) Patti, G. J.; Yanes, O.; Siuzdak, G. *Nat. Rev. Mol. Cell Biol.* **2012**, *13* (4), 263–269.

(5) Buescher, J. M.; Antoniewicz, M. R.; Boros, L. G.; Burgess, S. C.; Brunengraber, H.; Clish, C. B.; DeBerardinis, R. J.; Feron, O.; Frezza, C.; Ghesquiere, B.; Gottlieb, E.; Hiller, K.; Jones, R. G.; Kamphorst, J. J.; Kibbey, R. G.; Kimmelman, A. C.; Locasale, J. W.; Lunt, S. Y.; Maddocks, O. D.; Malloy, C.; Metallo, C. M.; Meuillet, E. J.; Munger, J.; Nöh, K.; Rabinowitz, J. D.; Ralser, M.; Sauer, U.; Stephanopoulos, G.; St-Pierre, J.; Tennant, D. A.; Wittmann, C.; Vander Heiden, M. G.; Vazquez, A.; Vousden, K.; Young, J. D.; Zamboni, N.; Fendt, S.-M. *Curr. Opin. Biotechnol.* **2015**, *34*, 189–201.

(6) Huang, X.; Chen, Y.-J.; Cho, K.; Nikolskiy, I.; Crawford, P. A.; Patti, G. J. *Anal. Chem.* **2014**, *86* (3), 1632–1639.

(7) Chokkathukalam, A.; Jankevics, A.; Creek, D. J.; Achcar, F.; Barrett, M. P.; Breitling, R. *Bioinformatics* **2013**, *29* (2), 281–283.

(8) Bueschl, C.; Kluger, B.; Berthiller, F.; Lirk, G.; Winkler, S.; Krska, R.; Schuhmacher, R. *Bioinformatics* **2012**, *28* (5), 736–738.

(9) Hiller, K.; Metallo, C. M.; Kelleher, J. K.; Stephanopoulos, G. *Anal. Chem.* **2010**, *82* (15), 6621–6628.

(10) Creek, D. J.; Jankevics, A.; Burgess, K. E. V; Breitling, R.; Barrett, M. P. *Bioinformatics* **2012**, *28* (7), 1048–1049.

(11) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78* (3), 779–787.

(12) Tautenhahn, R.; Patti, G. J.; Rinehart, D.; Siuzdak, G. *Anal. Chem.* **2012**, *84* (11), 5035–5039.

(13) Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; Fung, C.; Nikolai, L.; Lewis, M.; Coutouly, M.-A.; Forsythe, I.; Tang, P.; Shrivastava, S.; Jeroncic, K.; Stothard, P.; Amegbey, G.; Block, D.; Hau, D. D.; Wagner, J.; Miniaci, J.; Clements, M.; Gebremedhin, M.; Guo, N.; Zhang, Y.; Duggan, G. E.; Macinnis, G. D.; Weljie, A. M.; Dowlatabadi, R.; Bamforth, F.; Clive, D.; Greiner, R.; Li, L.; Marrie, T.; Sykes, B. D.; Vogel, H. J.; Querengesser, L. *Nucleic Acids Res.* **2007**, *35* (Database issue), D521–D526.

(14) 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. *Nucleic Acids Res.* **2013**, *41* (Database issue), D801–D807.

(15) Cheung, N.; Mitchell, P.; Wong, T. Y. *Lancet* **2010**, *376* (9735), 124–136.

(16) Simó, R.; Villarroel, M.; Corraliza, L.; Hernández, C.; Garcia-Ramírez, M. *J. Biomed. Biotechnol.* **2010**, *2010*, 190724.

(17) Simó, R.; Carrasco, E.; García-Ramírez, M.; Hernández, C. *Curr. Diabetes Rev.* **2006**, *2* (1), 71–98.

(18) Barba, I.; Garcia-Ramírez, M.; Hernández, C.; Alonso, M. A.; Masmiquel, L.; García-Dorado, D.; Simó, R. *Invest. Ophthalmol. Vis. Sci.* **2010**, *51* (9), 4416–4421.

(19) Simó, R.; García-Ramírez, M.; Higuera, M.; Hernández, C. *Am. J. Ophthalmol.* **2009**, *147* (2), 319–325.e1.

(20) Joussen, A. M.; Smyth, N.; Niessen, C. *Dev. Ophthalmol.* **2007**, *39*, 1–12.

(21) Capellades, J.; Navarro, M.; Samino, S.; Garcia-Ramirez, M.; Hernandez, C.; Simo, R.; Vinaixa, M.; Yanes, O. *Anal. Chem.* **2016**, *88* (1), 621–628.

(22) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. *Ther. Drug Monit.* **2005**, *27* (6), 747–751.

(23) Haug, K.; Salek, R. M.; Conesa, P.; Hastings, J.; de Matos, P.; Rijnbeek, M.; Mahendraker, T.; Williams, M.; Neumann, S.; Rocca-Serra, P.; Maguire, E.; González-Beltrán, A.; Sansone, S.-A.; Griffin, J. L.; Steinbeck, C. *Nucleic Acids Res.* **2013**, *41* (Database issue), D781–D786.

DISCUSSION

In the latest quarter of century, -omics technologies have evolved into a collection of well-suited tools to extract information about the phenome, genome and epigenome. Nonetheless, one needs to go downstream translation to get a complete systems view. This is especially relevant for multifactorial and polygenic diseases deriving in complex phenotypes. Proteomics and metabolomics are essential to characterize such complex phenotypes. However, despite the fact that the complementarity of metabolomics and proteomics data is clear on paper, there is not yet a consensus on how to integrate experimental fluxomics or metabolomics and proteomics datasets to interrogate metabolism systems-wise. In this regard, this thesis presents a novel methodology to do so using genome-scale metabolic network reconstructions. These networks provide a platform to interpret -omics data in a biochemically meaningful manner. They offer a powerful abstraction to simulate the set of metabolic fluxes through the network eventually defining the cellular phenotype.

By employing the connectivity of Recon 2 (a genome-scale human metabolic reconstruction) I have developed a quantitative proteomics-guided strategy to reveal potential metabolic alterations. The idea is that a coordinated imbalanced expression of enzymes, as detected per quantitative proteomics, would eventually have an effect on the concentration or flux of those metabolites involved in the reactions regulated by these enzymes. Based on the expression profile of enzymatic proteins and their connectivity in Recon 2, our method outputs a list of metabolites whose production and consumption are potentially dysregulated resulting in altered concentrations or imbalanced fluxes. This method does not take into consideration protein regulation dynamics such as post-translational regulation, feedback inhibition or high-order conformations. In addition, neither kinetic nor contraints-based modeling is used but, just the network topology to predict the impact of a perturbation in the absence of kinetic parameters thereof neglecting the dynamic component of the metabolic network. It should be stated that our approach is not aimed to determine the absolute flux through this network as a formal fluxomics approach would do but rather to unbiasedly predict a set of metabolites with altered concentrations.

Here, conversely to traditional proteomics analysis where proteins are statistically treated as if they operated independently, the levels of proteins associated to a metabolite transformation according to network topology are evaluated as a whole. This leads to a novel and much realistic framework to interrogate metabolism. In this thesis, I have demonstrated this novel approach to successfully predict a set of metabolites with proved imbalanced fluxes or abundances and that resulted overlooked by traditional statistical analysis. Moreover, this has the potential to also be applied to transcriptomics data to the same end.

We hypothesize, however, that due to the usually low correlation between transcriptome and proteome (27–40%), the predictive value of this approach will decrease with gene expression data.

This novel method is not without limitations. On one hand, it requires a rather large coverage of the proteomics data so as to be able to map the maximum number of enzymatic reactions in the metabolic network. Metabolites whose reactions are left uncovered can not be tested. On the other hand, the statistical power of our binomial test dramatically decreases with metabolites that present a lower number of interconnections (usually end-products of synthesis or degradation pathways) and therefore our method is not sensitive enough to predict them.

Next, in order to confirm predicted flux disturbances, I have developed two novel strategies aimed at high-throughput experimental determination of metabolic fluxes based on stable isotope labelling experiments. The first strategy consists of a GC/(CI)MS-based analytical method directed towards high-throughput analysis and relative quantification of isotopologues of those intermediate metabolites in central carbon metabolism pathways: glycolysis, TCA cycle, pentose phosphate pathway and urea cycle. This experimental method is complemented by isoSCAN, an open-source package allowing automatic isotopologue quantification, isotopologue assignment, normalization and fractional enrichment, from either high or low-resolution GC/MS data. The second strategy is called to expand metabolic network coverage beyond this central carbon metabolism and therefore it uses an untargeted LC/(ESI)MS-based analytical approach to detect and quantify isotopologues. Again, to ease and automate data analysis I developed geoRge, an open-source package enclosing a statistically-oriented data analysis workflow to detect enriched isotopologue distributions in untargeted LC/MS data. These two strategies are called to simplify and automate isotope labelling experiments allowing high-throughput measurement of fractional enrichment in a wide range of metabolite intermediates. This fractional enrichment can in turn be considered as a qualitative measure of flux. Thus, both strategies have significantly broadened the coverage of traditional stable isotope labelling experiments allowing for an unbiased mapping of fluxes through multiple metabolic pathways.

Altogether, these three contributions provide novel tools to comprehensively interrogate metabolism. These tools have been successfully applied to the study of different phenotypes associated with cancer and diabetes where they have been shown to play a pivotal role in revealing, explaining and confirming major mechanistic insights.

CONCLUSIONS

CONCLUSIONS

Herein I list the main conclusions of this thesis, related to each objective:

Objective 1:

- Metabolic networks are an excellent framework to study the effects of protein expression on metabolism.
- Annotating metabolic reactions with protein entities is an appropriate strategy to integrate metabolomics and proteomics.

Objective 2:

- Chemical ionization with isobutane as a reagent gas in gas chromatography coupled to high-resolution mass-spectrometers is an ideal setup for stable isotope labelling experiments.

Objetive 3:

- Statistical testing is a recommended option when performing the unbiased detection of enriched isotopologues in stable isotope labelling metabolomics experiments.