UVIC UNIVERSITAT
DE VIC

Master of Science in Omics Data Analysis

Master Thesis

Integrating quantitative proteomics and metabolomics in a cellular model of diabetic retinopathy

by

Jordi Capellades Tomàs

Supervisor: Oscar Yanes Torrado, Yanes Lab, Universitat Rovira i Virgili

Co-supervisor: Jordi Planas Cuchi, Department of Systems Biology,

Universitat de Vic

Department of Systems Biology

University of Vic – Central University of Catalonia

September of 2014

Context

The use of nuclear magnetic resonance spectroscopy in biological matrices started in the mid-1980s. However, it was in 1999 when the concept termed *metabonomics* was first coined and was formally defined by Jeremy K Nicholson and colleagues as *the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification.* A little bit later, in 2001, Oliver Fiehn defined the term *metabolomics* as *a comprehensive and quantitative analysis of all metabolites in a system.* During this period of time, metabolomics slowly evolved into utilizing new platforms such as gas/liquid chromatography coupled to mass spectometry. Currently, metabolomics has proven useful in a lot of different purposes, from medical to environmental.

Nowadays metabolomics enables high throughput interrogation for low molecular mass organic endogenous metabolites of large tissues and biofluids sample sets with limited *a priori* knowledge of the metabolites of interest. However, owing to the enormous chemical diversity of the metabolites in addition to the wide range of physiochemical properties: dynamic range concentrations, pH, polarities, etc, none of the analytical platforms mentioned above is yet able to cope with the full metabolome coverage.

Two types of metabolomics approaches exist: targeted and untargeted metabolomics. The former, takes advantage of the modern high-throughput mass spectometry technology to select those ions that are previously chosen, based on prior knowledge or research interests; the latter, does not need foundations but tries to find a solution to the biological problem after statistical analysis. The main benefit of untargeted metabolomics is that it can lead into discovering new metabolic pathways or new roles for known molecules.

Any metabolomics experiment comprises several basic steps: (1) sample preparation, different biological matrices involve different sample preparation protocols which also will depend on the analytical platform used; (2) instrumental analysis, considering that each analytical platform has its own drawbacks and advantages, a metabolomics experiment implies the use of more than one platform, commonly nuclear magnetic resonance and a mass spectrometry; (3) data pre-processing and multivariate data analysis, multivariate data analysis algorithms are the most common tools to analyze metabolomics-derived data, among these algorithms principal component analysis (PCA), Partial Least Squares – Discriminant Analysis (PLS-DA), Orthogonal Projections to Latent Structures (O-PLS) are the most used ones, plus a previous pre-processing step is necessary involving peak alignment, peak picking, deconvolution and data dimensionality reduction; (4) metabolite identification, identification of putative markers is the step requiring most effort in spite of some recent advances in metabolite library annotations.

Metabolomics encompasses a top-down approach consisting on comprehensive and simultaneous systematic profiling of multiple metabolite levels and their systematic changes using high-throughput sample analysis with computer-assisted multivariate pattern-recognition techniques. Metabolomics has emerged as a complementary technology to other -omics disciplines, in particular genomics, transcriptomics and proteomics, which are concerned with the measurements of DNA, mRNA, proteins and their interactions. Unlike these disciplines metabolomics is in its early development stages and it is currently strongly driven by technological developments. Metabolomics has its own specific preprocessing steps and is, along with proteomics, the most close study of phenotype that allows quantification, commonly semiquatification, in omics sciences.

The paper, in which this work is based, is still being written. All the difficulties that have been overcome during the internship are described in Appendix A.2. The following text, except for some parts of *Methods* section, has been completely written by me using LaTeX. My role in the creation of this article was mainly data processing and writing scripts for data interpretation, also, my knowledge on biochemistry helped in the understanding of biological problem and providing reasonable mechanisms that could characterize it.

Metabolic diversification in human retinal pigment epithelium cells induced by hyperglycemic and hypoxic conditions

Miriam Navarro, Lucrece Matheron, Sara Samino, Nuria Canela, Maria Vinaixa, Miguel A Rodriguez, Jordi Capellades, Marta Garcia-Ramirez, Cristina Hernandez, Salvatore Cappadona, Rosa Ras, Antoni Beltran, Albert JR Heck, Rafael Simo, Shabaz Mohammed and Oscar Yanes

Abstract
Diabetic retinopathy is the leading cause of visual loss in individuals under the age of 55. Most investigations into the pathogenesis of diabetic retinopathy have been concentrated on the neural retina since this is where clinical lesions are manifested. Recently, however, various abnormalities in the structural and secretory functions of retinal pigment epithelium that are essential for neuroretina survival, have been found in diabetic retinopathy. In this context, here we study the effect of hyperglycemic and hypoxic conditions on the metabolism of a human retinal pigment epithelial cell line (ARPE-19) by integrating quantitative proteomics using tandem mass tagging (TMT), untargeted metabolomics using MS and NMR, and ^{13}C-glucose isotopic labeling for metabolic tracking. We observed a remarkable metabolic diversification under our simulated in vitro hyperglycemic conditions of diabetes, characterized increased flux through polyol pathways and inhibition of the Krebs cycle and oxidative phosphorylation. Importantly, under low oxygen supply RPE cells seem to consume rapidly glycogen storages and stimulate anaerobic glycolysis. Our results therefore pave the way to future scenarios involving new therapeutic strategies addressed to modulating RPE metabolic impairment, with the aim of regulating structural and secretory alterations of RPE. Finally, this study shows the importance of tackling biomedical problems by integrating metabolomic and proteomics results.

Contents

1. Introduction

Diabetes is a high prevalence condition, the number of people affected is sorrowfully growing. The World Health Organization estimated that it could reach the 360 million mark by 2030 [1]. Being a chronic condition with a great variety of complications, diabetic diseases are greatly studied all over the world. The complications are caused by a deregulation of the glucose uptake originated by the chronic hyperglycemia. This occurs in cell types such as retinal epithelium, renal glomerulus, and peripheral nerve cells [1].
There are four biochemical processes that are common in affected cells in response to the surrounding glucose concentration.

- The polyol pathway is based on a family of aldo-keto reductase enzymes that can use as substrates a wide variety of carbonyl compounds, i.e. sugars, and reduce these by NADPH to their respective sugar alcohols, known as polyols [2]. NADPH depletion has been hypothesized to cause an increase of intracellular reactive oxygen species (ROS), since it is required for regenerating reduced glutathione (GSH), an important detoxifier of ROS [2].

- Advanced glycation end products (AGEs) are formed by a non-enzymatic reaction of glucose and other glycating compounds derived from glucose. In diabetes, AGEs are found in high amounts in the extracellular matrix. Besides, intracellular production of AGE precursors can damage cells by three general mechanisms. First, intracellular proteins modified by AGEs have altered function. Second, AGE-modified extracellular matrix components interact abnormally with other matrix components and with matrix protein receptors that are expressed on the surface of cells. Finally, plasma proteins modified by AGE precursors bind to AGE receptors on cells such as macrophages, vascular endothelial cells, and vascular smooth muscle cells, causing inflammation or vascular damage [2].
- Protein kinases are a family of kinase enzymes that modify other proteins by chemically adding phosphate groups to them, known as phosphorylation. There are at least 11 isoforms of Protein kinases C (PKCs) widely distributed in mammalian tissues, they are able to phosphorylate various target proteins. The activity of the classic isoforms is dependent on both calcium ions (Ca^{2+}) and phosphatidylserine (PS) and is greatly enhanced by diacylglycerol (DAG). Persistent and excessive activation of several PKC isoforms operates as a third common pathway mediating tissue injury induced by diabetes-induced ROS [2].
- Hyperglycemia and insulin resistance-induced excess fatty acid oxidation also appear to contribute to the pathogenesis of diabetic complications by increasing the flux of fructose 6-phophate into the hexosamine pathway. In this pathway, fructose 6-phosphate is diverted from glycolysis to provide substrate for the rate-limiting enzyme of this pathway, ending in formation of UDP-N-Acetylglucosamine, which is used in postranslational modification on cytoplasmic and nuclear proteins [2].

All these mechanisms seem to result in the upregulation of a sole process, mitochondrial overproduction of ROS [2]. In diabetic cells, which are continously surrounded by a high intracellular glucose concentration, there is more glucose-derived pyruvate being oxidized in the Kreb's cycle or tricarboxylic acid cycle (TCA cycle), increasing the flux of electron donors (NADH and FADH2) into the electron transport chain. As a result, the voltage gradient across the mitochondrial membrane increases until a critical threshold is reached.

At this point, electron transfer inside complex III is blocked, causing the electrons to back up to coenzyme Q, which donates the electrons one at a time to molecular oxygen, thereby generating superoxide ions, i.e. ROS.

In 2012, an epidemiologic study found that there were approximately 93 million people suffering from diabetic retinopathy (DR), 17 million with proliferative DR (the advanced phase of DR), 21 million with diabetic macular edema (a condition in which retina is damaged due to swelling), and 28 million with vision-threatening diabetic retinopathy worldwide [3]; the overall prevalence was 34.6% for any DR. DR is a condition affecting the vascular epithelium of the retina, the retinal pigment epithelium (RPE) is an especialized epithelium lying in the interface between the neural retina and the choriocapillaris where it forms the outer blood-retinal barrier (BRB).

The main functions of the RPE are the following: transport of nutrients, ions, and water, absorption of light and protection against photooxidation, reisomerization of all-trans-retinal into 11-cis-retinal (a crucial molecule for vision), phagocytosis of shed photoreceptor membranes, and secretion of essential factors for the structural integrity of the retina [4].

The main consequence of DR is a decrease in retinal blood circulation caused by a loss of vascularization in response to the hyperglycemia stress. DR is characterized by pericyte loss followed by increased vascular permeability and progressive vascular occlusion due to high glucose concentration. Loss of pericytes results in empty, balloon-like spaces on the wall of the retinal capillary. Endothelial cells try to repair the damaged vessel by proliferating on the inner vessel wall [5].

The inner BRB consists of the basement membrane and the fusion of membranes between retinal endothelial cells forms tight junctional complexes to help stop the outward flow of circulating proteins. The BRB breakdown begins with the loss of tight junctions between adjacent microvascular endothelial cells. As barrier breakdown proceeds, the basement membrane of the capillaries thickens and the capillaries become rigid [6]. Therefore, the vascularization of retinal epithelium drops, leaving the cells in hypoxic conditions. Glucose metabolism also depends on the homeostasis of oxygen since it is necessary for oxidation in glycolysis and electron transport in the mitochondria, under oxygen-insufficient conditions, fermentation is favored.

Metabolomics is the systematic study of the unique chemical fingerprints that specific cellular processes leave behind. Metabolites are small molecules that are chemically transformed during metabolism and, as such, they provide a functional readout of cellular state. Unlike genes and proteins, whose function is subject to epigenetic regulation and post-translational modifications respectively, metabolites serve as direct signatures of biochemical activity and they are therefore easier to correlate with phenotype. Given its sensitivity, high-throughput and minimal sample requirements, untargeted metabolomics has wide applicability across a countless biological questions. Despite its relatively recent emergence as a global profiling technology, untargeted metabolomics has already increased the understanding of comprehensive cellular metabolism and been utilized to address a number of biomedical issues [7].

Although untargeted metabolomics can be performed by using either nuclear magnetic resonance (NMR) [8] or mass-spectometry (MS) [9] technologies, liquid chromatography coupled with mass spectometry (LC-MS) enables the detection of the most metabolites and has therefore been the technique of choice for global metabolite profiling efforts [7, 10]. Proteomics is the study of the protein content of a biological sample, it provides biologists the ability to monitor global protein expression and quantitative data on the molecular basis of cellular change [11]. Evolving from two-dimensional gel electrophoresis, the current central platform for proteomics is tandem mass spectrometry (MS/MS) but a number of other technologies, resources, and expertise are required to perform meaningful experiments; including protein biochemistry, genomics, and bioinformatics [12]. MS provides excellent means for quantitative proteomics whereby the most common and accurate quantitative approaches utilize stable isotopes. The tandem mass tag (TMT) approach consists in pairs, or more, of TMT-tagged peptides are chemically identical, like the isotope tags used in other methods, but unlike other isotope tags, the TMTs also have the same overall mass and comigrate in chromatographic separations and, thus, will act as more precise reciprocal internal standards, which leads to more accurate quantification [13]. This brings advantages for global analysis of protein samples, because this should allow more proteins to be identified in a single analysis in the same time as other techniques while using conventional instrumentation [13].

The objective of this study was to determine metabolic changes in retinal epithelium cells due to hyperglycemia or/and hypoxia in order to characterize diabetic retinopathy from a metabolic insight. Furthermore, it was intended to integrate both metabolomics and proteomic results for better understanding of the disease profile.

2. Methods

2.1 RPE Cell Culture

Materials

ARPE-19 is a spontaneously immortalized human RPE cell line obtained from the American Type Culture Collection (Manassas, VA). D-Glucose was from Sigma (Madrid, Spain. Whitley H35 Hypoxystation from Nirco (Madrid, Spain). LC/MS grade methanol (MeOH) and acetonitrile (ACN) and analytical grade chloroform ($CHCl_3$) were purchased from SDS (Peypin, France). Water was produced in an in-house Milli-Q purification system (Millipore, Molsheim, France). Formic acid, ammonium fluoride, N-methyl-N-trimethylsilyltrifluoroacetamide, methoxamine hydrochloride and pyridine were purchased from Sigma-Aldrich (Steinheim, Germany). Myristic-d27 acid and succinic acid-2,2,3,3-d4 where from Isotec Stable Isotopes (Miamisburg, U.S.A.). A set of 13 even saturated fatty acid methyl esthers (FAMEs) from C8:0 to C30:0 were acquired from Sigma-Aldrich, NuChekPrep (Elysian, U.S.A.) and Molport (Riga, Latvia). Deuterated water (D_2O) and 5-mm NMR tubes were purchased from Cortecnet (Viosins Le Bretonneux, France). DMEM/F-12 basal medium was purchased from Life Technologies. Sequencing grade modified trypsin V511A was purchased from Promega and Lys-C 125-05061 from Wako.

The reagents for the quantitative proteomics experiments were: the Complete Mini EDTA-free protease inhibitor and the PhosSTOP phosphatase inhibitor cocktails were from Roche (Almere, The Netherlands), the 6-plex TMT labeling kit was from Pierce (Rockford, Ilinois), and all other reagents were from Sigma (Steinheim, Germany).

Cell culturing conditions

Cells were cultured under standard conditions in DMEM/F12 (1:1 mixture of Dulbecco's modified Eagle's medium and Ham's F12), 10% fetal calf serum (FCS) and penicillin-streptomycin. ARPE-19 cells from passage 20-23 were used and the media was changed every 3 days. Cells grown in these conditions constitute a monolayer that retains the functionality, polarity and tight junction expression of the human RPE [14]. For our study, cells were seeded in Petri dishes (10 cm) at $0.4{\cdot}10^4$ cells/mL and maintained in culture for 21 days with 5.5 mM or 25 mM of D-Glucose at 37^{o}C under 5% (v/v) CO_2 in an incubator. During the last 24 hours cells were subjected to serum deprivation (1% FCS). Serum deprived media were prepared with 5.5 mM or 25 mM of D-Glucose, and cultured in normoxic or hypoxic (1% O_2) conditions. Each condition was run in triplicate.

2.2 Quantitative Proteomics

Cell lysis and protein digestion

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

TMT labeling

100 μg of each sample were desalted and concentrated using C18 solid phase extraction (Sep-Pak Vac C18 cartridge 1 cm^3/200 mg, Waters), dried in vacuum and reconstituted in 120 μL of 200 mM triethylammonium bicarbonate (Sigma). Labeling was performed with the 6-plex labeling kit according to the manufacturer's protocol. Briefly, each labeling was carried out for 1 h at room temperature and quenched with 8 μL of 5% hydroxylamine. The four channels were mixed, dried in vacuum and resuspended in 10% formic acid.

Strong cation exchange fractionation

Peptides were fractionated by strong cation exchange (SCX) using a Zorbax BioSCX-Series II column (0.8 mm x 50 mm, 3.5 µm), as described in [15]. Solvent A consisted of 0.05% formic acid in 20% acetonitrile, solvent B of 0.05% formic acid, 0.5 M NaCl in 20% acetonitrile. The gradient was 0 to 2% B in 0.01 min; 2 to 3% B in 8 min; 3 to 8% B in 6 min; 8 to 20% B in 14 min; 20 to 40% B in 10 min; 40 to 90% B in 10 min; 90%B for 6 min; 90 to 0% B in 6 min. Fractions were collected once a minute and each dried in vacuum and stored at -20°C.

Mass spectrometry

SCX fractions were analyzed on an Orbitrap Q-Exactive (Thermo Fisher Scientific) connected to an UHPLC Proxeon Easy-nLC 1000 (Thermo Scientific). Peptides were trapped on a double-fritted trap column (Dr. Maisch Reprosil C18, 3 µm, 2 cm x 100 µm) and separated on an analytical column (Agilent Zorbax SB-C18, 1.8 µm, 40 cm x 75 µm), as described previously [16]. Solvent A consisted of 0.1 M acetic acid, solvent B of 0.1 M acetic acid in 80% acetonitrile. Samples were loaded at a pressure of 800 bar with 100% solvent A. Peptides were separated by a 110 min gradient from 10% to 40% solvent B at a flow rate of 150 nL/min. Full scan MS spectra were acquired in the Orbitrap (350-1500 m/z, resolution 35000, AGC target 3e6, maximum injection time 250 ms). The 20 most intense precursors were selected for HCD fragmentation (isolation window 1.2 Da, resolution 17500, AGC target 5e4, maximum injection time 120 ms, first m/z 100, NCE 33%, dynamic exclusion 60 s).

2.3 Untargeted Metabolomics

Metabolites extraction method

The culture medium was removed from cells and the dishes were placed on top of dry ice [17]. Cells were scrapped immediately and metabolites extracted into the extraction solvent by adding 2 mL of a cold mixture of chloroform and methanol (2:1 v/v). The resulting suspension was bath-sonicated for 3 minutes, and 2 mL of cold water was added. Then, 1 mL of chloroform/methanol (2:1 v/v) was added to the samples and bath-sonicated for 3 minutes. Cell lysates were centrifuged (5000 x g, 15 min at 4°C) and the aqueous phase was carefully transferred into a new tube. The sample was frozen, lyophilized and stored at -80°C until further analysis.

NMR analysis

The hydrophilic extracts were reconstituted in 600 µL of D_2O containing 0.67 mM trisilylpropionic acid (TSP). Samples were then vortexed, and centrifuged for 15 min at 6000 x g and 4°C. Finally, redissolved samples were placed into 5 mm NMR tubes. ^{1}H and ^{13}C NMR spectra were recorded at 300°K on an Avance III 600 spectrometer (Bruker, Germany) operating at a proton frequency of 600.20 MHz using a 5 mm CPTCI triple resonance (^{1}H, ^{12}C, ^{31}P) gradient cryoprobe. One-dimensional ^{1}H pulse experiments were carried out using the nuclear Overhauser effect spectroscopy (NOESY) presaturation sequence to suppress the residual water peak. The acquired spectral width was 12 kHz (20 ppm), and a total of 256 transients were collected into 64 k data points for each ^{1}H spectrum. ^{13}C-NMR spectroscopy was performed under approximately fully relaxed conditions (repetition time 8 seconds) and broadband proton decoupling. A total of 1024 scans and 64000 data points with a spectral width of 36 KHz (240 ppm) were acquired for each ^{13}C spectrum. Exponential line broadening of 0.3 Hz was applied before Fourier transformation and frequency domain spectra were phased and baseline-corrected using TopSpin software (version 2.1, Bruker).

LC/MS analyses

Fractions of 100 µL of each redissolved sample in deuterated water were placed into HPLC vials after NMR analysis with no need for solvent exchange as previously reported [18]. LC/MS analyses were performed using an UHPLC system (1290 series, Agilent Technologies) coupled to a 6550 ESI-QTOF MS (Agilent Technologies) operated in positive (ESI+) or negative (ESI-) electrospray ionization mode. Vials containing extracted metabolites were kept at -20°C prior to LC/MS analysis. When the instrument was operated in positive ionization mode, metabolites were separated using an Acquity UPLC (HSS T3) C18 reverse phase (RP) column (2.1 x 150 mm, 1.8 µm) and the solvent system was A_1 = 0.1% formic acid in water and B_1 = 0.1% formic acid in acetonitrile. When the instrument was operated in negative ionization mode, metabolites were separated using an Acquity UPLC (BEH) C18 RP column (2.1 x 150 mm, 1.7 µm) and the solvent system was A_2 = 1 mM ammonium fluoride in water and B_2 = acetonitrile, as previously reported [10]. The linear gradient elution started at 100% A (time 0-2 min) and finished at 100% B (10-15 min). The injection volume was 5 µL. ESI conditions: gas temperature, 150°C; drying gas, 13 L·min^{-1}; nebulizer, 35 psig; fragmentor, 400 V; and skimmer, 65 V. The instrument was set to acquire over the m/z range 100–1500 in full-scan mode with an acquisition rate of 4 spectra/sec. MS/MS was performed in targeted mode, and the instrument was set to acquire over the m/z range 50–1000, with a default isolation width (the width half-maximum of the quadrupole mass bandpass used during MS/MS precursor isolation) of 4 m/z. The collision energy was fixed at 20 V.

GC/MS analyses

Redissolved samples in deuterated water were lyophilized, dissolved in 50 µL of methoxyamine hydrochloride in pyridine (30 mg/mL) and incubated with agitation during 1 hour at 65°C. Trimethylsililation was done by adding 30 µl of N-methyl-N-trimethylsilyltrifluoroacetamide previously spiked with the FAMEs mix as internal standard. The samples were then shacked for 10 min and kept for 1 hour at room temperature. Derivatised samples were analysed in a 7890A Series gas chromatograph coupled to a 7200 GCqTOF MS (Agilent Technologies, Santa Clara, U.S.A.). Chromatographic column was a J& W Scientific HP5-MS (30 m x 0.25 mm i.d., 0.25

μm film) (Agilent Technologies). A volume of 1 μL of sample was automatically injected into a split/splitless inlet, which was kept at a temperature of 250°C. Helium was used as a carrier gas, at a flow rate of 1 mL/min in constant flow mode. The oven program was set at an initial temperature of 70°C for 1 min, then increased to 325°C at a rate of 10°C/min and held at 325°C for 9.5 min. Ionization was done by electronic impact, with an electron energy of 70 eV and an emission intensity of 35 μA. Source temperature was of 230°C. Mass spectra were recorded after a solvent delay of 6 minutes, after which the analyzer acquired in full-scan MS mode at a rate of 5 scan/sec, acquiring a mass range of 35-700 m/z.

2.4 Data Analysis

Proteomics Data analysis

Raw data was analyzed with MaxQuant(version 1.3.0.5) [19]. MS/MS peak lists were generated and searched with Andromeda against the Swissprot human database. Trypsin/P was chosen as an enzyme, with a maximum of 2 missed cleavages. Methionine oxidation was set as variable modification. Cysteine carbamidomethylation as fixed modifications, TMT6plex (Lys) and TMT6plex (N-term) as the reporter ion quantification method. The database search was performed with a precursor tolerance of 6 ppm for the main search (20 ppm for the first search) and a fragment mass tolerance of 0.05 Da. Match between run was enabled with a time window of 2 min. Peptide and protein FDR were set at 1%, and peptide score threshold at 60. The quantification and statistical processing was performed in Perseus (version 1.3.8.1) [20]. Proteins were grouped and reporter ion intensities were calculated for each of the TMT channels. Ratios were calculated and normalized on median. A significance B test was performed to determine significantly regulated proteins, where truncation was performed using p-values, with a threshold value of 0.05. A gene ontology (GO) [21] enrichment analysis of the resulting proteins was performed using on AmiGO [22].

NMR data analysis

^{1}H and ^{13}C NMR spectra were referenced to the chemical shift of TSP signal at 0.0 ppm. References of pure compounds from the metabolic profiling AMIX spectra database (Bruker), HMDB [23] and Chenomx databases were used for metabolite identification. In addition, we assigned metabolites by ^{1}H–^{1}H homonuclear correlation (COSY and TOCSY) and ^{1}H–^{13}C heteronuclear (HSQC) 2D NMR experiments, and by correlation with pure compounds run in-house. After baseline correction, specific NMR regions identified in the spectra were integrated using the AMIX 3.9 software package. Data processing, data analysis, and statistical calculations were performed in R 3.1.

LC/MS and GC/MS data analysis

LC/MS (ESI+ and ESI- mode) and GC/MS data were processed using the XCMS R package [24] to detect and align features. A feature is defined as a molecular entity with a unique m/z and a specific retention time (mzRT). XCMS analysis of these data provided a matrix containing the retention time, m/z value, and integrated peak area of each feature for every ARPE-19 sample. GC/MS data was normalized to the internal standard hexacosanoic (one of the FAMEs with the lowest coeficient of variance (CV)) was also performed. Quality control samples (QCs) consisting of pooled ARPE-19 samples from each four conditions were used in LC/MS and GC/MS analyses. QCs were injected at the beginning and periodically every 5 samples. Furthermore, samples entering the study were entirely randomized to reduce systematic error associated with instrumental drift. QCs were always projected in a PCA model together with the samples under study to verify that technical issues do not mask biological information. The performance of the analytical platform for each detected mzRT feature in ARPE-19 samples was assessed by calculating the relative standard deviation of these features on pooled samples (CVQC) according to Vinaixa et al. [9].
ARPE-19 samples were compared using the integrated peak area of each feature via a paired t-test and assigning a fold value to indicate the level of differential regulation due hypoxic and/or hyperglycemic conditions. Differentially regulated metabolites that were statistically significant after false discovery rate adjustment ($p<0.05$) between physiological-like and any of the pathological-like conditions detected by LC/MS were characterized by MS/MS. Differentially regulated metabolites ($p<0.05$) between detected by GC/MS were identified using the NIST and Fiehn mass spectral libraries. In addition, the retention time of pure standards were confirmed. Data pre-processing, data analysis, and statistical calculations were performed in R 3.1.

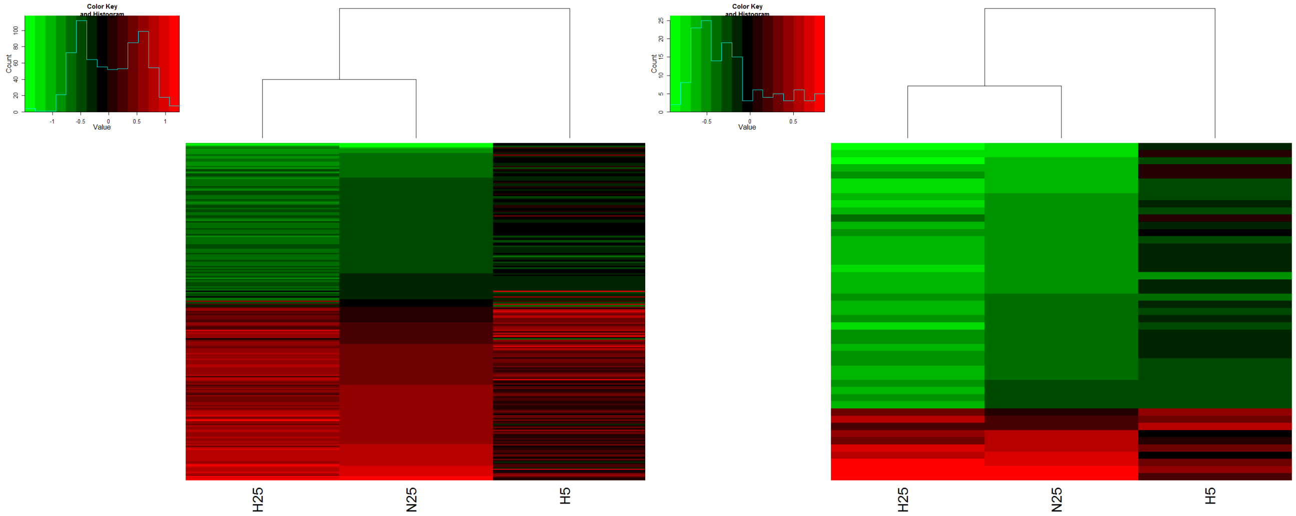
2.5 Data integration

Enzymatic proteins

In order to maximize the benefit from the quantitative proteomics results from a metabolic point of view, those proteins that have a known enzymatic function were filtered and thus could be mapped to Kyoto Encyclopedia of Genes and Genomes (KEGG) identifiers [25]. To do so, a complete list of the Reviewed Uniprot entries was obtained and using a Python 2.7 in-house script (see Appendix A.4), the European Bioinformatics Institute (EBI) Gene Expression Atlas API was interrogated to retrieve those Uniprot accession codes with an associated enzymatic function.

Pathway mapping

Pathwview R package [26] allows mapping color scale levels of expression of both metabolites and proteins in a KEGG pathway map. The resulting significant differentially expressed metabolites were mapped manually into KEGG compounds identifiers, which together with proteins, were used as the mapping input. A hypergeometric test was used to assess which pathways were significantly enriched.



(a) Heatmap for fold changes of 3260 proteins **(b)** Heatmap for fold changes of the 47 significant enzymes

Figure 3.1. Heatmaps from proteomics results sorted using N25 as reference

3. Results and Discussion

In this research, four sample models were designed: Normoglycemic (5.5 mM Glucose) and normoxic (5% O_2) (N5), resembling healthy status; high-glucose (25 mM Glucose) and normoxic (N25), hyperglycemia or early diabetes stage; normal-glucose and hypoxic (1% O_2) (H5), used to check hypoxia effects on ARPE-19 cells; finally, high-glucose and hypoxia (H25), simulating a proliferative DR condition.

3.1 Proteomics

Proteomics procedure was able to detect and identify a total of 5419 different proteins, of which 3260 were detected in all four sample conditions. After the significance B test, 233 proteins returned as signficant (FDR adjusted $p\text{-value}<0.05$) which were then categorized in up- and down-expressed for each comparison versus N5 condition using a $\log_2$ value transformation. Generally, the heatmap of the 3260 proteins shows that hyperglycemic conditions have a noticeable similarity in both upregulated and downregulated elements (see Figure 3.1). This was reinforced when filtering by enzymatic function, those significant proteins with an enzymatic function were 47, they also proved that H25 and N25 have a comparable enzymatic expression profile and mainly downregulated (green), rather than in H5 where the fold changes indicate that the differences in protein levels in comparison to N5 are much lower.

Using AmiGO tool, a gene ontology term enrichment was performed for each comparison versus N5. H5 returned the lowest number of differentially expressed proteins, as a consequence, the ontologies returned were few and poorly informative, which could be due to the fact that ARPE-19 cells are an immortalized cell-line that could not be thoroughly affected by hypoxia itself, or solely, but uncertainly, hypoxia does not cause important effects on protein expression. Yet, both hyperglycemic (N25 and H25) conditions have an increase of expression in proteins associated with cell development (GO:0048731) and physiological (GO:0032501) processes. As a response to the high glucose levels, there could be an increase in cell adhesion (GO:0005576) and activation of vesicle motion (GO:0031982), possibly to prevent the cell membrane from disrupting by the osmotic pressure. Generally, both high-glucose conditions suffered a decrease in mithocondrial protein expression which would cause a disruption of mithocondrial oxidative metabolism, probably driven by ROS toxicity. Interestingly, H25 had some specific metabolic effects, underrepresenting ontologies such as Acyl-CoA metabolism (GO:0006637), a process related to lipid metabolism.

Briefly, from a protein-level, in high-glucose conditions, independently of oxygen levels, ARPE-19 cells seem to respond by reducing cellular respiration capability and mitochondrial processes.

3.2 Metabolomics

Due to the wide range of chemical properties of cell metabolites, usage of different platforms is required if a good untargeted metabolomics coverage is desired. Here, three different devices were used; liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry (LC ESI-qTOF/MS) in positive and negative ionization mode, gas chromatography coupled to electron ionization (GC-EI-qTOF/MS), and NMR [10].

The objective in the untargeted metabolomics method was relatively quantifying either features in GC/MS and LC/MS following the identification of those which are significantly different between the glucose and oxygen concentration conditions. Alternatively, in NMR, the areas of the identified spectral regions are used as input for statistical testing.

Given that LC/MS-based untargeted metabolomics returns a great amount of features, in order reduce the multidimensionality, a Principal Component Analysis (PCA) was performed.

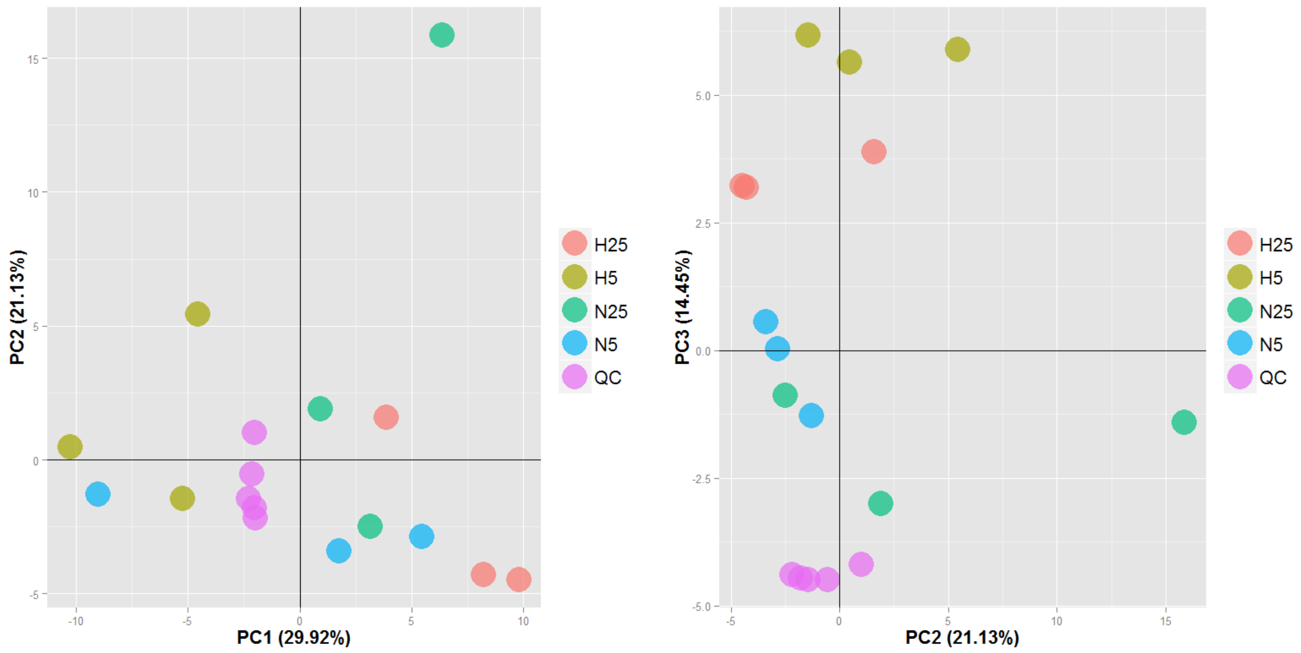


Figure 3.2. PCA plots obtained from negative phase LC-MS results after XCMS pre-processing

PCA involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for the remaining variability and so on. In our case, the proportion of variance explained by the first two components did not return any specific cluster of sample conditions, meaning that the difference between groups was not explained by the maximum variance, possibly due to the limited number of replicates or instrumental variability (Figure 3.2). On the other hand, using the third component of the PCA, having 64.11% of variation accumulated, it was possible to distinguish between the two levels of glucose concentration (Figure 3.2).

In total, 18 metabolites and 10 different lipid families were confidently identified (see Table A.1.1). Lactate and acetate, both anaerobic metabolism indicators, were significantly increased in H5 but not in H25, this may indicate that this metabolic strategy may be regulated not only based on oxygen but also on the glucose supply. Succinate was also significantly increased in H5, it has been theorized to be an end product of anaerobic metabolism of glutamate and aspartate [27], this event correlates with the former anaerobic indicators.

In general, the aminoacids detected had the same pattern of abundance ratios among the conditions, they were found significantly increased in H5. An increase in intracellular amino acids under hypoxia may be attributed to several possibilities; an increase in amino acid uptake and/or a decrease in protein synthesis, a phenomenon which is considered to be part of the metabolic adaptation to hypoxia, where ATP-consuming reactions like protein synthesis, are dramatically decreased; another possibility, which is not mutually exclusive, is that during hypoxia cells undergo protein catabolism in order to provide the cells with metabolic support and increase autophagy was observed under hypoxic conditions [28]. Succinate levels, also a product of aminoacid degration, support the latter.

Glycogen, the glucose storage form in animals, was, increased in N25, but not in H25. In both hypoxic conditions, the glycogen levels had a tendency to decrease, possibly in order to increase glucose mobilization into anaerobic metabolism. Besides, pyridoxal and pyridoxine, two forms of vitamin B6, were increased in H5. They origin pyridoxal-phosphate (PLP) which is a key cofactor for glycogen phosphorylase that catalyzes the rate-limiting step in glycogenolysis [29].

A high variety of features were found to be differentially abundant and positively identified in the lipid fraction in both high-glucose samples. In N25, glycerides, except for monoglycerides, lysophospholipids and phospholipids were upregulated. In addition, these effects were enhanced by hypoxia in H25 comparison (see Table A.1).

Polyol pathway which is known to be increased in cells affected by DR produces sorbitol. This metabolite was found to be significantly increased in N25 condition, which resembles early diabetic phase. The increase in intracellular osmolality, due to shunting of glucose into the polyol pathway and the consequent sorbitol accumulation, may lead to compensatory depletion of the endoneurial osmolytes taurine and myo-inositol in order to maintain osmotic balance [30], though, in our results only taurine seems to follow this regulation since myo-inositol is significantly increased in H5. Furthermore, nicotinate D-ribonucleoside, which is a central intermediate in nicotinamide metabolism that is closely related to redox power pool, was significantly increased in N25, possibly to generate cofactors NADH and NADPH and compensate for their consumption in the polyol pathway.

3.3 Integration

Using Pathview package both the significant proteins and metabolites were mapped for each comparison. In the case of metabolites, the compound identities (KEGGID) were manually obtained from the database. Alternatively, Uniprot accessions are automatically annotated by the Pathview package. Those comparisons in which they, protein or metabolites, were not significant were transformed into 0, as a way to indicate that the $\log_2$ ratio it is not meaningful. Afterwards, each KEGG pathway enriched according to the hypergeometric statistic was manually interpreted. A total of 35 pathways returned as significantly enriched, containing pathways such as: Glycolysis / Gluconeogenesis, Citrate cycle (TCA cycle)3.3, Glutathione metabolism, . . . (see Table A.1.2)

4. Conclusions

In the present study we have examined the effect of high glucose concentration with or without hypoxia in human RPE cells in culture by integrating differential protein expression and quantitative analysis of metabolite pools by untargeted metabolomics using MS and NMR. This unique approach allowed to connect metabolic processes and provide an improved understanding of the mechanisms regulating metabolic functions in RPE cells due to pathological hyperglycemic and/or hypoxic conditions. Interestingly, metabolomics and proteomics results paralleled with each other in all conditions, confirming that teaming proteomics and metabolomics is a great strategy for studying any biological problem.

In this research, the most changes were caused by high-glucose levels, though the expected metabolic changes in H5 were observed: increased lactate and acetate production, typical from hypoxia adaptation; also, increased aminoacid production maybe caused by protein catabolism in order to fill the central carbon pools to fulfill the cell needs; and some clues indicating that glycogen is degrated in this physiological condition.

Moreover, nicotinate D-ribonucleoside was found higher in N25 which would indicate a favorable metabolic flux from glucose into the pentose phosphate pathway and into nicotinate and nicotinamide pathways, possibly to generate NADH and NADPH cofactors and compensate for their consumption in the polyol pathway (confirmed by increased sorbitol production in N25) and to prevent the early stages of glycolysis from saturating. The consumption of NADPH by aldose reductase, the first and rate-limiting enzyme in the polyol pathway, results in less cofactor being available for glutathione reductase, which is critical for the maintenance of the intracellular pool of reduced glutathione, thus, ROS could affect the mitochondria (see Figure 4.1).

Additionally, more evident changes were perceived in high-glucose conditions, which resemble better the DR cell status. Mitochondria are the central metabolism machinery in any eukaryotic cell. In this work, there were both protein and metabolic proofs that high-glucose, with or without hypoxia, decreases mitochondria protein concentration. For instance, GO enrichment analysis using AmiGO tool found that processes and cellular components linked to this organelle were decreased in N25 and H25; besides, metabolomics demonstrated lower levels of TCA metabolites and lipid families, except monoglycerides.

RPE constitutes the outer BRB and is essential for neuroretina survival, and consequently, for visual function. Osmotic pressure becomes dangerous for cell survival in high-glucose conditions, GO enrichment processes found vesicle, actin filaments and extracellular regions upregulated in both N25 and H25, possibly to avoid membrane destruction. This theory is supported by the decreased taurine metabolite, which is known to be a compensatory mechanism.

Omics integration is a focused topic in systems biology research. In our study, Pathview was a simple and direct ap-

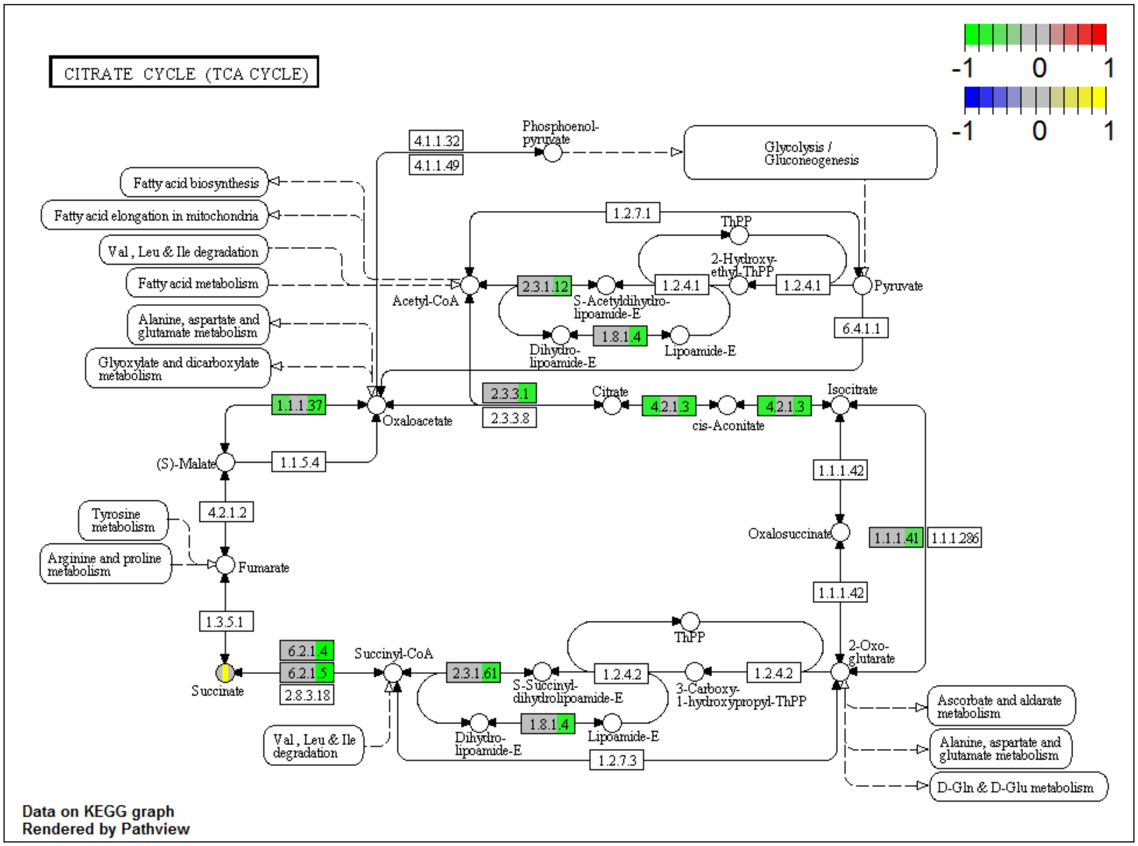


Figure 3.3. Example result of Pathview result using significant enzymatic proteins and identified metabolites as input for TCA Cycle KEGG Pathway. Top right shows a legend for values. Each metabolite (circles) and protein (rectangles) is divided in the number of comparisons; from left to right: N25, H5, H25.

proach to accomplish our needs of incorporating proteomics and metabolomics quantitative data for easier understanding and facilitate interpretation. Even though some protein were incorrectly mapped since the tool uses the enzymatic code instead the protein specificity, this happened in lysosomal alpha-glucosidase which according to the enzymatic activity is correctly mapped but not when considering that alpha-glucosidases, commonly found in plants and bacteria, are only found in animal lysosomes where they play a role for glucose-chain breakdown. Consequently, omics integration requires deep understanding of biochemistry or molecular biology, in order to avoid misinterpretation.

All things considered, hyperglycemia seems to be the primary root of DR pathology, rather the subsequent loss of vascularization, leading into the production of ROS that have been identified as cell damage and inflammation inducers that may as well cause the aged macular edema.

Acknowledgments

I am grateful to all the members of Yanes Lab for showing me the tips and tricks from the great metabolomics world. Also, I would like to thank the University of Vic and the Masters' professors for creating this Master of science which was highly enjoyable and exciting.

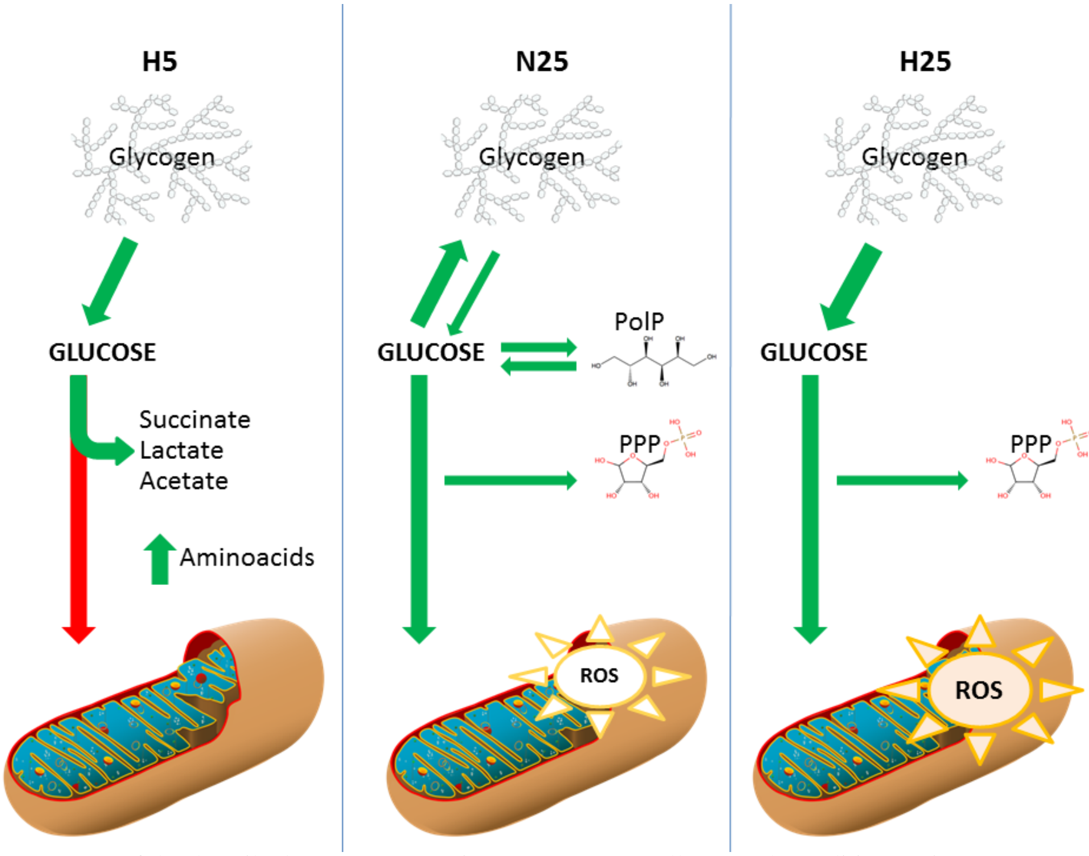


Figure 4.1. Summary of the described processes ocurring on ARPE-19 cells on each condition. H5 promotes anaerobic metabolism and possibly induces proteolysis; both high-glucose conditions, suffer ROS stress (higher in H25) and try to release pression from glycolisis pathway by activating the pentose phosphate pathway (PPP), yet, only N25 is able to mantain glycogen levels and activate polyol pathway (PolP) to even reduce the glycolisis products influx into the mitochondria.

References

[1] S. Wild, G. Roglic, A. Green, R. Sicree, and H. King, "Global prevalence of diabetes: Estimates for the year 2000 and projections for 2030," *Diabetes Care*, 2004.

[2] F. Giacco and M. Brownlee, "Oxidative stress and diabetic complications," *Circulation Research*, 2010.

[3] Joanne W Yau et al., "Global prevalence and major risk factors of diabetic retinopathy," *Diabetes Care*, 2012.

[4] R. Simó, M. Villarroel, L. Corraliza, C. Hernández, and M. Garcia-Ramírez, "The retinal pigment epithelium: Something more than a constituent of the blood-retinal barrier—implications for the pathogenesis of diabetic retinopathy," *Journal of Biomedicine and Biotechnology*, 2010.

[5] C.-J. Chiu and A. Taylor, "Dietary hyperglycemia, glycemic index and metabolic retinal diseases," *Progress in Retinal and Eye Research*, 2011.

[6] M. Garcia-Ramírez, M. Villarroel, L. Corraliza, C. Hernández, and R. Simó, "Measuring permeability in human retinal epithelial cells (arpe-19): implications for the study of diabetic retinopathy," *Methods in Molecular Biology*, 2011.

[7] G. J. Patti, O. Yanes, and G. Siuzdak, "Metabolomics: the apogee of the omic triology," *Nature Reviews Molecular Cell Biology*, 2012.

[8] D. S. Wishart, "Quantitative metabolomics using nmr," *Trends in Analytical Chemistry*, 2008.

[9] M. Vinaixa, S. Samino, I. Saez, J. Duran, J. J. Guinovart, and O. Yanes, "A guideline to univariate statistical analysis for lc/ms-based untargeted metabolomics-derived data," *Metabolites*, 2012.

[10] O. Yanes, R. Tautenhahn, G. J. Patti, and G. Siuzdak, "Expanding coverage of the metabolome for global metabolite profiling," *Analytical Chemistry*, 2011.

[11] S.-E. Ong, L. J. Foster, and M. Mann, "Mass spectrometric-based approaches in quantitative proteomics," *Methods*, 2003.

[12] I. A. Brewis and P. Brennan, "Proteomics technologies for the global identification and quantification of proteins," *Advances in Protein Chemistry and Structural Biology*, 2010.

[13] A. Thompson, J. Schäfer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, T. Neumann, and C. Hamon, "Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by ms/ms," *Analytical Chemistry*, 2003.

[14] M. Garcia-Ramírez, M. Villarroel, L. Corraliza, C. Hernández, and R. Simó, "Measuring permeability in human retinal epithelial cells (arpe-19): Implications for the study of diabetic retinopathy," *Methods in Molecular Biology*, 2011.

[15] J. Munoz, T. Y. Low, Y. J. Kok, A. Chin, C. K. Frese, V. Ding, A. Choo, and A. J. Heck, "The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells," *Molecular Systems Biology*, 2011.

[16] A. Cristobal, M. L. Hennrich, P. Giansanti, S. S. Goerdayal, A. J. Heck, and S. Mohammed, "In-house construction of a uhplc system enabling the identification of over 4000 protein groups in a single analysis," *The Analyst*, 2012.

[17] M. Yuan, S. B. Breitkopf, X. Yang, and J. M. Asara, "A positive/negative ion–switching, targeted mass spectrometry–based metabolomics platform for bodily fluids, cells, and fresh and fixed tissue," *Nature protocols*, 2012.

[18] A. Beltran, M. Suarez, M. A. Rodriguez, M. Vinaixa, S. Samino, L. Arola, X. Correig, and O. Yanes, "Assessment of compatibility between extraction methods for nmr- and lc/ms-based metabolomics," *Analytical chemistry*, 2012.

[19] J. Cox and M. Mann, "Measuring permeability in human retinal epithelial cells (arpe-19): Implications for the study of diabetic retinopathy," *Methods in Molecular Biology*, 2011.

[20] Jürgen Cox et al., "Perseus software." `http://www.perseus-framework.org`, 2011.

[21] Michael Ashburner et al., "Gene ontology: tool for the unification of biology. the gene ontology consortium," *Nature Genetics*, 2000.

[22] Seth Carbon et al., "Amigo: online access to ontology and annotation data," *Bioinformatics*, 2009.

[23] David S. Wishart et al., "Hmdb 3.0–the human metabolome database in 2013," *Nucleic Acids Research*, 2013.

[24] m. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment and identification, "Smith, colin a and want, elizabeth j and o'maille, grace and abagyan, ruben and siuzdak, gary," *Analytical chemistry*, 2006.

[25] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, 2000.

[26] L. Wiejun and C. Brouwer, "Pathview: an r/bioconductor package for pathway-based data integration and visualization," *Bioinformatics*, 2013.

[27] H. Taegtmeyer, "Metabolic responses to cardiac hypoxia. increased production of succinate by rabbit papillary muscles," *Circulation research*, 1978.

[28] Christian Freezza et al., "Metabolic profiling of hypoxic cells revealed a catabolic signature required for cell survival," *PLoS One*, 2011.

[29] Livanova Natasha B et al., "Pyridoxal 5'-phosphate as a catalytic and conformational cofactor of muscle glycogen phosphorylase b," *Biochemistry (Moscow)*, 2002.

[30] Martin J Stevens et al., "Aldose reductase gene expression and osmotic dysregulation in cultured human retinal pigment epithelial cells," *American Journal of Physiology*, 1993.

1. Appendix

A.1 Supplementary Material

A.1.1 Supplementary Table

Table A.1. Table of identified metabolites in both extracted fractions

KEGGID	METABOLITE NAME	LOG$_2$(N25.FC)	LOG$_2$(H5.FC)	LOG$_2$(H25.FC)
C00018	Pyridoxal phosphate	0.07	**1.22**	0.69
C00025	Glutamate	0.01	**1.60**	-0.14
C00033	Acetate	0.23	**0.88**	-0.67
C00041	Alanine	-0.23	**0.97**	-0.90
C00042	Succinate	0.12	**1.36**	-0.11
C00079	Phenylalanine	**-0.54**	**0.90**	-0.33
C00082	Tyrosine	-0.33	**1.52**	-0.15
C00123	Leucine	-0.19	**0.91**	-0.85
C00137	myo-Inositol	-0.01	**2.00**	-3.21
C00157	Phosphatidylcholine	**-1.87**	*	**-2.73**
C00165	Diacylglycerol	**-1.27**	*	**-2.01**
C00182	Glycogen	**3.02**	-1.17	-9.41
C00183	Valine	-0.18	**0.95**	-0.79
C00186	Lactate	0.37	**1.44**	-0.05
C00245	Taurine	**-0.84**	0.66	**-1.49**
C00314	Pyridoxine	0.55	**2.50**	1.25
C00350	Phosphatidylethanolamine	*	*	**-3.41**
C00407	Isoleucine	-0.21	**1.10**	-0.60
C00422	Triacylglycerol	**0.58**	*	**-2.03**
C00681	Lysophosphatidic acid	**-0.69**	*	**-2.49**
C00794	Sorbitol	**1.74**	0.15	0.07
C01885	1-Acylglycerol	**1.43**	*	**4.82**
C04230	1-Acyl-sn-glycero-3-phosphocholine	**-2.11**	*	**-2.70**
C05841	Nicotinate.d.ribonucleoside	**1.95**	-1.18	-0.85
C05973	2-Acyl-sn-glycero-3-phosphoethanolamine	*	*	**-1.88**
C05974	2-Acyl-sn-glycero-3-phosphoserine	**-0.90**	*	**-1.91**
C02737	Phosphatidylserine	**-1.61**	*	**-4.07**
C00550	Sphingomyelin	**2.10**	*	*

Bold values returned significant after t-test. (*) indicates that the values could not be obtained because the metabolite was not detected in the given condition, this especially happened in the organic extraction phase which contains lipids that are more difficult to identify in LC/MS.

A.1.2 Supplementary Table 2

Table A.2. Table of identified metabolites and fold changes (FC) in both extracted fractions

KEGG Pathway name	Hypergeometric test p-value
Glycolysis / Gluconeogenesis	$7.6e^{-4}$
Citrate cycle (TCA cycle)	$4.1e^{-15}$
Fructose and mannose metabolism	0.02
Galactose metabolism	0.02
Fatty acid elongation	$3.1e^{-9}$
Fatty acid degradation	$1.1e^{-23}$
Synthesis and degradation of ketone bodies	$7.7e^{-4}$
Ubiquinone and other terpenoid-quinone biosynthesis	0.02
Purine metabolism	$5.3e^{-5}$
Pyrimidine metabolism	$3.4e^{-10}$
Alanine, aspartate and glutamate metabolism	$8.8e^{-3}$
Cysteine and methionine metabolism	$2.3e^{-5}$
Valine, leucine and isoleucine degradation	$7.1e^{-17}$
Valine, leucine and isoleucine biosynthesis	$9.5e^{-3}$
Lysine degradation	0.01
Phenylalanine metabolism	0.04
Phenylalanine, tyrosine and tryptophan biosynthesis	$7.0e^{-5}$
beta-Alanine metabolism	0.04
Taurine and hypotaurine metabolism	$7.2e^{-5}$
Selenocompound metabolism	0.01
Cyanoamino acid metabolism	0.01
Glutathione metabolism	$3.6e^{-3}$
Starch and sucrose metabolism	$1.7e^{-3}$
Amino sugar and nucleotide sugar metabolism	0.04
Glycerolipid metabolism	0.01
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	0.04
Glycerophospholipid metabolism	$1.8e^{-7}$
Pyruvate metabolism	$4.7e^{-6}$
Glyoxylate and dicarboxylate metabolism	$1.2e^{-7}$
Propanoate metabolism	$1.4e^{-10}$
Butanoate metabolism	$1.2e^{-3}$
Nicotinate and nicotinamide metabolism	$3.9e^{-4}$
Sulfur metabolism	$2.1e^{-3}$
Aminoacyl-tRNA biosynthesis	$2.6e^{-6}$

A.1.3 Script example for pre-identification data analysis

```
library(xcms)
library(reshape2)
library(ggplot2)
library(outliers)
library(gridExtra)
library(mvoutlier)
lf <- list.files(pattern=’[.]mzXML’, recursive = T) # Get files
xset<-xcmsSet(files=lf,method="centWave",ppm=15,peakwidth=c(5,20),noise=1000)
xset<-group(xset)
xset2<-retcor(xset,method="obiwarp",profStep=0.1)
xset2<-group(xset2, mzwid=0.015,minfrac=0.5,bw =5)
xset3<-fillPeaks(xset2)
nsamples <- table(sampclass(xset3))

X1 <- groupval(xset3, value = "maxo")
classv <- xset3@phenoData$class
classv <- as.factor(gsub("CELL_Glc1._","",classv))

meanintensities <- apply(X1, 1, function(x) tapply(x, classv,mean))

M2  <- t(apply(meanintensities,1,sort,decreasing=T))
M2 <- melt(t(log10(M2)))
names(M2) <- c("Features","Group","Intensity")

ggplot(data=M2,aes(x=Features, y=Intensity, colour=Group))+
  geom_line()

idx_intensity <- which(apply(meanintensities, 2, function(x) any(x > 10ˆ3.5)) == TRUE)
(length(idx_intensity)/ncol(meanintensities)) * 100

cvcl <- rep(c("Sample", "QC"), times = c(2*3*4, 5))
CV <- t(apply(X1, 1, function(y) tapply(y, cvcl, function(x) (100 * (sd(x)/mean(x))) ) ))
idx_cv <- which(CV[, "Sample"] > CV[, "QC"])
(length(idx_cv)/nrow(X1)) * 100

Ib <- intersect(idx_intensity, idx_cv)
(length(Ib)/nrow(X1))*100

D <- data.frame(t(X1))
colnames(D) <- as.character(1:ncol(D))

D  <- D[,Ib]

D <- D[grep("_Glc13_",rownames(D),invert=T),]

Dnorm <- apply(D, 2, function(x) (x/max(x)) )

pca <- prcomp(Dnorm, scale = F)

library(car)
scatter3d(x=pca$x[,"PC1"],y=pca$x[,"PC2"],z=pca$x[,"PC3"],xlab="PC1",ylab="PC2",zlab="PC3",
          point.col=rep(1:5,times=unname(nsamples)),grid.lines=3)

summary(pca)$importance[, 1:4]

classv <- as.factor(c(rep("H5",times=3),rep("N5",times=3),rep("H25",times=3),rep("N25",times=3),
  rep("QC",times=5)))

scores <- data.frame(pca$x[, c("PC1", "PC2")], classv)

g1 <- ggplot(data = scores, aes(x = PC1, y = PC2, colour = classv)) +
  geom_point(alpha = I(0.7), size = 10) +
  geom_vline(xintercept = 0) + geom_hline(yintercept = 0)+
  ylab(paste("PC2"," (",round((summary(pca)$importance[2, 2]*100),digits=2),"%)",sep=""))+
  xlab(paste("PC1"," (",round((summary(pca)$importance[2, 1]*100),digits=2),"%)",sep=""))+
  theme(legend.title=element_blank(),legend.text = element_text(size = 16),
        axis.title = element_text(face="bold",size=16),
        axis.title.y = element_text(face="bold",size=16))
```

```
scores <- data.frame(pca$x[, c("PC2", "PC3")], classv)
g2 <- ggplot(data = scores, aes(x = PC2, y = PC3, colour = classv)) +
  geom_point(alpha = I(0.7), size = 10) +
  geom_vline(xintercept = 0) + geom_hline(yintercept = 0)+
  ylab(paste("PC3"," (",round((summary(pca)$importance[2, 3]*100),digits=2),"%)",sep=""))+
  xlab(paste("PC2"," (",round((summary(pca)$importance[2, 2]*100),digits=2),"%)",sep=""))+
  theme(legend.title=element_blank(),legend.text = element_text(size = 16),
        axis.title.x = element_text(face="bold",size=16),axis.title.y = element_text(face="bold",size=16))

grid.arrange(g1,g2,ncol=2)

lier <- outlier(Dnorm, opposite = FALSE, logical = TRUE)
lier[lier == TRUE] <- 1; lier[lier == FALSE] <- 0
sort(rowSums(lier),decreasing=T)

D2 <- D[grep("QC",rownames(D),invert=T),]
classv2 <- as.factor(as.character(classv)[grep("QC",as.character(classv),invert=T)])

pvalues <- apply(D2,2,function (x){
  sapply(levels(classv2)[-2],function(y){
    a <- try(t.test(x[which(classv2==y)],
      x[which(classv2==levels(classv2)[2])],var.equal=F)$p.value)
    if(is(a,"try--error")){a <- 1}
    return(a)
})
})

pvalues <- as.data.frame(t(pvalues))

fc.test <- sapply(levels(classv2)[-2],function(y){
  apply(D2,2,function (x){
    case <- mean(x[which(classv2==y)])
    control <- mean(x[which(classv2==levels(classv2)[2])])
    FC <- case/control;
    FC2 <- -control/case
    FC[FC<1] <- FC2[FC<1]
    return(FC)
  })
})

fc.test <- as.data.frame(fc.test)

positions <- sapply(1:nrow(fc.test),function(x){
  a <- c(any(pvalues [x,]<0.05),any(abs(fc.test[x,])>1.5))
  a <- all(a ==TRUE)
})

D3 <- D2[,which(positions==TRUE)]

featureinfo <- xset3@groups[as.numeric(colnames(D3)),c("mzmed","rtmed")]
rownames(featureinfo) <- colnames(D3)
```

A.1.4 Script for Pathview usage

```
pathway.enrichment.plot <- function(metabolite.data=NULL,gen.data=NULL) {
  I <- nrow(metabolite.data); if(is.null(I)) {I<- 0}
  I2 <- nrow(gen.data);if(is.null(I2)) {I2<- 0}
  library(pathview)
  library(KEGGREST)
  data(korg)
  organism <- "homo sapiens"
  matches <- unlist(sapply(1:ncol(korg), function(i) {
    agrep(organism, korg[, i])
  }))
  kegg.code <- korg[matches, 1, drop = F]
  pathways <- keggList("pathway", kegg.code)
  pathways <- pathways[1:91]

  #Llegim el total de compounds a humÃ  que hi ha a la KEGG per a enrichment
  if(!(exists("totalcmp.humans"))) {
    t <- read.table("http://rest.kegg.jp/link/cpd/hsa")
    totalcmp.humans <- length(unique(t[,"V2"]))
  }

  if(!(exists("totalgene.humans"))) {
    t2 <- read.table("http://rest.kegg.jp/list/hsa", sep="\t")
    totalgene.humans <- length(unique(grep("EC:",t2[,2])))}


  map <- gsub("path:", "", names(pathways))  # remove ’path:’
  p.enrichment <- NULL
  pen <- NULL
  map.name <- NULL
  for (i in 1:length(map)) {
    map2 <- map[i]

    pv.out <- try(pathview(cpd.data = metabolite.data, gene.data = gen.data,
        gene.idtype = "UNIPROT",
                           pathway.id = map2, species = kegg.code,
                           out.suffix = "enriched", keys.align = "y",
                           key.pos = "topright", kegg.native = T, match.data = T,
                           same.layer = T,
                           multi.state= T))

    if(is(pv.out,"try-error") | (!is.list(pv.out))) {
      file.remove(list.files(pattern=map2))
    }else{
      tp <- grep("height",colnames(pv.out$plot.data.cpd))
      tp2 <- grep("height",colnames(pv.out$plot.data.gene))
      c <- sum(!is.na(pv.out$plot.data.cpd[,tp+1]))
      c2 <- length(unique(pv.out$plot.data.cpd$kegg.names))
      g <- sum(!is.na(pv.out$plot.data.gene[,tp2+1]))
      g2 <- length(unique(pv.out$plot.data.gene$kegg.names))


      file.remove(paste(map2,"png",sep="."))
      file.remove(paste(map2,"xml",sep="."))

      ##Per a compunds
      if(c==0 & g==0){
        file.remove(list.files(pattern=map2))
        next
      }
      if(is.null(c2)){c2 <- 1e20}
      if (is.null(g2)){g2 <- 1e20}
      p.enrichmentcp <- phyper(c-1, c2,
                               totalcmp.humans, I, lower.tail = F)
      p.enrichmentg <- phyper(g-1, g2,
                              totalgene.humans, I2, lower.tail = F)

      p.total <- p.enrichmentcp* p.enrichmentg

      if (p.total<0.05) {
        pen<-c(pen,p.total)
        map.name <- c(map.name,unname(pathways[i]))
      }else{file.remove(list.files(pattern=map2))
      }
    }
  }
  enrichemnt.pval <- data.frame(pen)
  rownames(enrichemnt.pval) <- map.name
  return(enrichemnt.pval)
}
```

A.2 My internship

During my internship in Yanes Lab, I was required to accomplish different objectives on which this work is based. In April, the NMR and GC-MS data were already analyzed, but the LC-MS were carried out during that month.

SUMMARY: Firstly, I was asked to reproduce the untargeted metabolomics workflow from preprocessing to statistics that they had already performed (see Appendix A.1.3). Since identification requires experience and mastery of chemistry skills, that I performed assisted by my colleagues. In parallel, I was asked to develop a script in R language using mzR package to facilitate them the task of feature identification, into metabolites, from the MS/MS results (see Appendix A.3). Finally, I was required to replicate the proteomics analysis, using R, which was performed in an external lab. Plus, I also created a Python 2.7 script to access an API in order to assess which Uniprot entries belong to enzymatic proteins (see Appendix A.4), which was necessary for using Pathview.

The metabolomics data analysis is based in the workflow we learned in *Metabolomics* course. It started by transforming the raw output files of LC-MS equipment into *.mzXML* files, which can be read by XCMS function, which first analyses the peaks obtained and transforms them into data that R can work with. XCMS is capable of aligning different samples by creating bins of peaks and then groups peaks together across samples by creating a master peak list and assigning corresponding peaks from all samples, these groups define thousands of *features* (mzRT), an m/z charge in a given retention time. This grouping across all samples allows for statistical and value comparison between samples.
In order to avoid noise values, mean intensities were calculated for each sample group (N5, N25, etc.), and were used to filter all those below the mid value of the transition in the sigmoid formed by the mean values. In order to filter out those features that had did not vary highly between sample groups a $\log_2$ fold change was calculated, any sample not above $\log_2(1.5)$ was leaved out. Metabolomics studies normally apply t-test or ANOVA statistical analysis, in this case, a t-test was used since it was intended to always compare against a control condition (N5).
Before further testing, a PCA approach was used to simplify the complex matrix of features, the first two components did not show any clustering or separation of samples, which was unexpected, and thus I decided to use the third component (see Figure 3.2). Based on this lack of variance explained, possibly due to the limited number of replicates, and the low number of features surpassing a threshold of 0.05, a multiple comparisons correction was not applied (Actually, if applied, 0 features had a corrected p-value below 0.05). Appendix A.1.3, shows the essentials of the R script for the analysis mentioned above, this was performed twice, once for each type of ionization, negative and positive, since they can detect different ion types but the file structure is identical.
Using metabolomics databases, such as METLIN or HMDB, the resulting significant features were explored for putative metabolite identity based on m/z mass searches. Then, using MS/MS experiments, they were identified by analyzing, manually, the fragmentation pattern obtained for each feature and by comparing the given pattern to the result from a standard solution of the given molecule. Once the metabolite identiy was confirmed, the KEGG ID code was manually annotated.
Proteomics methodology and data pre-processing was performed by PhD Shabaz Mohammed team in Oxford. The TMT proteomics approach is semiquantitative and thus, the output for each replicate given was a $\log_2$fold change versus the control condition.
A significance B-test were the statistics applied to assess significancy, this test was recommended by the proteomics experts, the tests consists in comparing the values for each protein versus all the values of the same condition, meaning that those protein counts that are out of the total distribution of values are significantly changing; each comparison is performed like a Welch's test. Since no R package is available for gene ontology enrichment using Uniprot accession identifiers as input, it was performed on AmiGO online tool.
A Yanes Lab team member, PhD Maria Vinaixa, found out about the Pathview package, that allowed to map both metabolite and enzymatic data from the KEGG Pathways Database. It was suggested to test this package with that data, but firstly, we were interested in knowing which significant proteins had an enzymatic activity associated. To do so, I came up with the script shown in Appendix A.4, which I partly borrowed from *Programming and database management for Bioinformatics* course.
After the mapping was correctly performed, we started with the data interpretation, which was based on biochemistry knowledge and bibliographic research using PubMed. One of the conclusions in which I mostly participated was the one referring mitochondrial disruption.

A.3 MS/MS spectrum simplifying script

As mentioned above (Section 3), metabolomics bottleneck is feature identification using MS/MS experiments. To do so, the researcher sets the fragmentation system to fragment the desired features according to the resulting mzRT from statistics. Then, the resulting fragments pattern is analyzed using knowledge and databases, the problem is that, normally, the software used to visualize the pattern is slow and it becomes a tiring job. Thus, I was asked to create an R script to get a resulting clean fragmentation pattern from the mzXML files (the same file type used in XCMS) by outputting a pdf file for each MS/MS mzXML input file; indicating the mass (the mz of the feature) followed by the plot of the improved fragmentation pattern. An example of the output is shown in Figure A.1.

```
### Options
table <- F # If the user needs the table with mz and RI to be added to the pdf
ppmthreshold <- 10 # ppm distance between mz to be joined

#### Base Functions ####
RelativeTransform <- function(x){ #Transform to Relative intensity
  maxvalue <- max(x[,2])
  for (i in 1:nrow(x)){
    x[i,2] <- 1000*(x[i,2]/maxvalue)
  }
  return(x)
}

#

FilterMSMS <- function(x,prec,factor){ #Filter values lower than precursor and only higher than factor
   y<- x[intersect(which(x[,1]<prec),
              which(x[,2]>factor)),]
   return(y)
}

#

TestMSMS1 <- function(y,filename,precursor_vector){ #Check and inform if any precursor did
not provide fragmentation spectra
  vector <- unlist(sapply(y,nrow))
  if(any(vector==0)){
    positions <- match(0,vector)
    msg <-  print(paste("In",filename,"the following precursors did not fragment (No fragments
    lower than precursor mass available):\n"
                        ,precursor_vector[positions],".Please, check your MS/MS settings."))
    write(msg, file = "Report.txt",append = T, sep = " ")

  }else{
    positions <- -(1:length(precursor_vector))
  }
  return(positions/-1)
}

#### Simplify MSMS mzXML peaks ####

library(mzR)
lf <- list.files(pattern=".mzXML")

suppressWarnings(
for (x in 1:length(lf)){ # first counter: iterate over files
  filename <- lf[x]
  aa <- openMSfile(filename)

  h <- header(aa)
  msms <- h[h$msLevel==2,]
  precursors <- unique(msms$precursorMZ)

  out1 <- lapply(precursors,function(z){ # iterate over precursors list to get the scans

    q <- quantile(msms[msms$precursorMZ==z,"precursorIntensity"])["75%"]
    peak <- peaks(aa, scans=as.numeric(rownames(msms[(msms$precursorMZ==z&msms$precursorIntensity>q),])))

   if (!is.list(peak)){ # If there is only a scan passing the q threshold (is NOT list)
```

```
     peak <- RelativeTransform(peak)
     peak <- FilterMSMS(peak,z,50)

      if(is.vector(peak)){ # If only one peak passes
        p <- t(peak)
        p <- as.data.frame(p)
      }else{
        p <- as.data.frame(peak)
      }

    }else{ # If there is more than one (is list)
      peak <- sapply(peak,RelativeTransform)
      peak <- sapply(peak,function(x) FilterMSMS(x,z,50))

      # If a precursor loses all signals due to filtering:
      check_sums <- unlist(lapply(peak,sum))
      if (any(check_sums==0)){
        peak <- peak[-(which(check_sums==0))]
      }

      if (all(check_sums==0)){
        p <- matrix(c(0,0),nrow = 1,ncol=2)
      }else{
          if (is.matrix(peak)){ # If only one peak passes the FilterMSMS (does not need to average)
            p <- t(peak)
          }else{
            p <- do.call("rbind",peak) # To do the averages, get all the scans together
          }
          if (!(nrow(p)==1)){ p <- as.data.frame(p[sort(p[,1],index.return=T)$ix,])} # and sort
           them if there is more than 1
        }
    }

        m <- nrow(p)
        p2 <- data.frame("mz"=NA,"RI"=NA)

        for (i in 1:m){
          a <- which((p[,1]< (p[i,1]+ppmthreshold*(z/1e6))  & (p[,1] > (p[i,1]-ppmthreshold*(z/1e6)))))
          if (length(a)>1){
            m1 <- mean(p[a,1])
            m2 <- sum(p[a,2])
            p2[i,]<- c(m1,m2)
            p[a,] <- c(NA,NA)
          }else{
            p2[i,]<- p[i,]
            p[i,] <- c(NA,NA)
          }
        }

        p2 <- p2[!is.na(p2[,1]),]
        if (!(nrow(p2)==1)){
          p2 <- RelativeTransform(p2)
          p2 <- FilterMSMS(p2,z,10)
        }

    return(p2)
  }
)

#### Check if any does not have peaks
missings <- TestMSMS1(out1,filename,precursors)
out1 <- out1[missings]
precursors <- precursors[missings]

#### PDF Plots with or without table ####
  library(gridExtra)
  if (table){ pdfname <- "_Table.pdf"}else{pdfname <- ".pdf"}
  pdf(file=paste("Report_",ppmthreshold,"ppm_",strsplit(filename,".mzXML"),pdfname,sep=""))
  for (i in 1:length(out1)){ # iterate over averaged scans
```

```
    lbls <- out1[[i]]
    if (table){
    if (nrow(lbls)>20){
      while (nrow(lbls)>20){
        plot.new()
        title(paste("Precursor mass =",round(precursors[i],digits=4)),col.main= "gray50",cex.main=2)
        grid.table(round(lbls[1:20,],digits=5),show.rownames = F,
                   gpar.coretext=gpar(col = "black", cex = 0.8))
        lbls <-lbls[-(1:20),]
        }
      plot.new()
      title(paste("Precursor mass =",round(precursors[i],digits=4)),col.main= "gray50",cex.main=2)
      grid.table(round(lbls,digits=5),show.rownames = F, gpar.coretext=gpar(col = "black", cex = 0.8))
      lbls <- out1[[i]]
    }else{
      plot.new()
      title(paste("Precursor mass =",round(precursors[i],digits=4)),col.main= "gray50",cex.main=2)
      grid.table(round(lbls,digits=5),show.rownames = F, gpar.coretext=gpar(col = "black", cex = 0.8))
    }
    }
    lbls[lbls$RI<100,] <- NA
    plot(out1[[i]],type="h",xlab="m/z",ylab="Relative Intensity (to 1000)",xaxt="n")
    title(main=paste("Precursor mass =",round(precursors[i],digits=4)),col.main= "gray50",cex.main=2)
    axis(side=1,at=seq(1,ceiling(precursors[i]),10))
    text(x=out1[[i]]$mz-0.6,y=out1[[i]]$RI,
         labels = as.character(round(lbls$mz,digits=4)),srt=90,cex =0.8)
    Sys.sleep(0.1)

  }
  dev.off()
}
)
```

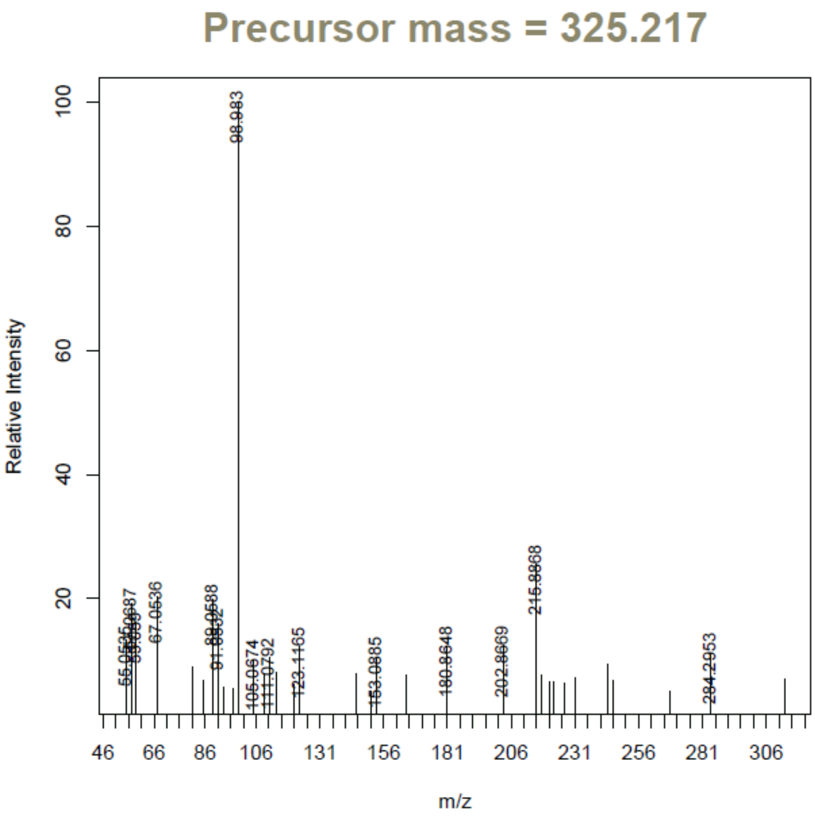


Figure A.1. Example output of a MS/MS spectrum after being simplified

A.4 Accessing EBI Gene Expression Atlas to get enzymes

The objective of this script was to provide a filter Uniprot accession numbers to check wether they have a known enzymatic function, and thus an enzymatic code (EC number). To do so Expression Atlas API was recursively accessed using as input the full list of Uniprot reviewed proteins (downloaded from Uniprot.org) and then searching its entry for an EC code, and if existed saving it.

```
import re, json, requests

p = re.compile(’[0-9]{1}\.[0-9]{1,2}\.[0-9]{1,2}\.[0-9]{1,3}’, re.IGNORECASE)

def makeRequest(ac):
_POSTHEADERS = {’Content-type’: ’application/json’, ’Accept’: ’text/plain’}
base_url=’http://www-test.ebi.ac.uk:80/gxa/api/deprecated?geneUniprot=’+ac+’&format=json’
r = requests.post(base_url,headers = _POSTHEADERS, timeout=9999999999999.99)
a = r.json()
return a

with open("uniprotreviewedlist") as file:
    for line in file:
        line = line.strip()
        a=makeRequest(line)
        if len(a[’results’])>0:
            b=a[’results’][0][’gene’]
            if ’ensfamily_descriptions’ in b:
                name=a[’results’][0][’gene’][’name’]
                description=a[’results’][0][’gene’][’ensfamily_descriptions’][0]
                print line
                if " EC_" in description:
                    c=p.search(description)
                    if c is not None:
                        d=c.group()
                        if len(d)>0:
                            print "added"
                            with open("enzymes_from_uniprotr.txt", "a") as myfile:
                                myfile.write(line+’\t’+name+’\t’+d+’\n’)
                                myfile.close()
```